HAAGA-HELIA
University of Applied Sciences

Targeting of Data Quality Monitoring: Which Subset of Business Master Data is Critical?

Tanja Penttinen-Santos da Silva

Thesis Report

BITe Thesis

26.10.2014

HAAGA-HELIA
University of Applied Sciences
29.10.2014

Abstract

BBA, Business Information Technology

Author or authors Tanja Penttinen-Santos da Silva	**Group or year of entry** 2008
Title of report Targeting of Data Quality Monitoring: Which Subset of Business Master Data is Critical?	**Number of report pages and attachment pages** 91+12
Teacher(s) or supervisor(s) Ralf Rehn	

This thesis forms a framework that does not currently exist for identifying the most critical data elements for data quality monitoring and a construct to represent these data. It was commissioned by an information management company to support the setup process of their data quality monitoring tool in a customer environment.

The thesis hypothesis is that it is not viable to aim for 100% data quality on all of an organization’s data. Instead the subset of data that is the most critical and offers the most benefit if of good quality should be targeted. The thesis suggests data quality monitoring as the means for data quality improvement. The main objective of the thesis is to define a generic subset of business master data that is most critical for data quality monitoring for most organizations regardless of industry.

The introduction and theoretical parts of the thesis build a big picture for the reader on the importance of data and their quality as well as introduce data quality monitoring. The main linkage between the theoretical and empirical parts of the thesis is the chapter explaining the connection between data quality and business processes. This also introduces the logic used in the thesis for defining in which business/data intersections critical data lie.

The empirical research was conducted in the summer of 2014. A preliminary data quality monitoring targeting construct was built based on the theoretical research and commissioning party representatives’ years of experience with data quality issues faced by organizations. Thematic interviews were conducted with data quality experts to verify/challenge the preliminary construct. Interviews were analysed to realign the construct.

As a conclusion, the final data quality monitoring targeting construct is introduced with recommendations for possible further development. The construct will be utilized as a basis for setting up organization-specific data quality monitoring. Conclusions also include additional approaches for identifying critical data.

Keywords
Data quality, Data quality monitoring, Business critical master data, Master data, Data management

Table of Contents

1 Introduction

> "Customers have been spoiled. Thanks to companies such as Amazon and Apple, they now expect every organization to deliver products and services swiftly, with a seamless user experience. -- They expect all service providers to have automated access to all the data they provided earlier and not to ask the same questions over and over again." (McKinsey&Company 2014.)

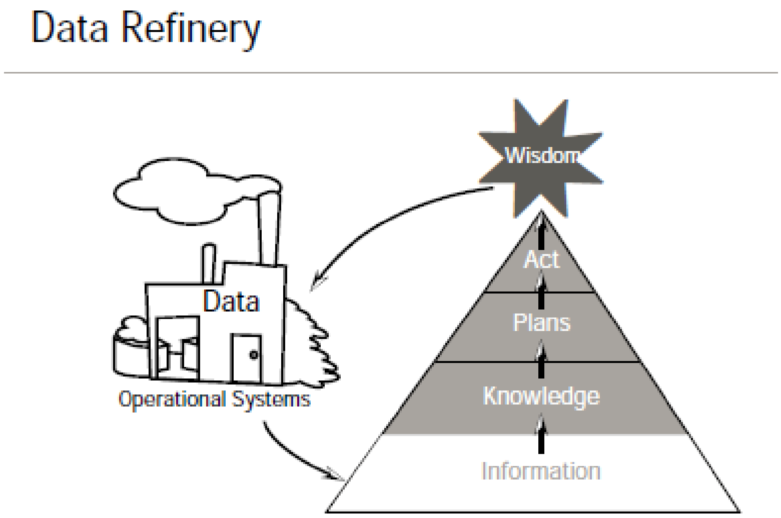
Consumers are expecting 24/7 service in a global omni-channel environment: services need to be available on the mobile, on PC, on tablet, face-to-face –regardless of the time, location or distance from service provider. On top of this, the expectation is of ever more personalized, targeted service offering – without compromising personal privacy, of course. There is great hype about big data and all its possibilities. Everyone wants to get on the business intelligence and high-performance analytics train. Many fail to realize what it is that is making all of this possible.

There is the state-of-the-art technology for organizations to utilize, of course. But, what good is a piece of technology if there is no content and no context? No data. No information. Moreover, data that are the right data. Consistent data. Complete data. Up-to-date data. Only then, can an organization make the most of the technology that is on offer and make even long-term, strategic decisions that carry signifigantly less risk than if no confidence could be put on the data behind it all. All of the "norms" of today's world, described in the previous paragraph, are facilitated by data – *data of good quality*. Today's organizations' data quality challenge summarized would possibly look like this: big data vs. row-level database. How to make all this data work both as a whole and on a detailed level?

"Amidst the increasing quantity of available information, the quality of information becomes a crucial factor for the effectiveness of organizations and individuals." (Eppler 2006, 1.) Reports show that organizations are struggling with their data quality so much that data are not trusted anymore, and resources are spent in vain for looking for verification to the data's accuracy and other quality matters (e.g. Redman 2013). The requirements for data and data quality are heavy and evergrowing. Organizations that were fairly lost before (e.g. Battini & Scannapieco 2006, 1; Eppler 2006, 1; From,

R. 12.5.2014), are now bewildered by the daunting task of making their data of good quality – and keeping it that way. This can especially be the case for organizations with no data management background, when the integration/system architecture is complex and business is siloed, functions operating seemingly independently from each other.

To start, organizations should realise that it is "better to have less data of good quality than more poor quality big data." (Aiken, P. 14.1.2014.) There is no point in trying to build a high-rise Big Data tower, if the foundations at the bottom of the their data structures are not solid. Only once foundations are solid, can more data be embraced and made to work for the organization. Pictured below, an example of the refining data goes through to become an enabler for the digitization era we are in, and how the base data should be in order to facilitate this refining.



Picture 1. Data Refinery – showing the importance of the data and information at the foundation of any data initiative. (Eckerson 2002, 5.)

At a recent Big Data seminar (e.g. From, R. 12.5.2014), one of the main issues presented regarding all the information that is available today, is not knowing what to do with it, how to utilize it. Aiken (14.1.2014) summarized it well: "Having more data does not substitute for thinking hard, recognizing anomalies and exploring deep truths. You need the right approach."

There are many techniques and methodologies aimed at improving and maintaining good data quality (e.g. data quality extensions of the Entity Relationship Model, Information Product (IP-MAP) Model, Total Data Quality Management (TDQM) Model). These models all operate on a top-down manner, meaning that an organization would first create a data governance/management strategy and start working its way down the organizational structure to take measures for improving data quality (e.g. Batini & Scannapieco 2006, 51-68.) The top-down approach is very heavy and devours resources, without many immediate results. The result of these approaches, however, is what should be aimed at: getting to the root causes of data quality issues.

All of the methodologies also agree on what one of the first steps of any data quality improvement initiative should be: recognise the business-/mission-critical data (e.g. Aiken 8.4.2014; Batini & Scannapieco 2006, 63-64). As Aiken (14.1.2014) put it: "Focus on your most important data assets and ensure your solutions address the root cause of any quality issues – so that your data is correct when it is first created." The problem here is that while it is all good and great to say that the solution should address the root cause of an issue, how to do this without deploying heavy methodologies? Aiken (14.1.2014) continues: "Experience has shown that organizations can never get in front of their data quality issues if they only use the 'find-and-fix' approach." Can a happy medium be found that would allow the combination of quick fixes and a structured data quality approach without taxing too much of organizational resources?

We are in the agile, digitization era. Developments need to happen rapidly and time to market needs to be minimal. Organizations are not willing to commit to heavy, long-lasting intiatives. This is especially true for something as abstract as data quality. "Fail-safe" approaches (if it does not work, it is possible to bury it quickly and move on without harm to business) are ever more popular in today's competitive world. Therefore, it is assumed that organizations would much more likely take on a bottom-up approach to data quality where the starting point is addressing single pain points of business - concrete, high-cost issues whose solving would provide clear value to business.

It is this thesis's driving idea that data quality monitoring can offer exactly the kind of light approach to data quality improvement as needed by organizations. It can be outsourced or done in-house with minimal resources, starting on the known pain points of the organization, possibly building to a deeper understanding and correction of the root causes. It is quick to setup and offers value from the get-go, while still offering the easy abort option should organizations not see the value of the monitoring.

Existing data quality improvement methodologies describe, at length, the different steps of improvement topics and techniques: data quality dimensions, metrics, calculations, etc. The central topic and starting point of all methodologies (e.g. Batini & Scannapieco 2006, 63-64) is the one question that none of the approaches or sources answers in pragmatic, easy ways: *What* data exactly should be concentrated on for most benefit? What is the *most business-critical data*?

The thesis operates on these three assumptions:

1. Organizations are at a loss as to what data quality issue to prioritize.
2. Organizations are not willing to invest in a top-down, heavy approach for something as vague as data quality but want fast, easy solutions.
3. There is not much, if any, existing research on what data exactly is critical to address for quality issues.

2 Thesis Background

2.1 Commissioning Party

This thesis was commissioned by Datpro Oy. Datpro is a small, privately-owned information and information quality management company established in 2010. The company has dealt with numerous customer companies who are struggling with data quality issues.

The owners and employees of the company have some half a century of experience in information management. One of the owners of the company is the Vice Chairman of the Finland division of the world-wide Data Management Association (DAMA) and has given speeches at the Massachusetts Institute of Technology (MIT). There is a lot of knowledge held in-house.

Datpro have recognised data quality monitoring and, in their pragmatic, straight-to-the-point manner, especially the target data for data quality monitoring as a gap in data quality research. Datpro are developing their own data quality monitoring software based on their knowledge from years of exposure to data quality issues.

The outcome of this thesis will serve as a basis for an initial kick-off session with customers when setting up Datpro's data quality monitoring as a service (DQMaaS) product, that is currently being developed. The outcome will kick-start the recognition of the customer-specific target data for data quality monitoring during this session by highlighting/bringing out for comparison the common critical data elements in many organizations across the board.

2.2 Thesis Hypothesis & Objectives

The thesis hypothesis is that an organization should not and cannot aim for 100% data quality but should instead concentrate on the assumedly small subset of data that, for their organization, is the optimal mix of data elements that is the most critical and will offer the most benefit if of good quality. This subset should be aimed for perfect data quality with.

This hypothesis aligns with a rule that many data experts (e.g. Aiken, P. 8.4.2014; Kontra, K. April 2014) commonly accept as a truth: The 80/20 rule of Pareto applies also to data quality – 80% of data quality issues are caused by 20% of the data. The rule can also be interpreted so that 20% of an organization's data is critical and 80% of the data is not of as much importance *in terms of data quality* (criticality for business can be high but data are not causing so many data quality issues).

Not only with big data, business intelligence and other major developments in a business's data environment but simply through its evolving organizational environment and customer engagements, data quality cannot be looked at as a one-off venture or a stable state of things. To make sure that the small subset of most critical business data stays in shape, it needs to be monitored for quality on a regular or continuous basis.

The main objective of this thesis is to, on a high-level, define the subset of data that is the most critical for an organization to monitor for data quality.

This thesis is not trying to suggest that there is a blanket template or approach for the targeting of data quality monitoring - every company's data and quality issues are different. However, during years of data work, it has become clear to the commissioning party representatives (and other experts that have spoken at length on the subject, e.g. Peter Aiken) that many, if not most, companies have data quality issues with at least partly the same pieces of data.

This thesis is attempting to form a framework for identifying those data elements that are the most beneficial to the organization to monitor for issues as they happen, and find root causes for. Most likely this set of data are also those data that cause data quality issues that are critical to the achievement of business objectives, and can offer one approach towards identifying the most critical data. The framework that is the aimed deliverable of this thesis, is meant to be applicable to as wide a range of different businesses and organizations as possible.

2.3 Research Question & Structure

The research question the results of this thesis are attempting to answer is:

- What is the most business-critical subset of all of a business's master data in terms of data quality and should therefore be targeted by data quality monitoring?

The thesis consists of four parts: the thesis background (chapters 1-2), the theoretical research (chapter 3), the empirical research (chapter 4) and the conclusions (chapter 5).

The introduction and thesis background chapters build a big picture for the reader on the importance of data and their quality. They describe the hypothesis, objectives and purpose of the work.

The third chapter explains the concept of data quality and data quality monitoring, what effects poor data quality can have for a business and to indicate how critical these effects can be. Through this theory, the report is leading the reader towards the more pragmatic objective of the thesis – the recognition of the critical dataset for data quality monitoring. Paving the way for recognizing this critical dataset is done by theoretical study of key business processes and the data that those processes utilize, attempting to demostrate the complexity of the business data environment in terms of data quality. This supports the recommendation of many literary sources that any data quality exercises should be started by recognising the key data within the key processes for the organization and its business objectives as a whole. The chapter on data quality and business processes (3.3) is the main linkage between the theoretical and empirical parts of this work.

The fourth chapter details the empirical part of the thesis work. It introduces a preliminary set of critical data for data quality monitoring that has been put together based on theoretical background and commissioning party expert experiences, and summarizes conducted research interviews with data quality experts to challenge/verify the previously built construct.

Last, conclusions and recommendations made based on the theoretical and empirical parts of the work on what should be the generic subset of data that an organization should monitor for data quality are presented in chapter five. This subset of data is presented in the form of a construct, a data quality monitoring targeting matrix. Recommendations are made for further development of the matrix.

2.4 Research Methods

The thesis is a combination of constructive research and grounded theory.

Constructive research was defined by the HAAGA-HELIA University of Applied Sciences Thesis Guidelines Working Group (2014, 21-22) as follows:

- “ focus on real-life problems to which it is important to find a solution
- the generated solution is an innovative construction
- the solution is linked integrally with acquired knowledge in the area

- the project includes an attempt to implement the construction, i.e. testing in practice (…due to the limited time available for completing a bachelor's thesis, it may be sufficient to interview experts on their views about the applicability of the solution)
- the process involves close cooperation between the researcher and people dealing with the matter in practice, with the aim of experiential learning"

The thesis is aiming to solve a very real problem: data quality issues experienced by a large number of organizations. There is no existing framework for data quality monitoring targeting, and therefore, a construct is generated to address this. The construct is preliminarily built in close cooperation between the researcher and the commissioning party and challenged/verified with research topic experts. As the topic is fairly new, and not many sources exist, some topics in this thesis (such as data quality monitoring) are also addressed by close cooperation with commissioning party representatives who are dealing with the matter in practise. The topic of this thesis also qualifies and compelled the researcher to use also the grounded theory research approach which is used when "sufficient data on the topic does not exists… or when a fresh approach to the topic is sought". Data are collected as detailed below, then interpreted and a theory is formulated. (HAAGA-HELIA University of Applied Sciences Thesis Guidelines Working Group, 25.) Formulating a theory and then build are construct to address/test it, in essence, is what this thesis does.

Data collection for the theoretical part of the thesis was done through content analysis of literary and electronic sources. This data collection method was chosen as good quality written sources for the majority of the thesis topics exist already and therefore do not need to be invented by this thesis. This data collection method was complemented by some open interviews/discussions on data quality and data quality monitoring with a commissioning party representative. Open interviews/discussions were decided on as a data collection methods for certain subjects (mainly, data quality monitoring) where no detailed literary or reliable electronic sources exist or were found by extensive searching. The topic of data quality monitoring is mainly covered by commercial sites offering data quality monitoring as a service and are therefore equal in value to the information provided by the commissioning party.

The commissioning party representatives and their expertise were utilized in some parts of this thesis. As detailed above, those theoretical data quality topics that did not have many existing literary or electronic sources, or where the sources were not objective, were discussed with the commissioning party (mainly one person, Lead Advisor Kimmo Kontra). The preliminary construct in the empirical part of the work was also built together with the commissioning party, through studying data quality assignments completed by the company previously and through discussion and/or e-mail exchange with company representatives on their previous projects (with Datpro and before). Any input received from commissioning party represesentatives was analysed critically.

Data collection for the empirical part of the thesis was done by open and thematic interviews. As mentioned in the introduction, no leading data quality literature or electronic sources address the research question of this thesis. Therefore, new information/point of view was needed to be seeked by other methods. Interviewing was seen as the best method for such an abstract topic as data quality. The number of persons able to answer data quality questions from a professional and wide perspective in Finland were estimated to be quite small, and for this reason also, e.g. questionnaires as a method were abandoned. The nature of the topic also requires personal interaction with participants. Interviewees were chosen in collaboration with the commissioning party whose employees have met many data professionals during their decades in information management. Interviewees were chosen based on their experience and proven knowledge in the research topic. The thematic interviewing method was chosen to encourage open discussion, while still being able to control the matters discussed. The expertise of interviewees was deemed so high that a more structured interview template was not needed to lead the discussion. The analysed results of the interviews were attempted to be re-verified with interviewees after all interviews were conducted, to further prioritize the criticality of the data elements of the concluded subset of critical data for data quality monitoring.

Comparative analysis method was used to compare the information from theoretical and empirical parts to make conclusion on the research question.

2.5 Thesis Scope

The factors affecting data quality in a deteriorating manner are not covered in-depth in this report, as the aim of data quality monitoring in the context of this thesis is to discover data quality issues that need to be investigated for root causes. This is, therefore, a consequent step from recognising those issues through monitoring the critical data to be identified by this thesis.

The techniques and methodologies of improving data quality are not covered in detail. For the purposes and hypothesis of this thesis, it is sufficient to acknowledge that numerous frameworks and methodologies exist, and those could well be the resulting action of organizations once they start with the bottom-up approach of using monitoring to uncover quality issues in data.

Any commercial data quality tools are not covered or promoted by this thesis. It is sufficient to recognise that many tools exist to assist organizations with their data quality efforts.

Data management/governance, although very central concepts in terms of keeping data of good quality, are also not covered in this thesis, due to the very specific nature of the topic.

The author would also like to emphasize that this thesis will not give any exact cost-benefit calculations for the dataset discovered to be the essential target for data quality monitoring. In 2009, the annual Information/Data Quality Salary and Job Satisfaction Report (Lintag, Pierce & Yonke; in Kontra 2010a, 5) stated that: "81% of respondents indicated that demonstrating the value of high quality data to their organizations is the single biggest challenge in data quality field". To do so on a generic level, without specificity to an organization, very likely is an even higher complexity task. Therefore, this thesis aims to give an indication as to the most likely source of benefit to an organization only.

3 Theoretical Background

3.1 Data and Data Quality

"Data and information are now as vital to an organization's well being and future success as oxygen is to humans. Without a fresh supply of clean, unpolluted data, companies will struggle to survive and thrive." (Eckerson 2002, 32.) This was most likely an accurate, but not widely recognised, statement in 2002, but today, more than a decade later, it will undoubtedly be considered fact by any data expert in the business world.

In any work related to data, the definition of data in that particular context needs to be made. Data by nature are detailed, and therefore, invite a detailed investigation into their essence. There are various definitions and classifications of data. Some consider data and information interchangeable terms, others insist on a clear division between their concepts.

Data, information and knowledge can be seen to form a hierarchy. In this hierarchy, data consist of bits, nuggets without context; they are the symbolic representation of a real-world state or event and the foundation of any information system. Data in themselves do not usually give value to their stakeholders or the system they are part of. Information is data that have been given a context (format and representation), and can develop into knowledge by human learning, rationalization and observation. Information is the value-added version of data. A typical elaboration of the data-information-knowledge hierarchy uses a number, say 1 000 000, as an example of data. It can mean anything, until represented in a certain way, e.g. "The turnover of company A is €1 000 000". It has become information. (Peltonen 2006, 8.) Eppler (2006, 22) described information as "potential knowledge that has to be internalized by the receiver". Knowledge would be the interpretation of information: knowing the significance of this statement of a company's turnover mentioned above in comparison with other information, and the ability to use it to create value (Peltonen 2006, 8). As picture 1 in the introduction chapter shows, knowledge is not necessarily even the tip of the "data -refining iceberg".

As the previous data-information-knowledge definition implies, data are often seen as fact, objective representations of a real-world state, whereas information is seen as a subjective interpretation of the data, moulded by context. However, most business-critical data such as taxonomic data (e.g. categorizations of products, classifications of customers) are actually an organization's perception of the real-world state. Can those therefore be classified as data or should it be information? (Kontra 2010b.)

In this thesis, data and information are considered to be interchangeable terms as information cannot exist without data (Pelkonen 2006, 8-9) and information can easily disguise itself as data (Kontra 2010b). It is information that adds value to an organization, also in turn having the ability to cause costs and losses. Knowledge is a thoroughly subjective concept and therefore not concentrated on in this thesis.

The classifications of data are also a living concept that vary from master data to transactional data, structured to unstructured data, raw to productized, elementary to aggregated, operational to analytical. Another layer of complexity is added by looking at the time dimension of data to determine whether data are stable or changing frequently, and affects the perception of data quality. (Batini & Scannapieco 2006, 6-9.)

This work concentrates on structured master data. In structured data, "each data element has an associated fixed structure", and can therefore be measured for monitoring more easily than unstructured data where no specific structure is specified (Batini & Scannapieco 2006, 6). Master data are the core business-critical data that are used by multiple business processes/functions and systems across an organization, e.g. customer and product data (e.g. Kolehmainen 2011). Master data are fairly stable, describing the characteristics of an object, and form the foundation for any operational or analytical activities to take place. In general, master data are "original" and not derived from (result of) other data, so e.g. although analytical data can be very business-critical to an organization and shared by many functions, they are still based on other data (master data) and would therefore not be considered master data. "Master data are the critical nouns of a business and fall generally into four groupings: people, things, places, and concepts. - - For example, within people, there are customer, employee, and salesperson. Within things, there are product, part, store, and asset. Within concepts, there are things like contract, warrantee, and licenses. Finally, within places, there are office locations and geographic divisions." (Haselden & Wolter 2006.)

Transactional data, e.g. sales orders are also a product of operational, day-to-day activities that utilize master data (e.g. the above-mentioned customer and product). Both master data and transactional data, along with other types of data, are then utilized for analytical activities and consequent data. Any quality issues in critical master data affect data in all consequent activities, and the issues multiply - growing in transactional use and becoming alarming in analytical use where data can potentially affect long-term plans of an organization. Master data are therefore a self-evident target for monitoring in an organization-wide initiative, to prevent the escalation of quality issues.

3.1.1 Definition of Data Quality

Data quality is an abstract concept that is somewhat complex to define. It is also a fairly subjective matter, as the same data can be of good or bad quality depending on one's point of view. On the whole, data quality is still quite a living concept, where commonly agreed terms and guidelines are still being defined. Some definite consistencies can be found between the different sources though, and these are becoming the norm.

Very often data quality can mistakenly be considered synonymous with accuracy of data (e.g. correct spelling) (Batini & Scannapieco 2006, 4). Data quality is much more. The most popular definition for good quality is one that was first introduced by Juran et al in 1974 and later adapted to be data-related: "Data that are fit for their intended uses in operations, decision-making and planning" (e.g. Pelkonen 2006, 9; Roebuck 2011, 1), and this further implies that they meet the requirements of their authors, users, and administrators (Aiken, P. 8.4.2014). Evans & Lindsay (1999 in Eppler 2006, 20) defined quality as "the totality of features and characteristics of a product or service that bears on its ability to satisfy given needs". This definition aligns with the approach where data are considered a product manufactured as any physical product, and this information product's users are its customers (Batini & Scannapieco 2006, 61). Data quality could in this context be an add-on product or feature that could fulfil customers' further requirements than basic functions can fulfil or exceed their expectations for the product.

Data represent real world objects in a most versatile way, being capable of representing e.g. measurements, events, characteristics of people and sounds (Batini & Scannapieco 2006, 6). From this can be derived another important criterium for how good quality data are: how correct or accurate their representation of the real-world objects that they refer to are (e.g. Roebuck 2011, 1).

The most simplified framework, and fitting for the approach of this thesis, for representating the types or levels of data quality is the semiotic framework introduced by Price & Shanks (2005): The framework divides data quality into three semiotic levels: syntactic, semantic and pragmatic, which respectively refer to form, meaning and application (or use) of data. The syntactic level data quality is usually fairly simple to implement in databases with different kinds of simple constraints, but semantic and pragmatic level data qualities are somewhat more complex to implement and can be more easily approached with data quality monitoring, if approachable in an automated way at all.

	Syntactic	**Semantic**	**Pragmatic**
Quality question assessed	Is information system data good relative to information system design (represented by metadata)?	Is information system data good relative to represented external phenomena	Is information system data good relative to actual data use, as perceived by users?
Ideal quality goal	Complete conformance of data to specified set of integrity rules	1:1 mapping between data and corresponding external phenomena	Data judged suitable and worthwhile for given data use by information consumers
Operational quality goal	User-specified accetable % conformance of data to specified set of integrity rules	User-specified accetable % agreement between data and corresponding external phenomena	User-specified accetable level of gap between expected and perceived

			data quality for a given data use
Quality evaluation technique	Integrity checking, possibly involving sampling for large data sets	Sampling using selective matching of data to actual external phenomena or trusted surrogate	Survey instrument based on service quality theory (i.e. compare expected and perceived quality levels)
Degree of objectivity	Completely objective, based on integrity conformance	Objective except for user determination of relevance and correspondence	Completely subjective, dependent on user and use

Table 1. Quality categorization information (Price & Shanks 2005)

Eppler (2006, 20-21) concluded quality to be of twofold nature: quality has a subjective (meeting expectations) and objective (meeting requirements) component, or a relative (satisfying needs) or an absolute (meeting specifications) dimension. Like Eppler, most methodologies and approaches related to data quality are concerned with categories of attributes (or dimensions) of data to evaluate data quality (e.g. Batini & Scannapieco 2006, Roebuck 2011), discussed in chapter 3.2.1.

3.1.2 Costs and Benefits of Data Quality

Data quality is considered an IT issue, and this is the main misconception of organizations. It is also not a business issue. Data quality affects and is affected by everyone in an organization. There should be partnership between the business and technical stakeholders of any data quality issues. (Harris, J. 2009.)

Information systems where all technical constraints possible are available, are just as likely to have data quality issues as the ones that are somewhat lacking in technical capabilities. The technical capabilities need to be harnessed according to business rules and processes. More often than not, data quality issues are a result of not understanding or facilitating the business requirements for data. (Kontra, K. April 2014.) As explained in the previous chapter, the quality of data is very much concerned with the context where data is used. If data is of poor quality, the biggest effects will also be felt in the business's operational and analytical activities using and creating the data, not in their IT department.

Data quality issues are behind the many everyday events that are often blamed on other factors, not making the connection to the root cause. For example, if a parcel is delivered to the wrong address or delayed, it is blamed on the malfunctioning post office. Instead, the more likely cause is incorrect address data in the address database. (Battini & Scannapieco 2006, 1.) In the context of data quality, a mistake like this is unlikely to be unique, and a further investigation should be started into the matter. In most occurrences of the above example though, the address would assumedly either not be corrected at all or only be corrected for that one instance of an address. In either of the two cases, costs would be incurred to the organization sending or delivering the parcel: the correct address for the delivery needs to be found, it will possibly be corrected to the database, a new delivery of the parcel will need to be organized and executed. In 2002, a report stated that 23.6 percent of all U.S. mail was sent to incorrect addresses, partly due to the high volume of Americans (some 45 million) moving every year (Roebuck 2011, 2). Possibly this number has decreased somewhat with the growing awareness and developed tools of data quality, but without a doubt an issue still exists, even with this simple a piece of data. So, although the extra activities caused by the incorrect address in the above example might seems like small tasks that only take a few minutes, when you consider the volume of this kind of mistakes, it would surely be more cost-effective to launch a proper investigation into the matter of incorrect addresses in the database and fix the root cause of them being created/becoming outdated in the first place (partly from Batini & Scannapieco 2006, 1).

All poor quality data cause costs or losses to an organization, either directly or indirectly (e.g. Batini & Scannapieco 2006; Funk, Lee, Pipino & Wang 2006). Hristova et al. (2013) recognised four main types of effects/costs of poor data quality:

- Costs and reduced productivity (fixing data is expensive and takes time that could be spent more productively)
- Bad business decisions (when data used in business analytics or intelligence reporting is incorrect, potentially leading to serious financial implications)

- Low customer satisfaction/public image (providing customers inconsistent/incorrect data or bad customer service due to poor quality data results in decreasing trust and market share losses)
- Expensive and dragging systems implementation/maintenance process (solutions built on incorrect data or data built for a solution instead of vice versa result in constant maintaining and unstability)

As mentioned in the previous chapter, the higher up the issues with data quality are allowed to escalate, the greater the problems become, and the bigger the effects/costs rise. Logically, the data whose poor quality costs the company the most (directly or indirectly through negative impact of data on the business), are most likely the ones that the organization would gain the most out of improving the quality of. Unfortunately in the field of data quality exact calculations are quite difficult to make, although there is a substantial amount of literature on cost/benefit analysis including indicative formulas for calculation. There are many "invisible" costs such as loss of image and loss of potential customers that organization might not even be aware of. Benefits of data quality (monitoring) are implied on a higher-level in the practical part of this thesis and in the conclusions, through identification of the business-critical dataset that is recognized through pain points of business operations. As this thesis concentrates on identifying the data elements whose monitoring would be most beneficial on a generic level, this kind of high-level approach to cost/benefit analysis is sufficient and all that can be expected: detailed cost/benefit analyses cannot be done on a generic level.

As background information, it is worthwhile mentioning some types of costs and benefits that have been recognized in relation to data quality. English (1999, 209-212) divided costs from poor quality data as:

1. **Process failure costs**

   Process failure costs incur when processes are stopped due to poor quality data, such as in the example given of an incorrect address data resulting in the failed delivery of correspondence. The costs can be irrecoverable (e.g. resending of delivery, costs of first delivery are never recovered), liability or exposure costs (e.g. not complying with authority regulations due to poor data quality) or recovery costs (e.g. winning back the trust of dissatisfied customers).

2. **Information scrap and rework (i.e. maintenance) costs**

   Information scrap and rework refer to manufacturing terms that apply to data also: defective information needs to be cleansed (reworked) or marked as error/rejected (scrap). The costs incur from redundant data handling and support, searching for missing information, re-running failed processes, workarounds and decreased productivity, verification, software rewrite, cleansing and correction as well as software for cleansing data.

3. **Loss and missed opportunity costs**

   Lost and missed opportunity costs incur due to income not realized due to poor data quality. These costs are not incurred only from immediate sales lost but also from customer lifecycle point of view, where sales could have been considerable. Lost opportunity implies losing an existing customer, missed opportunity implies losing also prospect customers either due to alienating existing customer (and not being recommended by them to potential customers) or ineffective development & marketing.

4. **Assessment and inspection costs**

   Assessment costs incur from assuring processes are running properly (e.g. data quality monitoring) and must lead to process improvement to add value, inspection costs incur from assessing data quality.

   Any data quality assessment method such as data quality monitoring is a cost item and not a value-adding initiative in itself: it is what an organization does with the results of the assessment that can add value. It is an objective of data quality assessment also to minimize the cost of the actual assessment.

5. **Process improvement and defect prevention costs**

Process improvement and defect prevention costs that incur from actually doing something to prevent poor quality data from existing and therefore creating the most benefit for an organization are often mistakenly concentrated on when starting a data quality improvement initiative.

The real costs that must be considered are costs resulting from not creating and maintaining quality data in the first place. These costs have often been accepted as the normal costs of business. (English 1999, 213.) The last two cost types above are the only acceptable costs of data quality (Pelkonen 2006, 34).

Batini & Scannapieco (2006, 94) classify benefits gained from data quality improvement in three:

1. **Monetizable**

   Monetizable benefits can be directly connected to increased revenues or decreased costs.

2. **Quantifiable**

   Quantifiable benefits cannot be directly expressed in monetary terms but are related to some measurable dimension, e.g. time. For example reduced wasted time by organizations is a quantifiable benefit. In some cases the quantifiable benefits can be converted to monetary benefits if reliable conversion functions exist.

3. **Intangible**

   Intangible benefits cannot be expressed in a numeric way, therefore, they cannot be measured. This kind of benefits include increased customer/employee satisfaction, increased service quality, etc.

On a high-level, the cost/benefit calculation for data quality improvement activities is simple. Costs of data quality issues and improvement activities need to be weighed against the benefits that are estimated to be had from the improvement activities. These benefits are made up of such elements as cost or loss avoided by improvement as well as possible extra income, e.g. money saved by decreased need for additional maintenance due to errors, high customer satisfaction due to good quality data enabling good customer service resulting in keeping customers loyal/buying more, attracting new customers with correctly targeted marketing that is due to correct analytics done with good quality data, etc. The value of the improvement activities is positive if the benefits outweigh the costs. If the value of the improvements is negative, it is better to do nothing. (Batini & Scannapieco 2006, 88-95; Funk et al. 2006, 16.)

3.2 Data Quality Monitoring

"You can't manage what you don't measure" is an old management saying that is believed to be true today (e.g. Reh 2014). Of course, there are many that say the complete opposite and claim that the most important things cannot be measured (e.g. Ryan 2014). In the world of data quality, the truth lies somewhere in between. There are important aspects of data quality that are very hard to measure, such as the subjective perspective or reputation of data and data quality. However, with the help of dimensions (described in the following chapter), the attempt is very much to measure as wide and varied amount of aspects of data as possible. The method of how to measure is a different matter: in subjective dimensions, the best way might be through surveys and interviews, whereas with more objective matters, using predefined quality targets as a benchmark, quality can be measured by monitoring. (Partly adapted from Funk et al. 2006, 33-40.) Subjective quality assessment and monitoring should involve a Subject Matter Expert (SME) or similar to assist in the evaluation of what the targets are and what is important to the organization (Roebuck 2011, 11).

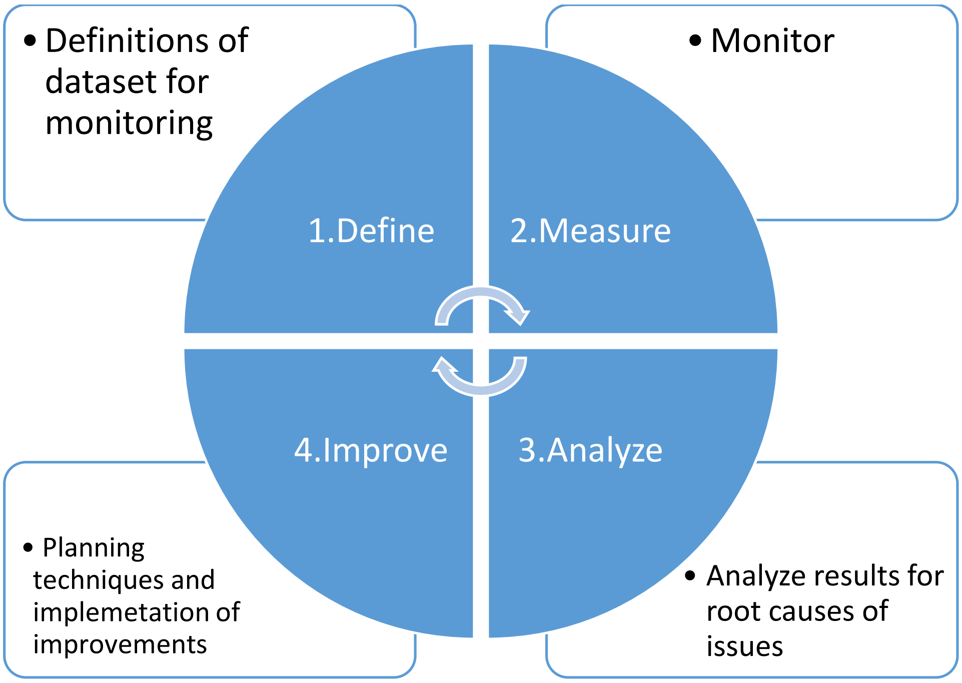
Monitoring in general means "observing and checking the progress or quality over a period of time or keeping something under systematic review" (Oxford Dictionaries 2014). "Data quality monitoring is the process of examining your data over time and alerting you when the data violates any business rules that are set." (Oracle 2009). The important matter "is not to measure the quality of data in isolation, but to measure the quality of data within the relative context of a specific business use, or in other words, to measure the ability of a data provider to service the needs of a data consumer" (Harris, J. 2011). Already 15 years ago, English (1999, 139) concluded that the biggest pitfall of data measurement systems was not measuring the right things, that would "have a positive impact on knowledge workers and customer satisfaction".

Monitoring data quality is a way to assess the quality of a selected dataset at any given time, by e.g. comparing business-critical data in an organization's systems against data quality targets (e.g. "address data needs to be complete and accurate") or rules (e.g. a naming convention or pattern for value) set for that data. These can be in place to reach semantic and pragmatic levels of data quality. In practise, a way to measure data quality is often through checking compliance of data to set business rules (e.g. "If this material's origin is USA, it needs to have this additional field filled. The value for that field needs to exist in this pre-defined allowed list of values."), to make sure data are "fit for their intended use" (i.e. of good quality). Using the fitness for use criterion establishes a connection to the business impact of data quality measurement instead of simply measuring the potential business impact by aligning to the real-world object the data are based on. Additionally, meaningful metrics that provide business insight need to be used, otherwise measuring data quality is pointless. Analysis should be done between existing data, business needs and technology; this should serve as a basis for data quality business insight. (Harris, J. 2011.)

Business rules that are in need of monitoring are most likely too complex, and therefore expensive, to implement with traditional database constraints or custom-built solutions directly into the database or application (expenses include labour costs and other direct or indirect costs such as cost of updating custom solutions, all for an uncertain amount of ROI as is often the case with abstract data quality). Especially rules for data elements that exist across system landscape in various different solutions, are challenging to implement directly into the systems. (Kontra, K. April 2014.) If an issue with a data element is seemingly small (e.g. previously mentioned example of a letter or delivery sent to wrong address due to incorrect address information), the threshold to act on it is high if no "quick and easy" solution is at hand. Often organizations also do not impose data quality rules such as mandatoriness of a value for system fields due to process limitations, e.g. not all (mandatory) data are available at the time of creation of a data object and it is not wanted that dummy values are used, etc. These fields need to then be monitored in one way or another to make sure they are eventually filled in.

(Unfilled fields would eventually be noticed when a process fails due to the incompleteness of the data, but by that time potentially high costs, that could have been avoided, have already been incurred.)

The data quality monitoring approach an organization takes could for example follow the well-known Deming quality management cycle (also known as the PDCA cycle) that was refined to the data-oriented TDQM (Total Data Quality Management) cycle (Eppler 2006, 21), pictured below. The cycle consists of first defining what to measure, followed by doing the actual measuring, analysing the results of the measuring and finally, making improvements on data based on results. Then the cycle starts again, with the same or new measuring definitions. Once data quality errors have decreased to a tolerable level or are not found in the measuring anymore, an organization might want to adjust the measuring to include other data. The results of this thesis will form part of the planning/definition part of the cycle for data quality monitoring as a means of measurement.


• Definitions of dataset for monitoring
• Monitor
1.Define
2.Measure
4.Improve
3.Analyze
• Planning techniques and implemetation of improvements
• Analyze results for root causes of issues

Picture 2. Total Data Quality Management (TDQM) cycle, with original names presented inside the pieces of pie but with explanations added from data quality monitoring point of view.

Much as the cyclic approach above implies, the working hypothesis for data quality professionals is that it is not a one-time effort – it is not viable that once data have been harmonized and data quality processes tuned, the data will stay of good quality forever. To achieve sustainable and effective results from data quality improvements, data quality needs to be continuously monitored and reported. It is also commonly acknowledged among data quality professionals that data quality issues are often noticed too late, once damage has already been done. (e.g. Informatica 2014; Uniserv 2014.) Data quality monitoring offers a way to make sure that the issues are noticed before having an effect on business, and that they do not become issues again – as long as one knows where to start looking for critical data quality issues.

Regardless of if an organization has a new data quality management plan, has just redesigned its processes to improve data quality or has no data quality activities in their past or present, they can monitor their data for quality. They can either use it to see how effective the new plan/processes are or use it to identify issues in business-critical data that would be beneficial to correct/investigate on.

The benefit of data quality monitoring is quite clear: an organization will have good quality data and all the benefits (or lack of costs) that come along with that.

3.2.1 Data Quality Dimensions

Data quality dimensions are characteristics or aspects of data that can be used to evaluate their quality. Dimensions are a fundamental part of any data quality approach or initiative, as they enable measurement and comparison of data to quality targets and requirements. They are therefore something that should and naturally will be considered when setting up data quality monitoring, consequently after it has been decided what data elements need to be monitored. Some dimensions are easier to measure objectively than others. Dimensions are qualitative (unmeasurable, no value can be assigned to them) in nature until combined with one or more metrics and associated measurement methods, when they become quantifiable (measurable). (Batini & Scannapieco 2006, 19.) If data quality metrics are aligned with a data-driven business strategy, this will provide the traditionally missing link between data quality and business performance (Harris 2011).

Due to the changing nature of data (unstructured, semi-structured), the domain-specificity and evolving technologies and requirements, there is no universally agreed set of data quality dimensions (Batini & Scannapieco 2006, 49). One of the classifications of data/information quality dimensions (MIT Total Data Quality Management presented in Pelkonen 2006, appendix 3) is shown in table 2.

Natural criteria	Data quality can be evaluated by the extent by which….
Free-of-Error	Data are correct and reliable.
Objectivity	Data are unbiased, unprejudiced and impartial.
Believability	Data are regarded as true and credible.
Reputation	Data are highly regarded in terms of its source and content.
Contextual criteria	Data quality can be evaluated by the extent by which….
Relevancy	Data are applicable and helpful for the task at hand.
Value-added	Data are beneficial and provide advantages from their use.
Completeness	Data are not missing and are of sufficient breadth and depth for the task at hand.
Appropriate amount of data	The volume of data is appropriate for the task at hand.
Ease of manipulation	Data are easy to manipulate and apply to different tasks.
Representation criteria	Data quality can be evaluated by the extent by which….
Interpretability	Data are in appropriate languages, symbols and units, and the definitions are clear.
Understandability	Data are easily comprehendible.
Concise representation	Data are compactly presented.
Consistent representation	Data are presented in the same format.
Availability criteria	Data quality can be evaluated by the extent by which….
Accessibility	Data are available or easily and quickly retrievable.
Security	Access to data is restricted appropriately to maintain its security.

Table 2. MIT TDQM information quality criteria (Pelkonen 2006, appendix 3)-

Determining which data quality dimensions to focus on is a next step from the results of this thesis. It is first imperative to define what subset of data is to be the target of monitoring, before it is explicitly decided which dimensions or characteristics of those data should be measured and how. It is however important for the reader to understand the dimension-aspect of data quality as it is central to any data quality activities, including monitoring.

3.3 Data Quality and Business Processes

Data plays an integral part of all business processes across an organization. Often the data in these processes are viewed as by-products of the processes of buying, manufacturing or selling the services or items that are conventionally thought of as products, instead of as products in their own right. Data are now often approached in the same way as any other products of an organization: data are designed according to customer needs, they are manufactured using pre-defined processes or bought from a supplier and then re-sold or given to (internal or external) customers to use (e.g. Batini & Scannapieco 2006, 61). This data process should be considered as a supporting process expanding across the whole of an organization's process structure.

An organization's business processes can be divided into operational, supporting and managerial processes, all consisting of several sub-processes or activities (Wikipedia 2014). These processes form three layers of the value proposition for the organization. The operational processes involve everyday operations of an organization and have a direct effect on parties external to the organization, e.g. the sales process that involves many of the operative functions and directly involves the customer. Supporting processes on the other hand are (usually) internal to the organization and have an indirect effect on external parties of the organization, e.g. data maintenance process that ensures that the data are updated and available at critical times, which affects customers through many channels and across functions, such as efficient and accurate sales and marketing. Managerial processes have an analytical nuance and involve strategic decision-making, affecting external parties in the long run. (Partly from Butel et al. 2005.)

Very often, the everyday business takes a vast amount of time to take care of and as it is the part of business that directly affects the customers, an organization's main source of income, it is also commonly thought of as the most valued part of business. While it is true that a business would not succeed without its customers, and that relationships with them should be valued, the value of supporting processes should not be underestimated. They facilitate the organization to run its business effectively and to focus its long-term plans and strategies correctly and profitably. These are also the processes that enable an organization to exist in ever more competitive and demanding markets. (Partly from Butel et al. 2005.) Unnoticed by many organizations, all processes not only use but rely on data, and moreover, on data of good quality. Many operational issues lead back to data, a fact that is often overlooked.

A pre-requisite for being able to recognise the datasets critical to an organization, to implement any data quality activities successfully, is identifying the key business processes, i.e. those processes and their activities that are critical to the organization's running and profitability. (Adapted from many data quality methodologies described in e.g. Aiken, P. 8.4.2014; Batini & Scannapieco 2006; Eppler 2006; Funk et al. 2006.) These processes and activities should be the starting point for focusing the targeting of data quality activities.

3.3.1 Main Business Domains of an Organization

The concept of a business process can be simplified as "an inter-linked, often logically sequenced set of work activities which translate inputs into outputs in order to deliver something of value for the business and/or the customer" (Failte Ireland 2013, 5). Key business processes are those processes that "have maximum impact on the success of an organization – real value-creating processes that customers and stakeholders are concerned with" (Alagse Consulting 2014).

Recognising an organization's key processes is always a subjective matter because it is shaped by an organization's individual environment and business. It is vital to know your business well to make the processes work in alignment with the organizational strategy, including making its data work for the organization towards its strategic goals. (Partly adapted from Failte Ireland 2013, 3-7; Aiken, P. 8.4.2014.)

Identifying key processes can be done in many ways, one is detailed below:

1. Identify critical success factors (CSF's) for achieving business objectives
2. Identify key performance indicators (KPI's), the metrics to measure CSF's

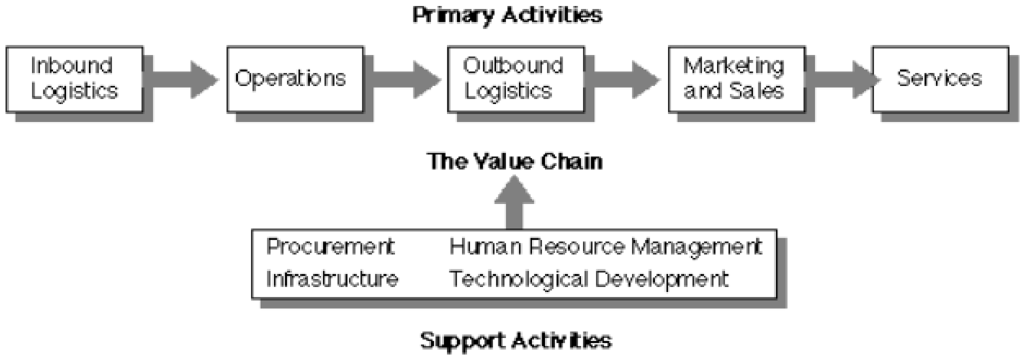
3. Identify processes that deliver above CSF's or KPI's
4. Group or ungroup related or un-related activities so that they describe the activities that get done.

Those groups of activities are the key business processes of the organization. (Alagse Consulting 2014.)

Although it was mentioned before that each organization has their own unique environment, a generic set of key business processes should not be too challenging to form as regardless of an organization's line of business, it will always have certain core functions that form the foundation that the business is built on. These are the basic activities and processes that create value to the organization. For most organizations, these activities are essentially the same, although each industry and environment has different emphases and might require slight alterations to the standard or fit into the model differently. (Butel et al. 2005, 39.)

An often used, textbook model for determining the business processes of an organization is the value chain model introduced by Michael Porter in 1985. The model recognises an organization's functions (=value activities) to consist of primary and support activities. It divides a company into strategically relevant main activities, that all can further divide into several separate activities that vary according to industry. (Porter 1985, 33-43.)

The needs of today's world might require some adjustments and additions to Porter's model, but in essence, the gist is still the same. The assumption of this thesis is that although almost 30 years old, Porter's model is still a valid model for the very core elements of an organization's structure. The same way as the master data are the foundation of organizational data, the activities in Porter's model are still the foundation for an organization's business. The more modern division of processes into operational, supporting and managerial introduced in the previous chapter goes together with Porter's model: the operational processes fall under primary activities and supporting and managerial processes under supporting activities.



Picture 3. Michael Porter's value chain from 1985 (University of Cambridge Institute for Manufacturing 2014).

According to Porter (1985, 39-41), the primary activities of an organization are five:

1. **"Inbound logistics:** Activities associated with receiving, storing and disseminating input to the products, such as material handling, warehousing, inventory control, vehicle scheduling, and returns to supplier.
2. **Operations:** Activities associated with transforming inputs into the final product form, such as machining, packaging, assembly, equipment maintenance, testing, etc.
3. **Outbound logistics:** Activities associated with collecting, storing, and physically distributing the product to buyers, such as finished goods warehousing, material handling, delivery vehicle operation, order processing and scheduling.
4. **Marketing and sales:** Activities associated with providing a means by which buyers can purchase the product and inducing them to do so, such as advertising, promotion, sales force, quoting, channel selection, channel relations and pricing.
5. **Service:** Activities associated with providing service to enhance or maintain the value of the product, such as installation, repair, training, parts supply, and product adjustment."

The support activities Porter (1985, 41-43) divided into four categories / main activities:

1. **Procurement:** Activities associated with purchasing inputs to be used in the companies' primary activities, not the purchased goods themselves. Procurement is also present in all of the support activities, such as purchasing outsourced technology or legal support. It is for this tendency to be present in all other organizations' activities that Porter sees procurement as a support, and not a primary activity.
2. **Technology development:** Technology in the value chain context means everything from know-how, to procedures to physical equipment, and is therefore also involved in every activity of an organization. The variety of technologies in a company is wide,

the complexity high and the technology structure often contains many levels. Technology development signifies a variety of activities that aim to improve the product and the process, and is more commonly known in many companies as research and development, a term which has a narrower associated breadth.

3. **Human resource management:** Activities associated with recruiting, hiring, training, developing and compensation of personnel. Again, this activity is present across the primary and support activities, in different forms, either serving a single activity or the entire spectrum of activities.
4. **Firm infrastructure:** Activities associated with the management, planning, finance, accounting, legal affairs, governmental affairs and quality management of an organization.

A business domain in this thesis is used to describe a business process or group of processes that together perform a function in an organization. Business domains can include any of Porter's primary or support activities as well as their sub-activities. Each domain consists of its own input-transformation-output processes (Adapted from process definition from Butel et al. 2005, 39).

3.3.2 Data across Business Domains

Historically, organizations have been operating in silos, each main function running on its own, without much cross-functional interaction (apart from possibly shared support activities), much as Porter's value chain model from 30 years ago implies. Those same functions still exist today, however, an organization is more and more expected to give a lifecycle service from first customer contact to delivery of service or product to after sales support. To achieve the seamless service required, a connecting thread needs to run through the activities described in the previous chapter.

The need to de-silo organizations shows especially in their growing data quality improvement requirements - data that were created for one purpose only, all of a sudden should serve the usages of all functions across the lifecycle of the product or customer relationship. Data already are not of good quality as they are not "fit for their intended use". Additionally, due to this functions will add their own interpretation of a data element when the existing data are not suitable for their use, or the data might be exactly right for all purposes but existing data might not be checked, and another copy of the same data might be created. This might be repeated several times over the lifecycle of the data. Obvious quality issues, such as duplication (same real world object is represented by more than one record in a database), inconsistency (e.g. same data value used for contradictory purposes) and many more, arise.

Most businesses still have a very functional division in their system architecture where different functions have their own systems (or operate the same system but separate modules) that, in the worst case scenario, have no integration between them. Even if there is an integration, data in the different systems/modules can be used very differently and have different maintenance policies. Data transformations in integrations are one of the critical points for data quality, especially on the semantic level (Peltonen 2006, 27).

To grow internal competence for creation of a more cross-silo culture is a challenge for organizations (Gulati 2007). The same applies for the push to have cross-silo data of good quality in an organization. “People tend to see only the data that is in front of them. There is little cooperation across boundaries. [Organizations should] achieve a more complete picture and facilitate cross-boundary communications.” (Aiken, P. 8.4.2014.) Aligned with the de-siloed approach now often used, organizational data should also be able to support all different functional activities. The very definition of master data, that many, if not all, master data are shared by different processes and activities, most likely existing in two or more cooperative systems (e.g. Kolehmainen 2011), supports the importance of cross-function, cross-system data.

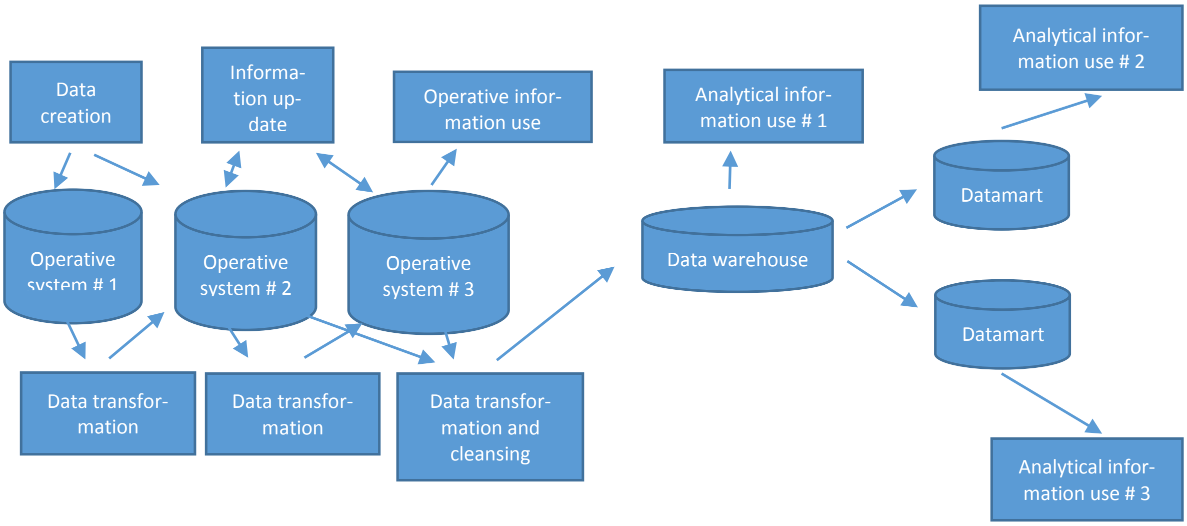
All organizational data need to work together to be able to serve all of an organization’s individual activities’ requirements efficiently, enable the lifecycle approach to products and support the organization as a whole to function profitably, making sure the customer is satisfied. To be able to streamline an organization’s data lifecycle, and to achieve good quality data, it is important to understand both the key business domains (as discussed in the previous chapter) and the key data domains of the organization (discussed in the next chapter). Logically, and based on the 80/20 rule of data quality issues mentioned earlier, the subset of critical data for data quality should be found by combining and analysing the data needs of those identified key business and data domains.

3.3.3 Main Data Domains and Flows in an Organization

Regardless of the business an organization is in, there are several data domains or objects that are nearly always present. For this reason, they are often called master data. Hence, these data are also the main data domains of an organization. As mentioned before, these master data are "the critical nouns of a business and fall generally into four groupings: people, things, places, and concepts. Further categorizations within those groupings are called subject areas, domain areas, or entity types. For example, within people, there are customer, employee, and salesperson. Within things, there are product, part, store, and asset. Within concepts, there are things like contract, warrantee, and licenses. Finally, within places, there are office locations and geographic divisions." (Haselden & Wolter 2006.)

An approach that can be used to assess those data domains' criticality for an organization is to evaluate each domain's volume in the business data, static or transactional. An organization that has three customers might not consider customer as important data, as opposed to a company that has 1000 customers. (Haselden & Wolter 2006.) However, if those few customers for the organization mentioned first, made 100 000 sales orders a year, their importance might increase to a higher level than for the company that has 1000 customers but each of them are one-time customers.

An indicator to the importance of data in an organization is also where the data flows within the organization and what are the processes involved in their lifecycle. English (1999, 160-161) discussed a model called the information value (and cost) chain whose objective is "to determine all business processes and applications, and all who create or update a group of data along with the process dependencies". The model is more commonly known as simply the information value chain and according to the International Association for Information and Data Quality (2014), it consists of "the end-to-end processes and data stores (…) involved in creating, updating, interfacing and propagating data of a specific type from its origination to its ultimate data store –". This is a model based on Porter's value chain idea, and although there are many representations of this model that resemble the traditional value chain introduced earlier, in practise, the model closely resembles a system architecture map where the information stakeholder's role and context of use are considered (an example is shown below). It models the data entry, manipulation and transformation into information along the value chain. Systems and use can be divided into operational and analytical due to the different nature of requirements and processes to data. (Pelkonen 2006, 18.)



Picture 4. A pragmatic representation of the information value chain (Pelkonen 2006, 18).

This model is very useful in data quality activities involved in assessing data quality and analysing the causes for poor data quality. Understanding the information value chain is key in identifying and analysing cross-domain data - either integrated or separately operated - that are very likely a critical source of data quality issues. Every instance of a data element that represents the same real-life object, whether separately across systems or as a duplicate within a system, should be assessed/monitored, to be able to perform a full, organization-wide analysis on the data element's quality and to recognise where improvements are needed to help make data better serve the business.

4 Empirical Research

The previous chapter covered the basics of data quality and data quality monitoring as well as giving an overview of what could be considered the main business processes and data domains that would be the home for the most critical subset of data for a business to monitor for data quality.

This chapter takes the theory to the field, to research topic experts from different industries, to address the issues of data quality through pain points of actual businesses – data quality issues critical to the achievement of their business objectives.

4.1 Overview of Field Study

The field study for this thesis was conducted during the summer of 2014. The study involved a series of one-to-one interviews conducted either at interviewees' current place of work or at the commissioning party office.

Preliminarily during the spring of 2014, a group of 17 interviewee candidates were listed with the help of the commissioning party – experts and enthusiasts in data and data quality fields with years of experience from different organizations in varied industries. Out of the 17 candidates, 13 persons were contacted first to see if they were willing to take part in the interviews. Altogether 10 persons agreed, and were then sent a short overview email of thesis and interview contents and the interview schedule in order to confirm the interview time and location.

Out of the 10 interviews that were agreed, 9 interviews were conducted. One interviewee had to cancel due to personal reasons. In these 9 interviews, altogether 10 persons were interviewed (one interview was with two persons due to main interviewee's request to have a second person present).

The interviews conducted were thematic: the themes of the interview were decided in advance, and further, specifying questions were asked during the interviews (HAAGA-HELIA University of Applied Sciences Thesis Guidelines Working Group 2014, 26). Interviewees were sent an introductory e-mail about the thesis topic and interview in mid-June 2014 (1-3 weeks before their respective interview). The e-mail sent can be found as attachment 1 of this report. For the interview a PowerPoint presentation was prepared and presented to interviewees as a basis for discussion. The presentation contained a few quotations from thesis introduction and background chapters about thesis objectives, an explanation on the purpose of the interview, a preliminary matrix with some example data elements to give an example of the many points of view that could be used for assessing the criticality of a particular data element, as well as a further filled-in version of the matrix to be shown at the end of the interview to possibly spark some further discussion. The further developed preliminary matrix is presented in the next chapter and the matrix along with some of the slides used as basis for discussion in the interview can be viewed in attachment 2.

Interviewees were asked to provide input on what they think are the most important data and business domains, then elaborate on the critical data elements for data quality monitoring from those data and business domains' points of view. This was already asked from them in the e-mail sent earlier to give them time to prepare the answers (attachment 1). In the interview, they were then also asked to discuss/verify the further developed version on the data quality monitoring target matrix.

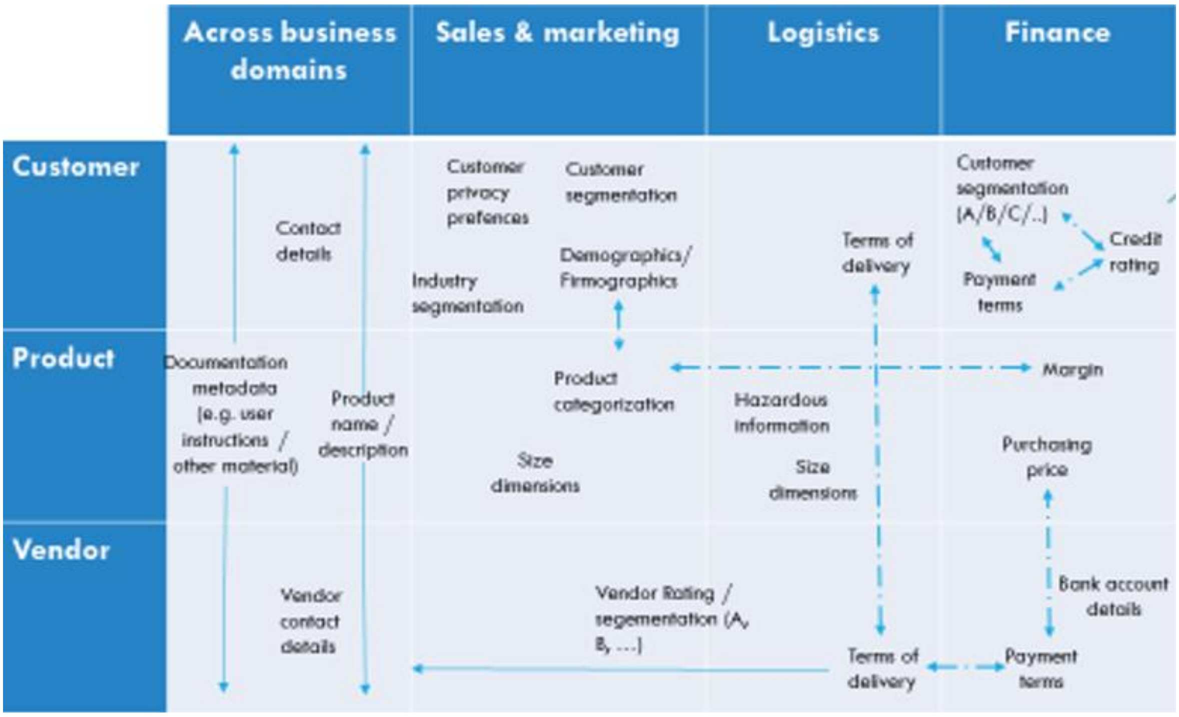
Interviewees were asked to discuss the interview topic based on their past and current experiences, thus the interview answers are not specific to any particular organization or even industry for most of the interviewees. Interviewees will remain anonymous as no added value is to be had from naming them. Interview results are to be interpreted as subjective as data quality is a subjective matter and interviewees are speaking from their own experience and knowledge. No absolute truth is intended to be found but knowledgeable conclusions can be drawn from analysing the interview results based on the years of accumulated experience in data and data quality management of the interviewees.

As the interviews were quite informal and the prioritization within data elements given was not always clearly indicated, at the end of the interview cycle, all received suggestions for the subset of data to be monitored were collected and e-mailed to all interviewees to rate between 0-5 (0 being "not important at all" and 5 being "critical"). This was to further help the writer to make conclusions on what elements can truly be stated as generic for most industries. Unfortunately the response to this e-mail verification round was quite poor: only three out of the ten interviewed persons replied. These replies were therefore not taken into consideration for the conclusions of this report.

4.2 Preliminary Targeting Suggestion for Data Quality Monitoring

As mentioned in the previous chapter, a preliminary matrix for the data quality monitoring target was put together for the interviews. This was used as a basis for the examples shown to interviewees at the beginning of the interview, and then shown to them fully at the end of the interview for verification/creating more discussion.

The matrix was created based on analysis of the theoretical part of this thesis as well as on discussions with the commissioning party's representatives that have years of experience with data quality issues from varied companies across industries. No formal interviews were held with representatives but open discussion, analysis of existing personal materials and email questions were used for data collection during this phase. The matrix, as presented to interviewees can be seen below, with explanations of the logic underneath.



Picture 5. Preliminary matrix for data quality monitoring target

The matrix was chosen as the representation format for the data quality monitoring target as it is a clear way of presenting the relationship between different data domains, business domains and the data elements within those.

The term data domain (vertical dark blue column in picture 5) is used to describe the data objects or areas that exist in a business data environment. For the matrix the three data domains that were considered the most commonly occurring and value-adding in key business activities were chosen: customer, product and vendor. Nearly all if not all business transactions/activities involve one of two external stakeholders: customer or vendor, without whom an organization would not be able to buy or sell goods, therefore, crippling the whole running of the business. That having been said, without product, there would not exist a business as what is a business that does not have anything to sell or promote? Product is used as an umbrella term for all materials bought, products produced and products/services sold in an organization. All of the three data domains also have a clear structure and exist in several if not all business information systems. They therefore fulfil the criteria for structured master data that was defined as the scope of this study in the theoretical part of this report.

The term business domain (horizontal dark blue column in picture 5) is used to describe the business functions, processes or sub-processes that exist in a business. The decision what business domains to include in the matrix was not as easy as for the data domains. Organizations have many key processes and the importance of the processes can greatly vary according to industry and size of business. After analysis of Porter's value chain and the key business process determination guidelines presented in the theoretical part of this report, the three business domains chosen for the matrix were sales & marketing, logistics and finance. The domains are closely related to the data domains chosen for the the matrix, partly exist in Porter's value chain model and are supported by the experiences of the commissioning party representatives. Sales & marketing was a clear winner for being part of the matrix due to the close connection to both customer and product data domains, involving the main value-adding objects in a business. In the Porter's value chain model inbound and outbound logistics were considered separate activities, but for the matrix these were combined as logistics due to the modern day way of thinking of logistics more as a chain rather than separate links. Finance was seen as a key function as it involves the cash flow and profitability calculations of a business, and although a support activity as such (not adding value on its own), has many (costly) data quality issues according to commissioning party's experience.

As the concept of master data implies, the data elements are mostly ones that are cross-domain, existing in more than one business function or more than one data object. For this purpose an ”across business domains” column was also added to the matrix in the end, covering data elements that cannot be pinpointed to be significant for one or two of the business domains only, but are important for all three. The discussions with commissioning party representatives revealed some single-domain data elements that are critical for data quality monitoring as well. However, a lot of the interest in data quality today is born out of multi-system environments being complex and the need to have this complexity managed to enable e.g. digitization. It was therefore a highly interesting issue to potentially be addressed by the interviews to see if any single-domain elements were raised as critical by the interviewees.

The dotted line arrows between some of the data elements in the matrix represent the need to monitor those data elements together to achieve greater benefit. For example, the vendor terms of delivery can directly affect the terms of delivery that can be given to customer in the logistics chain. Therefore, by monitoring the relationship between these two, possibly even in connection to a particular product, can reveal potential risks to customer delivery times leading to costly issues and customer dissatisfaction. In finance, the purchasing price and payment term relationship can be significant in cash flow management. For small purchasing prices it is most likely fine to have a short payment term but for bigger amounts, a longer payment term is required. At the same time, payment terms on their own are a valid target for monitoring, especially in a multi-system or multi-domain environment where the same vendor/customer could incorrectly have different payment terms across systems/domains.

4.3 Field Study Interview Summaries

The interviewees were not directed too much during the interview, therefore if other than master data were discussed (e.g. transactional data), this was not corrected due to not wanting to influence the train of thought of interviewees. If the interviewee started talking purely of data governance or other master data disciplines, the discussion was attempted to steer towards the thesis topic again.

The interviews have been summarized with unfiltered descriptions of what interviewees said, i.e. reference to for example transactional data might appear. There might also be inconsistency in interview summary for the same reason. The relevance of the interview results for the thesis scope will be assessed and the angle to be used will be chosen in the conclusions part of this report. As interview results differed quite much for each interview, it was decided that each interview would be summarized briefly on its own. If enough data elements were highlighted in the interview, a table summarizing the data elements is presented for the interview. A short description of the data elements is given if the definition is somewhat different to how the element could be generally perceived or how the other interviewees described and understood the element.

Short background descriptions of the interviewees can be found as attachment 3 of this report. This is to demonstrate that although the sample of persons interviewed was fairly small, there is a lot of expertise and experience held by them.

4.3.1 Interview 1

Interviewee stated the method for determining the significance of particular domains and elements for data quality monitoring to be tying the activity to a company's strategic goals. The data quality monitoring targets should support the overall strategic targets of the organization, and once those data elements have been identified that are critical to the strategic success, has the subset of data that should be monitored for data quality been found as well. The next step would then be to determine which of those data elements have data quality issues for the organization in question.

Another way of finding data quality anomalies in his opinion was to recognise the structure of data objects and through that finding the possible places where anomalies can exist, e.g. if a customer has a number of branches who all have their own instance in a system's database, and each of those branches has a contact person, there is a good chance that the same contact person is used for more than one branch. There should therefore be monitoring to see that all of the contact person's data is consistent, and has for example not been updated only for one branch.

Interviewee agreed with the three data domains specified in the example matrix - customer, product and vendor – specifying that those domains should be considered from a wide perspective, i.e. a customer could mean e.g. an applicant if the organization was a government social security institution or a student if the organization was a school or a university; product could mean not only physical items but also e.g. services or application forms (in the cases of the previous examples of applicant or student).

For the business domains the interviewee had a somewhat differing opinion from the example provided. He named the three most important business domains to be sales & marketing, processing/production and service architecture. The last business domain was particularly relevant to the public sector that the interviewee had most experience of. With processing/production the interviewee wanted to set on the same line the concepts of e.g. application processing in one organization to the more traditional assembly line production processes in another. He also mentioned that the finance business domain mentioned in the preliminary matrix was also important e.g. in the banking industry, although it would be more specifically named, e.g. controlling, risk management, etc.

On approaching the subject of data elements within the data and business domains, the interviewee emphasized the concept of consistency. By this he meant that those data elements that were present in several systems should be consistent, and therefore a key target for data quality monitoring.

The interviewee named the main data elements that were in his opinion key to data quality. Most of them aligned with what was also mentioned in the example matrix. Additional data elements that interviewee brought up were identification data (social security number or similar for individual persons, business ID/address for businesses, and different standardization codes for product, e.g. EAN) and contact person data (the issues concerning them existing in the first place, and their correctness). He added that contact person data could be seen as belonging to the umbrella element "contact details" that was specified in the exemplary matrix as well. The interviewee mentioned that metadata was also important not only for documentation as specified on the preliminary matrix but for e.g. product pictures and product descriptions.

The interviewee emphasized again the importance of the breadth of perspective also with the data elements, e.g. the benefit payments for social security institution's passed applications can be considered the same for the public sector as prices are for physical products in another industry, the different types of application can be comparable to product categorizations, etc. All the names of the domains and elements simply need to be adjusted to fit the terminology of the organization running the data quality monitoring.

4.3.2 Interview 2

The interviewee strongly believes that most data have several use cases and are relevant throughout processes/functions, and that most data's quality will be important to more than only one business domain. It is very unlikely to find critical data for data quality monitoring that is only relevant and critical for one process for example.

The key is to document the rules for the monitoring well: It is important to take into consideration what industry is in question and from which process's point of view the data are looked at. In the interviewee's opinion, there will most likely be a very small subset of data that can be the outcome of this study due to the fact that different industries can have very different emphases on critical data definitions.

Interviewee agreed with the three data domains in the exemplary matrix, however, she emphasized the need to understand the complexity of the data domains. She explained this by pointing out that e.g. a customer can be a business (B2B) customer, a personal (B2C) customer or a common term such as business partner can be used to identify both customers and vendors if there is no importance for a particular organization which one the partner is. For product, the data requirements differ and the emphases shift depending on if it is a physical product or a service product, what stage of the product lifecycle the product is in, etc.

For the business domains, the interviewee nominated sales & marketing, purchasing and logistics. Those business domains are the main functions of any business's activities. She emphasized that for example for sales, product data requirements and criticality vary greatly depending on the sales channel: sales in a shop are very different to sales online where customer has no physical contact with the product. In the latter case, the product data become hugely important as the decision to buy the product is based on the product information – and it needs to be of good quality.

The below table shows the summary of the data elements that came up in this interview as critical. Further elaboration on some of the data elements are given in the paragraphs after the table.

	Sales & marketing	Purchasing	Logistics	Cross-business domain
Customer	- Customer segmentation/typing - Bank account details - Contact details - Authorized persons - Allowed payment methods - Payment terms	-	- Contact details	-
Vendor	-	- Bank account details - Payment terms - Contact details	- Contact details	-
Product	- Product categorization/typing - Name/description - Size/dimensions - Return regulations	- Purchasing price	- Product categorization - Size/dimensions - Hazardous information	-
Cross-data domain	-	-	-	- Identification data

Table 3. Data elements brought up in interview 2.

Interviewee mentioned that product categorization is very critical in finance due to connecting the income from sold products to the correct financial structures and therefore steering the cash flow correctly, but interviewee did not raise finance as a critical business domain. In sales, product categorization is important especially in online shops where products are located and searched based on category. In logistics the categorization/typing can e.g. steer which warehouse the product is ordered to.

Bank account details were brought up by interviewee and could fall under the finance business domain, as they are used for making payments to customers. However, in this case, the payments made to the bank account come from bonus points collected during sales transactions, placing the data element under sales & marketing. The other use case for bank account details brought up by interviewee was to return money for any sales orders that were either not delivered due to corporation's own fault or returned by customer. In these cases the data element could fall under either sales & marketing or finance as well. Bank account details are also relevant for vendor under purchasing/finance. Payment terms in the above table could also fall under finance or sales/purchasing, the latter needing the term details when making sales deals or purchasing goods.

Contact details can contain several different data elements within the same umbrella term: postal address (visiting, delivery and mailing addresses), billing address, electronic contact details such as e-mail, telephone number (for text messages as well as calls), etc. Even the bonus card details for the retail corporation the interviewee currently works for, are considered contact details of a sort.

Customer segmentation is very relevant according to the interviewee. With customer segmentation as well, there are different types. The interviewee mentioned stable and volatile segmentation based on e.g. demographics that stay fairly stable and segmentation based on sales or browsing history that can change fairly frequently. Finance also segments customers based on their credit rating and payment history.

Size and dimensions of a product are a highly critical data element touching many business domains and in many different ways. The interviewee mentioned at least six different use cases varying from warehouse management to mode of delivery to freight management to shop presentation design. This data element could easily be put in the cross-business domain cell but as it is not a highly relevant field for purchasing, it was put separately in the other two business domains.

Hazardous information was also seen as important by interviewee due to legal requirements, e.g. flammable products need to be marked, there is a legal requirement to be able to report all flammable products held in warehouse at any given time and there are regulations to how many flammable products can be showcased at any one time in a shop. The rules for delivering and returning of hazardous products are also different.

Identification data was mentioned by interviewee to be relevant and critical for all data domains across all business domains. Identification data elements can be various, depending on the company.

Sales pricing data are very critical pieces of data for all organizations. Interviewee does not however think it they are master data due to their dynamic nature, and adds that they would not be easy to monitor either. The purchasing price, on the other hand, although also possibly changing at some intervals, could possibly be considered as master data/monitoring candidate due to the fact that they are used as a basis for other data (e.g. sales pricing calculations).

4.3.3 Interview 3

This interview exceptionally had two interviewees due to a request from the main interviewee to have a member of the Global Master Data team present in the interview for a broader view on the thesis topic.

The interviewees considered the ways of identifying the importance of certain data elements over others, and concluded that transaction volumes would be a good way to do this based on fact. This would involve analysing of transactions to see which data and data elements were used most often in the transactions of a business. Also, when monitoring the data for quality, the target set of data that should be monitored should be scoped so that time would not be "wasted" on monitoring e.g. customers that bought from the business once three years ago. Instead, only e.g. the top X number of customers that provide 90% of sales together with the best-selling products, should be concentrated on for most benefit from both finding and correcting data quality errors and being able to discover the root causes for the data quality errors. Using this top-selling method in conjunction with the subset of critical data discovered by this thesis project, would be a solid effort in discovering and fixing a company's data quality issues now and in the future. If it is known that the most important data domains are e.g. customer, vendor and product and the most important business domains are e.g. sales & marketing, procurement and logistics, the top 100 customers and their transactions can then be looked at to locate the most used data elements/fields for a particular business domain, e.g. sales.

The interviewees also commented on the reasoning of why the data elements they nominated as the critical ones are so important. Data quality issues in these elements stop processes from running, cause an increase in reclamations, cause a decrease in customer image of company, show in accounting (profitability, cash flow, balance sheet, cost management, etc.) and make the company vulnerable when opening data up to external stakeholders.

For the data domains, the interviewees had quite a few suggestions. These were

- Customer
- Product
- Pricing
- Lifecycle (consideration of time and dynamism)
- Production Data
- Regulatory Requirements (tax authorities, customs, health & safety authorities, recycling authorities, etc.)

When approaching the subject of business domains, the interviewees brought up the process point of view. They explained that if it were only functions that were looked at, the situation might be quite simple with data quality. However, with so many processes crossing functional borders, the situation gets complicated due to different requirements for data quality. Those requirements are important to understand, e.g. regulatory requirements that warrant sanctions if not complied with, the cost of goods sold, etc.

They would add production to the business domains already present in the example matrix. They wanted to replace finance with that business domain as they were slightly questioning the need for finance as a business domain as finance usually carries simply the consequences of what happens in the production – logistics – sales processes. They would also replace the "logistics" business domain from the exemplary matrix with "supply chain management".

The interviewees admitted that it is usually the data elements that are considered connected to or affect finance that are emphasized. However, those data elements can usually be placed under other business domains. The data elements that the interviewees nominated for the data quality monitoring subset are listed in the table below. The interviewees mentioned three data domains for which no data elements were mentioned.

	Sales & marketing	Production	Supply Chain Management	Cross-business domain
Customer	- Payment terms - Customer segmentation	-	- Customer delivery routes - Delivery terms	- Name - Contact details
Product	- Product categorization/sub-categorization	- Country of origin - Subregion - Regulatory data	- Unit of measure + conversion factor - Regulatory data	- Name - Identification - Brand owner -
Pricing	- Sales pricing components - Discount components - Special pricing condition components	- Cost of production components	- Purchasing pricing components - Vendor bonus components - Vendor discount components - Tax data	-
Lifecycle	-	-	-	-

Production	-	-	-	-
Regulatory rq	-	-	-	-

Table 4. Data elements brought up in interview 3.

The regulatory requirements were mentioned as part of product data domain data elements. For the production business domain the requirements included recycling requirements, whereas for the supply chain management business domain these included allergy information, reporting requirements, etc. There are also accounting requirements that would be connected to finance through different processes (as all processes might have their own accounting requirements).

Instead of considering purchasing and sales prices as data elements, the interviewees said that it should be the different pricing components and pricing rules/formulas that are the used for data quality monitoring. The components are the data elements and the formulas would provide rules for monitoring the quality of the pricing data.

4.3.4 Interview 4

The interviewee agreed with the thesis hypothesis that it is not viable for an organization to aim for 100% data quality. Even if it were possible, there comes a point where it is simply not financially worth to keep improving the quality.

The interviewee stated that although the financial master data involved in external accounting should always be correct due to being a compliancy issue and to enable educated decision-making (and rarely in his experience are an issue), the data that directly involve a customer should be prioritized over internal accounting/controlling.

The interviewee's opinion is that the more complicated the business model of an organization is, the larger the volume of critical data for data quality is. The emphasis on what are critical data depends on the industry the organization is operating in. The interviewee has experience from two organizations selling very different types of products: packaged consumer goods and customized equipment manufactured from raw materials/semi-finished components. For the first case, the business model is fairly simple and the products are mass-produced without a need to identify a single product which makes the master data area somewhat simpler than for the latter case where for each product/equipment there is a need to have the product uniquely identifiable through a serial number whether mass-produced or not.

The interviewee sees master data to be born out of the information requirements of processes. Master data are needed to run processes, and if data are not of good quality, that interrupts running of processes. Master data and their quality also need to be tied to the management model of the organization: its reporting needs (management and external), its organizational structures, etc. In certain industries, the legal and regulatory requirements are also heavy influencers. Those could even be considered their own business domain.

Moving onto the exemplary matrix, the interviewee agreed with the data domains chosen. He added that the customer domain could even be divided to B2B and B2C domains due to different data requirements. For product, he said he agreed with it as long as it included all different types of materials as well: raw materials, semi-finished goods (produced from raw materials), components (bought semi-finished goods), finished products, packaging materials and marketing materials. These two data domains along with vendor belong to the supply chain, the operative heart of an organization. (Note by author: The supply chain means a combination of all different companies and stakeholders involved in a particular product during procurement, production, handling and distribution (Investopedia 2014).

With that thought, he would also change the logistics business domain to supply chain that he sees to include sourcing, production, warehousing, logistics and customer service (which is in the middle ground between sales and supply chain actually). He would change the finance business domain in the example matrix to controlling structures/managerial reporting. By this he means the financial structures such as charts of accounts, cost centres, etc. Finance is a function that utilizes data from other functions but the financial structures also need to be available for other functions to use. The information for running an organization is born from finance (management accounting), and so are the statements required by law (external accounting). Therefore those areas of finance are critical for data quality and the results of this study as well. He agrees that sales & marketing is one of the key business domain, and further specifies marketing in his opinion to also include product management.

Delving into data elements that are important for data quality monitoring, the interviewee brought up the following elements collected in the table below.

	Sales & Marketing	Supply Chain	Controlling Structures & Managerial Reporting	Regulatory Requirements	Cross-business domain
Customer	- Contact details - Credit rating - Credit limit - Payment terms - Minimum order quantity	- Contact details - Delivery details - Allowed packaging sizes for products	- Payer details - Customer structure	-	-
Product	- Metadata - Physical details - Pricing component/Prices - Product categorization	- Physical details	- Product categorization	-	- Traceability data - Quality, Health & Safety data - Hazardous information
Vendor	-	- Contact details - Pricing components / pricing - Delivery terms	-	-	-

		- Payment terms - Bank account details - Product list			
Cross-data domains	-	-	-	-	- Identification data

Table 5. Data elements brought up in interview 4.

Interviewee specified that identification data is the number one data element in all data domains. This is the key to not having duplicate records in the systems, it ties together with credit risk management and enables efficient supply chain management. Products can have several identifying data elements such as the EAN code (international article number), system-specific product number, customer-specific product number, GTIN number (Global Trade Item Number), etc.

The contact details for customer and vendor include not only the physical and electronic contact details but also the contact person for the customer or vendor, these are especially important for sales and sourcing. Delivery details for customer include the terms of delivery, delivery locations, the details of allowed delivery times, etc.

The product physical details are all factors to do with the physical being of the product in terms of sales and the supply chain. This includes:

- Packaging sizes
- Packaging variations (e.g. shelving unit that contains x amount of some product, retail box that contains x amount of shelving units, master box that contains x amount of retail boxes, etc.)
- Weight
- Dimensions

The interviewee does not think one can say with certainty that pricing is or is not master data. In his opinion it depends a lot on the organization's process management and type of products sold. For example, the prices of fast-moving consumer goods change more often than those of durable consumer goods. There are often annual list prices (both from vendor to organization and from organization to customer) that stay the same for a whole year, also depending on the industry, customer/vendor discounts can be given on an annual basis. Different kinds of fees charged, e.g. minimum order fee when customer orders under a certain amount of goods, are also stable. Campaign pricing is often then more volatile. If pricing is of good quality, a lot of costs can be decreased by e.g. using automatic payment of purchase invoices, etc. so it is a critical piece of data to monitor. The interviewee would add both pricing components and prices themselves into the matrix.

Interviewee adds that more important than the classification of pricing data as master data or not, is the stability of processes used to maintain pricing data. It is also vital to have process controlling structures (e.g. auditing) in place to prevent possible fraudulent actions from taking place. These kind of auditing data also qualify as a type of metadata.

Product categorization in the interviewee's opinion is also a financial factor connected to controlling structures. Behind the categorizations there are often structures that connect products to certain cost and profit centres, allocating the flow of cash from sales to correct accounts, etc. These allocations also affect the profitability calculations and sales reporting.

The regulatory requirements for a product are many according to interviewee's experience. Those can be divided in three:

- **Traceability:** Authorities need to know e.g. the country of origin of a material even though for the end product it makes no difference whatsoever. The serial number is a key to tracing products globally, but also the batch where the particular material or product came from needs to be known in case a problem occurs even years from sales, causing the need to locate all other materials or products from the same batch.
- **Quality, Health & Safety:** Authorities need to have it noted down that all necessary quality, health & safety guidelines have been followed.

- **Hazardous:** Authorities require for organizations to mark and handle hazardous materials in appropriate ways.

These regulatory requirements are not only followed in organizations because they are legally bound to do so, but making sure those regulations are fulfilled is also in the organizations' own best interest due to helping to avoid unnecessary risk of dangerous situations for customers, leading to potentially catastrophic consequences and complete loss of business.

4.3.5 Interview 5

The interviewee believes that the data that are worth monitoring for data quality are the ones that are shared by more than one function/process. Sometimes some single business domain data elements can be raised to the company-wide level but in general those data elements are only important within a business domain and not worthy of too much effort.

For the data domains, the interviewee agreed with product (without it you have nothing to sell) and customer (without customers, you will go down very quickly). With product she mentioned that the productization of services is still a difficult matter, and often the physical products are easier in the data sense than service products. She would replace the vendor data domain (or add a fourth data domain, preferably) with employee that would include not only internal personnel but all outsourced/temporary employees that work in the same premises, use the same systems, and to all extents look like a company's internal employee to external parties. The interviewee also mentioned that depending on the company, customer and vendor could simply be represented by the same master record, as long as the role in which the record acts in each process is identified.

Interviewee mentioned that employee data are becoming an increasingly critical area for many organizations who are losing a lot of money messing with simple matters' data such as who they employ and what rights those employees have within the organization. Especially public administration has been criticized for inefficiency, and the employee data quality has been raised to high priority to combat this. Employees can also directly influence the customer image of an organization if they are not given the correct guidelines to work with from the very first minute, not to mention being a major security risk if correct access rights are not given to the employee. The ability (and associated data) to audit the doings of an employee is also critical.

The business domain the interviewee would add to the example matrix is human resources (HR). When asked which business domain she would replace, it was a more difficult question that she could not answer as business domains in the matrix are all valid. Additionally, the interviewee questioned that the data in the preliminary matrix are operational data that are created naturally from business operations, but what about analytical data which are not necessarily created naturally by any function and are not needed by any function but which are still needed for reporting. As the goal for data quality and data management in general is to have all data correct from the time of creation, these data need to exist before reporting needs arise. The data might exist in customer master or product master, but knowing the reporting dimension connection enables the locating of these data. It is up to the organization, whose (which function's/process's) responsibility the maintenance of these "extra" data is. These data may not be critical for operational running of company but are critical for e.g. the management decision-making and profitability calculations which, in the long run, enable the business to flourish and grow. The interviewee does not really see reporting as a business domain but as a third dimension to the matrix.

In the end, the interviewee did not nominate to take any of the exemplary business domains out, simply added the HR and brought up the matter of reporting to be considered in the conclusions of this study. The data elements the interviewee nominated as critical for data quality monitoring are detailed below. A lot of the elements are critical for more than one business/data domain but as they are not critical or valid for all domains, they could not be placed under the "cross-business/cross-data domain" column/row.

	Sales & Marketing	Logistics	Finance / Reporting	Human Resources	Cross-business domain

Customer	- Name - Contact details - Customer preferences - Segmentation / demographics	- Name - Contact details - Customer preferences	- Name - Contact details - Segmentation	-	- Customer privacy
Product	- Name - Sales pricing component	- Name - (Purchasing pricing components)	-	-	- Regulatory requirements
Employee	-	-	- Cross-references to customer data	- Contact details - Access data - Role data - Skill & competency data - Regulatory requirements	-
Vendor	-	- Name - Contact details	- Name - Contact details - Cross-references to customer data	-	-
Cross-data domain	-	-	-	-	- Identification data

Table 6. Data elements brought up in interview 5.

Even though the interviewee recognises that some organizations do not necessarily force uniqueness of records - e.g. some organizations allow duplicate records for the same customer for process reasons - she strongly believes that a representation of a real-world object such as customer or product should only exist in a system once. If there is a need for e.g. multiple customer records for one legal entity in the real-world, there should be an “uber” or umbrella code connecting all of them or at the very least, mappings should exist between the different records that make it 100% clear which customer records are representing the same legal entity.

The interviewee thinks contact details are a wide set of data, and even include bank account details, especially for vendor. Also the eInvoice address is more and more important today.

Customer preferences contain all kinds of data about what the customer would or would not like, e.g. privacy preferences, subject preferences to help with targeted marketing, etc. Those data are very important and affect the image of an organization greatly – either in a positive or negative way depending on whether data about preferences are correct or not. Preferences might also include matters of rights, e.g. the right that a customer has to buy certain types of products or the right for the organization to hand over the customer’s address to external parties.

The employee access data refers to the data that grants or revokes an employee either physical or digital access to an organizations premises and resources. According to interviewee these data are ones that no one usually wants to take responsibility for but that are extremely critical in all organizations. Employee data that are important are also the details of the role the employee has in the organization, as well as the skills, competencies and permits the employee possesses.

The interviewee also emphasized the need for cross-referencing in cases where e.g. employee is also a customer or a vendor is also a customer. These are important in finance especially, where otherwise e.g. the organization could keep paying bills to the vendor, not realising that the same vendor as a customer is not paying their bills.

4.3.6 Interview 6

Interviewee considers scoping the most important part of any data management or data quality activity. One should not submit to the assumption that only a small portion of data can be fixed: if you know how to scope correctly, you can handle great amounts of data or at least the most critical subset of data. It is also key to recognise what are the data that are classed as critical or strategic. Criticality can be looked at from a few points of view, e.g. business-critical or system-critical. Further scoping can be done using rules. In interviewee's experience, data quality monitoring is usually started with management initiative of setting the strategic data for monitoring.

The significance of good quality data can be seen in smoothly running business processes, correctly running systems and the ability to take out proper reports. A way of considering what are the most critical data in an organization is to think of its operating model – if a company is operating globally for example, the key is to fix the data that are global first. Secondly, out of the global data, you look at the key processes and the global data that those processes need. The key processes are the ones that carry out an organization's strategy.

The interviewee sees it so that each data domain has a global and a local portion instead of "cross-domain" being its own data domain. For example, payment terms for a customer can be global data if agreements are global or local if agreements are local. For these kind of data elements where a list of allowed values is often the preferred way of managing data, he also emphasized that managing and monitoring of values in these lists is a very important data quality activity.

Going into discussion on the key data domains for data quality monitoring, the interviewee nominated five domains: product, customer, vendor, financial data and employee. For product, he mentioned the importance of the definition of product as the definition can be and is different depending on the point of view: for a production organization, the definition deals more with the designing and creation of products; for a sales organization, the definition is more commercial-minded and for sourcing, the emphasis is potentially on raw materials or components. The financial data would include such data elements as charts of account, cost centres, etc. depending on the organization.

For the business domains the interviewee again nominated five domains: sales & marketing, sourcing, supply chain management, finance and human resources. He added that basically sourcing is often viewed as part of supply chain management but in his organization sourcing deals with such volumes of raw material purchases that it is considered a business domain/function of its own. According to interviewee, human resources have become more and more important for many organizations: there is a need to know who is where, what they are doing, how much they are being paid, how they are supported in skill development, what the organization's future plans are for them, etc. In large groups of companies where human resources data are to be managed on a group-level this can be very tricky.

Within these domains, the interviewee brought up the data elements presented in the table below.

	Sales & marketing	Sourcing	Supply Chain Management	Finance	Human resources	Cross-business domain
Customer	- Contact details - Contact person - Name - Group details - Buying permits - Customer segmentation	-	- Contact details - Contact person - Name - Delivery terms & monitoring data	- Contact details - Contact person - Name - Group details	-	-
Product	- Sales pricing components / characteristics - Standards	- Purchasing pricing components - Standards	- Product line - Production details	- Product categorization	-	-

	- Product categorization					
Vendor	-	- Contact details - Delivery terms & monitoring data	-	- Contact details - Bank account	-	-
Financial data	-	-	-	- Charts of account - Profit centres - Cost centres	-	-
Employee	-	-	-	-	- Contact details - Bank account - Salary - Skills - Permits - Legal requirements	-
Cross-data domain	-	-	-	- Internal organizational details	-	- Identification data - Metadata - Quality monitoring data

Table 7. Data elements brought up in interview 6.

By customer group details the interviewee meant the information about if the customer company belongs to a group. Delivery terms & monitoring data refers to both the actual terms of delivery that define who is responsible for the delivery and possible faults, as well as the monitoring of deliveries, e.g. if a delivery is late for a customer then there needs to be a note in the customer data, etc. This is important also for the vendor data domain.

For product data domain, the interviewee specified that the data elements that are critical depend on the type of products in an organization, as there are a lot of variations. In the interview, he said he would try to bring up as varied a bunch of elements as possible, not concentrating only on one type of product. Production details in the product data domain refer to the data about the production process of the product that assists in e.g. optimizing the process. Product standards mean different identifying codes of a product, e.g. the EAN code. They are also included under the umbrella term identification data but interviewee wanted to emphasize the importance of these different standards. The type of standards that are important depend on the organization's line of business.

According to the interviewee, pricing components (both sales & purchasing) are definitely data elements to monitor for data quality. Those always involve formulas which provide a basis for the monitoring. There is no denying that pricing is a key element in any organization's data assets.

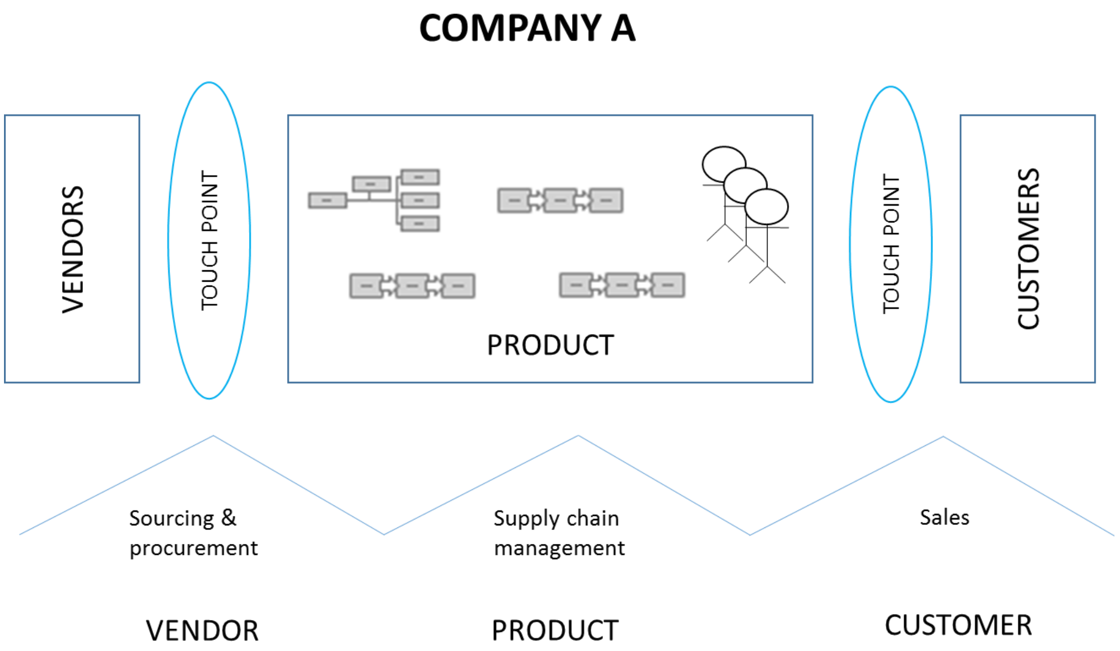
Interviewee sees data concerning the quality of different processes and objects to be critical to most organizations, especially where safety is an issue. Quality is a factor that should be recorded across business domains and data domains but the emphasis of the data varies based on the organization's industry. In the interviewee's organization, especially the physical quality of products and the quality of the production process are important. There are regulations determining the level to which these details need to be recorded (most likely they are stored in a system), and sanctions can be incurred if these data are missing.

The internal organizational details data element in the product/finance domains refers to the data regarding e.g. the responsible buyer, the responsible sales representative, persons making purchases or sales, etc. that affect the allocation of costs and income to correct organizational units or steer any queries to the correct people.

4.3.7 Interview 7

The interviewee considers the way to prioritize the most important data to be using key performance indicators (KPIs). KPIs are the measures that an organization uses to determine their success at reaching their strategic and operational goals. Each company has their own way of organizing and so the most important processes can vary a lot from business to business, therefore also the KPIs. Once the company-level KPIs are determined, the data quality KPIs that support those higher-level KPIs need to be determined. The master data that are critical to be monitored for data quality are those that support the data quality KPIs.

The interviewee agreed with the data domains in the exemplary matrix: customer, product and vendor. As for the business domains, he had a differing opinion. In his opinion finance is not one of the key business domains but is a domain that depends on the other business domains: It is only viable to build structures for finance once other business domains are in order. Finance expands all other business domains, having a role in all of them in one way or another. He nominated the business domains to be sourcing (external processes, a touch point from internal to external processes), supply chain management (internal processes and structures) and sales (external processes, a touch point from internal to external). Those are the key business domains that are common with most organizations but he also mentioned regulatory requirements (especially in the pharmaceuticals industry where he works currently) and human resources to be important. The interviewee drew a picture to illustrate his logic on the connection of the key data and business domains (picture 6 below). In the interviewee's opinion, the "cross-business domains" section could be presented as a separate list, and taken away from the actual matrix.



Picture 6. Connection of key data and business domains in an organization by interviewee 7.

The data elements the interviewee mentioned to be the most important within these domains are detailed in the table below.

	Sourcing	Supply chain management	Sales	Regulatory requirements
Customer	-	- Contact details	- Contact details - Customer type - Customer segmentation - Customer structures - Contractual terms	- Permit data
Product	- Product categorization	-	- Product categorization	- Characteristics
Vendor	- Contact details - Bank account details - Contractual terms - Vendor segmentation	-	-	-

Table 8. Data elements brought up in interview 7.

As per interviewee's suggestion, the cross-business domain data elements, presented separately below, are:

For customer / vendor:

- Identification data
- Name

For product:

- Identification data
- Units of measure
- Dimensions

The interviewee added that within the customer data domain, the identification data can be used not only for uniquely identifying the customer but also for e.g. customer relationship validity determination (based on valid VAT number).

Customer structures in the sales/customer domains of the matrix refer to different groupings of customers either externally or internally, e.g. according to which group of companies customer belongs to. The regulatory requirements differ depending on the data domain. For customer data domain, the regulatory requirements that are important to monitor are the permits the customer holds for buying certain types of products, e.g. hazardous products. For product, the regulatory requirements include many different kinds of characteristic data, e.g. hazardousness, storage requirements (storage & transportation temperature, sensitivity to light, etc.), classification (as e.g. a narcotic substance), etc.

Contractual terms in turn refer to any kind of terms agreed to by contract either with a customer or with a vendor, these can include e.g. payment terms, delivery terms, pricing (if contractual), etc.

The interviewee mentioned that the only pricing that he considers more stable (and possible to consider as master data) is the wholesale price which is determined by the authorities. He mentioned that this is at least true in his experience. Other pricing data are of course extremely important but not to be considered master data in his opinion.

In line with his view that finance expands all other business domains, he mentioned also that the finance-side structures such as charts of accounts, cost and profit centres, bank masters, etc. are very important but it would be difficult in his opinion to place them under any particular business domain.

4.3.8 Interview 8

The interviewee started by mentioning that the banking and insurance industry where he is in is very customer-centric. A lot of the data used are received directly from the customer and therefore, it is key that the processes are maintained correctly as there is not much else an organization can do in that kind of scenario to affect the quality of the data. With that in mind, customer data and its availability are extremely critical for the kind of service product companies that operate in his line of business. You need to know your customer and be able to use data about the customer seamlessly throughout your operations, even with different kinds of regulatory restrictions hindering the sharing of data sometimes also inside one organization.

The banking and insurance industry is very much becoming, and aims to be, more and more digital. According to the interviewee, one of the first steps in digitizing a process is to determine measures for monitoring e.g. data quality. In his opinion, the customer data are the trickiest to monitor as you need to get under the customer's skin, whereas product data are somewhat easier as the most of the complexity exists within the product structures.

The customer-centricity affected the data and business domains suggested by the interviewee greatly as well. He nominated two data domains only, customer and product/service. He also emphasized that his view on products will also be from the customer point of view, instead of e.g. product development or other aspects where he has not been involved with as much. In the future, he would see that vendor as an extension of the internal organization could become critical as organizations like his are running a more and more networked business.

The business domains the interviewee nominated were sales & marketing, finance/risk management and customer relationship management/customer service. He sees that sales & marketing is the business domain in which there is most potential for data quality monitoring and improvement as data here are key to making sure the organization is communicating the right things to the right customer. Previously, in his line of business, the focus has heavily been on risk management but this approach has proven to have a negative effect on sales & marketing.

The data elements mentioned by the interviewee are collected in the table below.

	Sales & marketing	Finance/Risk management	Customer relationship management/customer service	Cross-business domain
Customer	- Contact person - B2B customer classification - B2C characteristics - Products / services - Needs - Customer contact/visit - Contractual data - Marketing and privacy consents / preferences - Asset data - Value network	- Regulatory requirements - Risk management calculation components - B2B customer classification - B2C characteristics - Location data - Value network	- Products / services - Needs - Customer contact/visit - Contractual data - Bank account details - Location data	- Contact details - Customer economy - Customer segmentation - Identification data - Customer relationship status - Start and end date of customer relationship - Documentation metadata
Product / service	- Product categorization - Pricing components	-	- Product categorization	- Name - Documentation metadata

Table 9. Data elements brought up in interview 8.

The interviewee mentioned that although contact details are an important data element for all business domains, the emphasis within the domains can vary, e.g. the street address is mostly important only for customer service who send customer the paper-versions of the terms and conditions of contracts, etc. In the other domains, email and phone are mainly used as a means of contact.

Risk management calculation components refer to those data that are used to perform calculations for determining e.g. the solvency, credit risk rating and profitability of a customer. These enable risks for the organization to be better managed, and also, if the data for the components are wrong, the risk profile of a customer can be incorrect and affect the incoming cash flow of the organization.

The B2B customer classification refers to different classifications of companies such as industry classification, legal configuration of the company and sector classification (part of regulatory requirements as well, due to being required by an authority). B2C customer characteristics include such things as the personal customer's lifecycle phase and their socioeconomic status.

Customer economy data element includes details about a customer's economy:

- A personal customer's income and expenditure, customer's household, household income and expenditure, etc.
- A company's turnover, profit, number of employees, etc.

Customer segmentation can be done in several ways, e.g. based on the

- value of a customer (how many products of the organization the customer is currently using, what is the potential for this customer),
- profitability of a customer,
- customer profiling (e.g. single, just married, family with kids, retired, etc.),
- customer typing (e.g. self-employed, agriculture entrepreneur, foundation, association, etc.), or
- customer relationship status (e.g. based on length of relationship, etc.).

The interviewee said that further data requirements for a customer could also depend on these different classification and segmentation data of a customer. For example, a current customer needs more data about them than a potential or former customer.

For sales & marketing, customer relationship management and customer service domains it is important to know what products and services the customer currently has/uses and with what terms. These data combined with other data about customer help to determine where opportunities with that particular customer lie. By gathering data about the customer's needs now and plans in the future, this opportunity recognition can be taken even further. In practise though, the monitoring of these data for data quality needs to be done using indicative factors, data from such real-life events as customer visits, customer contacts, etc. that help to understand how e.g. current the data had about customer is. The customer visit/contact data are also important to monitor for completeness, i.e. have all required data been filled out for each event, as well as for metadata, e.g. specifying the context of the event.

Contractual data in the context of this interview means the customer-specific terms and conditions agreed for services, the lifecycle data of a contract (potential, just started, running out, etc.), margins, interest payments, etc.

The interviewee saw identification data as critical but said that in his company, this was not a big data quality problem as for the type of business they are in, this is so critical that they had to have the data 100% correct and therefore very good mechanisms in place to stop it from being an issue.

Marketing and privacy consents/preferences include all data about what are the levels of privacy customer wishes to keep, whether they can be contacted for marketing and through which channels, where they want to receive their communications from the organization (e.g. invoices) and which channels customer would like to receive confidential information through.

Location data refer to data about the location of e.g. a customer's branches (B2B) or property (B2B and B2C). This affects for example the value of a property and makes the organization's operations more effective and proactive, e.g. in cases of natural catastrophes location data help the organization to map the effected properties of its customers effectively (to be able to offer customer pre-emptive customer service, as well as help themselves to prepare for future insurance claims and other possible actions from customer).

For sales & marketing it is key to also know about the customer's total assets to evaluate e.g. what assets have services against them, and where there is potential to sell more services. From sales & marketing and risk management points of view, it is important also to know about the value network of customer, i.e. its customers and partners.

Documentation metadata is important in the banking and insurance industry as there are many documents around that are strictly regulated to be viewed only by a restricted audience. One of the key metadata elements is the date of expiry after which the documents cannot be used anymore.

4.3.9 Interview 9

The interviewee agreed with the three data domains suggested in the preliminary matrix: customer, product and vendor. For the three business domains she nominated sales, marketing and warehouse management due to the fact that at her current employer, these are the areas where most issues occur. She mentioned that she sees finance as being a part of all business domains, not as a business domain on its own right.

The interviewee mentioned that those data elements that are commonly considered as pain points for data quality (e.g. many data elements in the exemplary matrix) are quite well covered at her company through process management, that is, they are using e.g. a customer creation process that eliminates many of the possible data quality issues with data such as payment terms, credit ratings, etc.

In her organization the availability of data is an issue: data might exist but people are not aware of it to look for it. A lot more analytics and process improvements might be possible to do if these data were acknowledged. Also, many processes are done manually currently, e.g. assigning of costs for customer or product profitability, where it could be done more automatically if data in systems supported it.

The data elements that the interviewee brought up as the most important in her current environment, are detailed in the matrix below.

	Sales	Marketing	Warehouse Management	Cross-business domain
Customer	- Organizational structure - Profitability calculation components - Customer classification & components	- Customer classification & components	-	- Identification data
Product	- Organizational structure - Profitability calculation components - Pricing components - Product categorization - Regulatory requirements - Product naming	- Product categorization - Product naming	- Size/dimensions - Location data	- Identification data
Vendor	-	-	-	-

Table 10. Data elements brought up in interview 9.

The interviewee emphasized customer and product profitability as extremely important factors in sales currently. To be able to calculate these profitabilities, it is vital to know all the costs incurred by customers and products. Customer profitability refers to not only knowing the incoming revenue by a customer but also the internal costs incurred by the customer e.g. the salary costs of employees serving them. With product, it is not enough to know the purchasing price of the product or components but also e.g. the handling and warehousing costs. Costs are very much dynamic data that are hard to monitor but the organizational structures behind customers and products can be monitored to be correctly in place, so that e.g. the assignment of costs is done correctly, therefore helping the profitability calculations.

Another related data element would be profitability calculation components, in a similar was as pricing components can be monitored for data quality. There are many single data elements that contribute to customer, product or whole organization's profitability as mentioned above.

Product size/dimension data as well as location data (i.e. where the product is physically located) were raised as important data elements by the interviewee for optimization of warehouse management. She told that in her company, the measurement details of products were previously simply guesses but that the importance of these data are now acknowledged and huge effort has been put into correcting this data.

The identification data was considered important for all data domains across (relevant) business domains. For customer identification the interviewee wanted to point out that this also included the ability to identify a customer e.g. when browsing online. (The vendor was not marked with identification data on the matrix above as there are no relevant business domains for vendor present in the matrix.)

With customer classification, the aim would not only be to have the classification for typical uses in operations and analyses, but also to have the components for determining the classification for a customer of such good quality that classification could be done automatically based on those components.

5 Conclusions

Conclusions on the thesis topic and specifications for the final construct put together are detailed in chapter 5.1. The realigned targeting matrix for data quality monitoring that combines the results of the thesis project and answers the main research question is then presented in 5.2. Chapter 5.3. gives recommendations for the further development of the data quality monitoring targeting construct. Chapter 5.4. considers the proceeding of the thesis project from an academic learning, project management, objective-achieving and self-development point of view.

As with the interviews before, where the framework made during the thesis project was presented to the interviewees but was not forced to be considered the only direction for discussion, the final targeting matrix presented in chapter 5.3. does not force a particular point of view on the reader. There are therefore several data elements that are e.g. included in two business domains for one data domain or included in the matrix although possibly not as generic as originally intended by this project. The justifications for these decisions to set up the final targeting matrix in this way are given in the next sub-chapter.

The conclusions of this thesis are for recommendation only, and have to be adapted to an organization's environment and further investigated when used in a data quality monitoring endeavour. The final data quality monitoring targeting construct has been built on individual persons' experiences with several organizations across industries, as well as partly on the assumption that the main data in the main business functions will represent the main issues in data quality as well. All the experiences from the interviewees are from medium-large to large organizations. The smaller the organization, the more specific its problems, and therefore it is harder to generalize on those cases.

As was the purpose of this thesis, the theoretical part of the report "paved the way" for the reader to understand the results of the empirical part, giving background information on data quality that is common knowledge for persons working in the data and information management fields, such as the interviewees involved in the field study. There are some references to the theoretical part of this report in the conclusions, but mainly the theory's role is one of an enabler to understand and recognise the meaning in the results.

5.1 Results of thesis work

The original idea of the thesis, grounded in the theoretical part of this thesis and commissioning party's experiences in the field, was that by recognising the key business domains and the main data domains would provide a basis for discovering those data elements within these domains that were critical in terms of data quality.

Determining the importance of one data element in comparison with another in regards to data quality monitoring proved a thought-provoking topic during the field study interviews although not implicitly indicated as a topic for discussion. The interview invitations and information provided for interviewees prior to the interviews were clearly indicating the project's way of determining this, simply asking for ideas on the different business and data domains, as well as the data elements within those. This was also the approach taken during the interviews conducted.

Many of the interviewees had anyway given this a lot of thought and provided some further ideas for the approach to determining the key data elements. Some of those points of view are discussed here due to the fact that this is obviously an important topic, otherwise it would not have come up in most interviews – and also, the differences in opinion regarding this can be a reason for no unified ideas having been formed and written down in literary form before. Unless not viable for putting together a generic subset of business critical master data for data quality monitoring, these points of view are not agreed with or discarded. The targeting matrix that is the result of this thesis project is, however, based on the original approach to determining the important data elements.

A common point of view was to tie the data elements monitored to the strategic goals and KPIs of the organization. It was seen that it is often, and it should be, the management of the organization that initiates the data quality improvement process based on these factors. They would then also indicate the data elements and the associated rules for the monitoring.

Critical data elements were also considered the ones that ensured the smooth running of processes. If those data elements were not of good quality, that could stop processes completely and incur considerable costs for the organization. This coincides with the theoretical part of this thesis that specified that process failure costs are recognised to be one of the leading reasons for data quality criticality. This is also not differing too much from the method used by this project – considering the key business domains as part of determining key data elements. Moreover, it was stated during interviews that if an organization operates globally, the data that are global and shared across business units should be considered more critical than local data. After this, the key processes and the data they use should be looked at for the globally shared elements.

There was also an idea that by looking at the data objects' structures in an organization's system databases, it would be quite easy to determine the possible loopholes where data quality issues might arise. This method would not, however, specify if those data quality issues would be critical for the running of the organization, and is therefore not viable for the objectives of this project without further development.

Another suggestion was to look at the key business and data domains as done by the project, but also add an aspect of considering the volumes of business for each of those. For example, if the key business and data domains were sales & marketing with customer and product, one should look at the best-selling products and combine those with the most-buying customers to narrow the scope further. Once this scoping was done, the transactions within sales & marketing for this scope of products and customers should be analysed for the data elements that were featured the most. From those, it would be easy to determine the data elements that had or could potentially have data quality problems associated with them. As the results of this thesis are attempting to be as widely applicable as possible, this method of determining the critical data elements would obviously not work fully. This would definitely be something to consider when going forward with setting up data quality monitoring at a particular organization, at least as a basis for discussion with the organization.

It was also emphasized in several interviews, as well as already acknowledged by the project (being the reasoning for setting the scope of the work as *master data* which by definition are shared by functions and systems and furthermore to their *quality* which by definition is a subjective matter), that most critical data are cross-domain data, and that they can be looked at from multiple points of view with different results. One set of data can be of good quality from one perspective, but from another, it can be of very poor quality. This is where the criticality examination is paramount – what perspective is the most important to have good data quality from? In the final targeting matrix, most if not all perspectives that were mentioned in interviews have been included, i.e. if a data element was considered critical for sales & marketing by one interviewee and for finance by another, it has been included in both. It is again useful to have both these perspectives included in the matrix when setting up data quality monitoring for a particular organization, to open discussion on what perspective in particular is critical for that organization.

All data domains in the preliminary data quality monitoring targeting matrix (customer, product and vendor) were verified by interviews to be critical. Only one interviewee did not nominate vendor to be part of the critical data domains, and another one did agree with vendor being a critical data domain but as it did not apply to their area of expertise, they did not nominate it for their part. Many interviewees commented though that if one data domain was to be replaced, it would be vendor as it is not as all-encompassing as customer and product.

Many interviewees emphasized the diversity of the data domains: Customers can be B2B customers or B2C customers or something in between. Products can be materials (raw materials, packaging materials, marketing materials, etc.), production components, finished products, services or even benefit applications. The lifecycle stage of customer or product can also affect the data requirements that the domain has, and that are critical to it. Vendor was the only fairly unambiguous data domain, although even vendor can be seen as an extension of the internal organizational structure instead of a traditional external stakeholder, as specified in an interview. The main message was that even if consensus was reached with data domains within the interviews, data domain specifications need to be known to determine the critical data elements for that domain in a particular organization.

Additional data domains were also suggested. Two interviewees mentioned employee data as becoming more and more critical to organizations. Although these data are not connected to external processes transparently, several connections exist which are making these data critical to the running of a business. Data about employees or facilitating employees' work that are not of good quality, can cause an organization to lose their positive image with external stakeholders as well as put them at high risk of compromised security (both physical and virtual).

Other data domains suggested in the interviews were more varied and were only mentioned in one interview (e.g. production data, lifecycle). Some of them were considered by other interviews as data elements or business domains (e.g. pricing, financial data domain) so there was no consensus between the interviews on those and they were not selected for the final targeting matrix. There were also necessarily no data elements given for those other additional data domains (e.g. production data, lifecycle) so this is a further reason for them not having been considered for the final targeting matrix.

There was more discussion on the business domains. The only business domain that was agreed between all interviews was sales & marketing (although in one interview the business domain was divided into two business domains: "sales" and "marketing"). Most interviewees could justify the business domains they nominated very well, and therefore the only criteria for choosing the top three business domains for the final targeting matrix was which domains were mentioned in most interviews. This was to support the objective of building as generic a construct for data quality monitoring targeting as possible. This approach resulted in the critical business domains to be sales & marketing, supply chain management, and a combined business domain that can be called finance that consists of many financial processes mentioned in several interviews.

In these conclusions supply chain management is seen to consist of purchasing/sourcing, production, logistics (inbound and outbound), warehouse management and customer service. This then also covers a great deal of business domains mentioned separately in some of the interviews due to the particular set up of functions/processes in the interviewees' organizations (e.g. production and purchasing).

Finance was a strongly debated business domain. Many interviewees started by stating that finance is not really a business domain on its own but ended up bringing up many financial data elements or factors that were critical. Several business domains were suggested that would fall under the umbrella domain "finance" but depending on the organization, can be separate functions/processes. Some used the business domain management reporting, others e.g. risk management, controlling and financial structures. Financial data elements are tricky to pinpoint to the "finance" business domain because they can just as easily be placed under other business domains. This fact was mentioned in many interviews: finance is part of all other business domains and often contains data created by other domains, financial structures still being the first data that need to be in place for other data to be correctly connected to the fundamental structures of business.

As a conclusion, finance can be considered an omni-present domain that exists in some way in all main functions/activities of an organization. Most of the data elements usually connected to finance are actually derived from elsewhere – e.g. payment terms are a data element that exists in vendor/customer master data and then in purchase/sales order dynamic data. In the finance process, the payment that is the consequence of the purchase/sales order is simply paid/received according to those terms but does not offer any value on its own. If terms are not abided to, further processes are set off that are purely financial domain processes and this is where the data element suddenly becomes a vital trigger within finance alone. In this report's conclusions and in the final targeting matrix presented in the next chapter, these elements are placed under both: the different operational business domains and in the finance domain. As for the entire construct, the point of view for this can then be determined on a case by case when setting up data quality monitoring in an organization.

Other business domains were brought up, human resources being the only one that was mentioned more than once (mentioned by the two interviewees that also nominated employee for the data domains). In the same way as for the data domains, for business domains as well, there were nominations that were suggested as data domains or data elements in other interviews.

The diversity aspect of business domains was also mentioned but not emphasized as much as for data domains. One interviewee gave as an example that sales data requirements for a product differ greatly depending on the sales channel, whether online or in a shop.

There was a point of view that, although only mentioned by one interviewee, was such a valid and challenging issue it needs to be addressed also in these conclusions. This was the question of how to include analytical data needed for reporting in the targeting matrix (excluding the management reporting data that have been included in the finance business domain in these conclusions). Reporting is extremely important in today's world of business intelligence, big data and analytics. There is no question as to data needing to be of good quality for reporting to be effective, and poor quality data having potentially catastrophic effects. But sometimes data used in reporting are not naturally born out of any business domain but need to be artificially inserted into its data. All the data in the preliminary matrix are naturally born out of the normal running of a business. The analytical data needed for reporting might reside under a business domain like sales & marketing but is not used by sales & marketing, simply belonging to a data domain that is relevant to that business domain (e.g. the cost of goods sold data are needed to make profitability calculations (a financial task whose result is used for management decision-making), and might reside under sales & marketing due to being connected to customer profitability). These data are critical and need to be in order but the way in which to integrate it into the construct developed in this project remains an open issue. Certain analytical data elements, such as profitability calculation components (under umbrella term "Risk management calculation components"), were included under the finance business domain in the final targeting matrix. The interviewee would not add reporting as its own business domain but considers it a third dimension to the matrix.

On top of the above-mentioned data and business domains, each of those had a "common" element to them where the same data element would be critical or important for all data domains or all business domains. For this reason, as in the preliminary matrix was already specified for business domains, a cross-domain cell was added for both data and business domains. Certain data elements were critical for all data domains and all business domains.

One interviewee mentioned that in his opinion each data element should have a "global" and a "local" section, as all data elements can be either shared by all business units of an organization or stay local for a particular unit. This point of view is quite system-specific to SAP and does not add value to the results of this project (as a business domain connection would still remain to be established), therefore it has been not considered for the final targeting matrix.

The scope of this research was to take into consideration only master data elements, as specified in the theoretical part of this report. As research interviews were not stressed to be regarding master data only (apart from a reference to master data in the original e-mail sent to interviewees), some other suggestions for the monitoring targeting matrix were given. It is the conclusion of the researcher, that these data elements were valid and some of them are considered for the data quality monitoring targeting matrix either directly if possible or as a recommendation in chapter 5.3.

As suggested in the theoretical part of this report, the interviewees also concentrated on the semantic and pragmatic aspects of data quality, that are more common to have quality issues due to not always or easily being implementable by technical constraints in system databases or program logic. Identification data, for example, that in theory are easy to keep in order on a syntactic level by forcing uniqueness of IDs, are more often a cause for semantic and pragmatic data quality issues due to there existing e.g. duplication, that is, several instances (all with unique IDs) of a data object (like customer) for one phenomenon in real life. This kind of issues cannot be prevented by forcing system constraints but by processes and monitoring. Even with processes, there is always room for human error, and therefore, monitoring the quality of identification data is often the only way to guarantee to keep the quality of these data good.

None of the data elements featured in the preliminary targeting matrix were disagreed with by the interviewees. Some of the data elements would have been moved by interviewees to be under different business domains, e.g. payment terms were seen by many interviewees as a sales & marketing data element more than a financial data element, or expressed differently or on a different level, e.g. prices were not seen as appropriate elements for data quality monitoring but the components making up prices were, etc. These have been considered, and changes to these from the preliminary targeting matrix can be seen in the final targeting matrix.

During the interviews it became clear that to put together a targeting matrix that is applicable to organizations across most industries, the key is to generalize and to use "umbrella terms" for similar data elements that could then be interpreted according to an organization's needs and situation. This approach was reinforced by the diversity of data elements that was obvious from the interview results: the same data elements could be expressed by different terms by different interviewees, or contain a different amount of sub-data elements than for other interviewees. The industry of the organization hugely affects the point of view from which a data element is looked at – customer segmentation data for a retail company is totally different to a banking and insurance company where segmentation can be multi-dimensional for example.

As suspected, not many single-domain data elements (elements only relevant for one business domain per data domain) were nominated by interviewees for the critical subset of data. It was mentioned by a few interviewees that rarely are single-domain data as critical for data quality as data shared by domains. Some single-domain elements still exist in the final targeting matrix but were not as commonly mentioned in interviews as elements affecting many domains.

In the preliminary matrix, the relationships between different data elements were indicated, e.g. monitoring two data elements in conjunction with each other would provide most benefit. These connections did not come up in interviews much although certainly relationships exist. It was therefore decided to exclude these relationships from the final matrix, to be considered separately if and when needed by an organization during set up of data quality monitoring.

Some data elements in the final matrix can be somewhat overlapping with each other, but this is due to there being separating factors as well. For example, the payment information data element includes all data dealing with payments, e.g. payment terms. Payment terms can also be seen to be part of contractual terms data element as they are most commonly agreed by contract. However, both data elements also have other data that are not possible to combine with each other so the decision was made to keep both data elements separately. Same applies for delivery terms under delivery information and contractual terms, and pricing data that are under pricing components as well as contractual terms. Having both data elements is, again, also good in case one way of presenting the data is more eye-opening than another for organizations when setting up data quality monitoring.

Sales and purchasing prices were included as data elements in the preliminary matrix built prior to interviews. As verified by interviewees, pricing data are highly critical in all organizations. But, the pricing data themselves were considered by many as dynamic data and therefore volatile, with high frequency changes - not suitable or at the very least, highly challenging for data quality monitoring as such. One interviewee did mention contractual/annual pricing agreements that stayed stable or fairly stable (in case of contractual changes) for a whole year, therefore qualifying the price itself for data quality monitoring. Many interviewees challenged the previous logic of the thesis, and suggested that instead of pricing data, the pricing component data were to be monitored for data quality with the help of associated pricing calculation formulas that offer a way to verify the correctness of any pricing data selected. Due to the unanimity of interviewees that pricing data should be considered in one way or another, this suggestion to use the pricing components for the matrix was adapted without question.

In the preliminary matrix hazardous information was one of the data elements for product within logistics. This was a somewhat narrow view on an important matter, and served to fuel discussion on other similar data elements that were actually usually in place due to requirements by authorities or law. It was suggested by some interviewees and in the author's point of view is a valid point, that regulatory requirements could be seen as a business domain so that single data elements could be connected to a data domains (e.g. product has allergens and can be hazardous, but a customer is authorized to buy a product classified as a narcotic). Regulatory requirements does not, however, align with the definition of a business domain made in this report (i.e. "the business functions, processes or sub-processes that exist in a business"). Therefore, in the final matrix, regulatory requirements are considered a data element. This approach allows the data to be present wherever needed in the matrix and also leads to a more generic usage that does not consider a single organization or industry specifically. The umbrella element "Legal/regulatory requirements" is used and can contain data from various different regulatory areas such as health & safety, taxation, customs and quality. They are all critical data elements, and as one interviewee pointed out, elements that are thought to be important due to possible sanctions if not followed but that are actually extremely important also for an organization's image and for the fact that they contain a huge risk factor for both external and internal stakeholders of the organization. There is an appendix to the final matrix that specifies the different responses from interviewees that fall under this and other umbrella elements.

An interesting point was brought up about the data element product categorization that was already featured in the preliminary targeting matrix. Product categorizations tie to different financial structures such as cost or profit centres that navigate the costs and revenues from processes correctly in accounting. At the same time these categorizations can be used for e.g. lifecycle management, simple navigation in a web shop, as a type of ID, etc. For the multitude of uses, the product categorization data element was raised to be cross-business domain in the final matrix.

Many of the very customer-centric data elements mentioned in interview 8 were very valid in terms of important data for the type of financial organization in question. Whether those data elements can be applied to many other types of industries is another matter. Also, some of the data elements are not fully suitable for data quality monitoring as there might be difficulty in determining the measures to monitor them by. Some of those elements are on the borders of qualifying for the scope of this thesis, but have been included in the final targeting matrix regardless, with the justification that these are very "hot" topics in customer data management in many financial sector industries. As with the data matrix on the whole, it should be interpreted from the appropriate point of view when utilized for setting up data quality monitoring for an organization. Therefore, these data elements can be discarded if needed from any organization-specific endeavour.

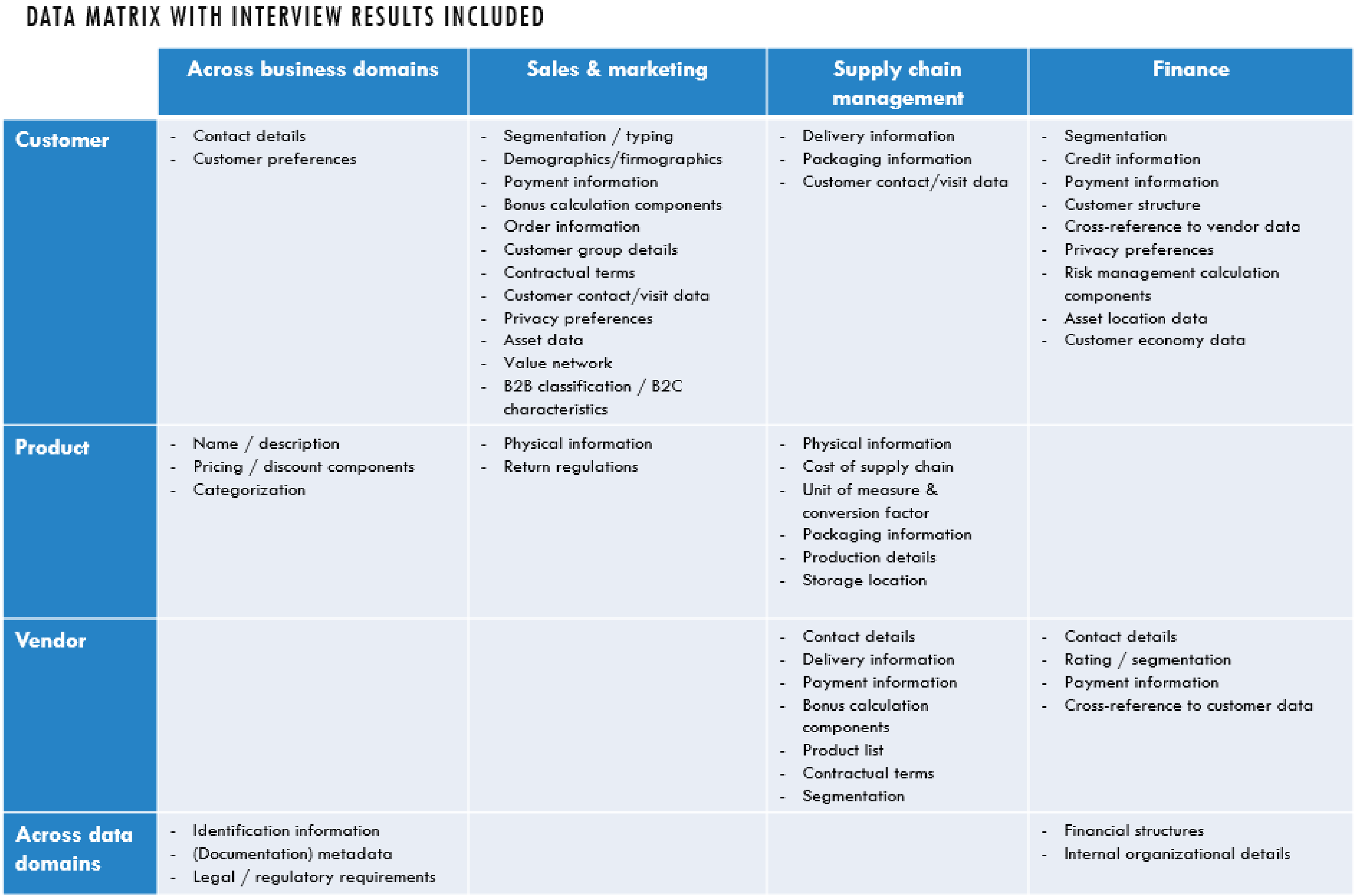
The only data domain that did not have any cross-business domain data elements is vendor. This verifies the statement by some interviewees that vendor is not as critical a data domain as customer and product but actually quite a business domain-specific one. However, no other data domain that could have data elements valid for all business domains was suggested, therefore vendor remains in the final matrix as well.

Those data elements that are not specifically discussed in these conclusions but are included in the final targeting matrix, were clear choices for the matrix based on the theoretical and empirical parts of the thesis, and, furthermore, not disputed in the interviewees or found to have disagreement on when analysing interviews.

5.2 Realigned Targeting Suggestion for Data Quality Monitoring

Based on the results of the empirical part of the thesis, a realigned targeting suggestion for data quality monitoring was put together (picture 7). This construct is the main result in terms of answering the research question of what the most business-critical subset of all of a business's master data in terms of data quality is. This subset should be targeted by data quality monitoring.

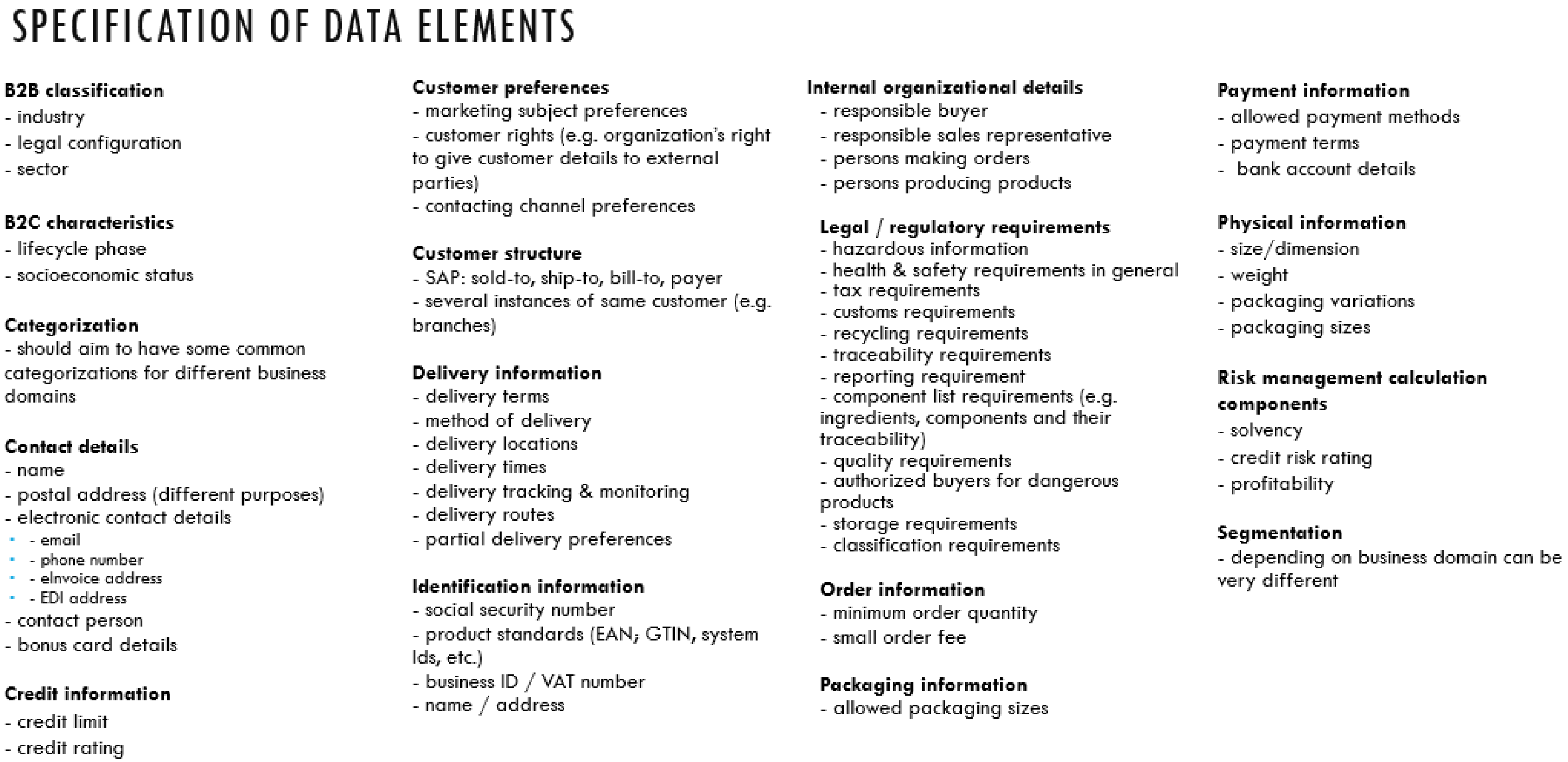
The justifications and specifications to the targeting matrix were given in the previous chapter. To adhere to the objective of the thesis, discovering a *generic* subset of master data, as varied a representation as possible of different perspectives expressed in interviews is incorporated into the construct. To the same extent of reaching genericity, many "umbrella terms" are used to describe data elements expressed in interviews. Many of the "umbrella terms" used in the matrix have been opened up in the picture 8 and give an extensive list of the exact data elements given in interviews.



DATA MATRIX WITH INTERVIEW RESULTS INCLUDED

	Across business domains	Sales & marketing	Supply chain management	Finance
Customer	- Contact details - Customer preferences	- Segmentation / typing - Demographics/firmographics - Payment information - Bonus calculation components - Order information - Customer group details - Contractual terms - Customer contact/visit data - Privacy preferences - Asset data - Value network - B2B classification / B2C characteristics	- Delivery information - Packaging information - Customer contact/visit data	- Segmentation - Credit information - Payment information - Customer structure - Cross-reference to vendor data - Privacy preferences - Risk management calculation components - Asset location data - Customer economy data
Product	- Name / description - Pricing / discount components - Categorization	- Physical information - Return regulations	- Physical information - Cost of supply chain - Unit of measure & conversion factor - Packaging information - Production details - Storage location	
Vendor			- Contact details - Delivery information - Payment information - Bonus calculation components - Product list - Contractual terms - Segmentation	- Contact details - Rating / segmentation - Payment information - Cross-reference to customer data
Across data domains	- Identification information - (Documentation) metadata - Legal / regulatory requirements			- Financial structures - Internal organizational details

Picture 7. Realigned (final) targeting matrix



SPECIFICATION OF DATA ELEMENTS

B2B classification
- industry
- legal configuration
- sector

B2C characteristics
- lifecycle phase
- socioeconomic status

Categorization
- should aim to have some common categorizations for different business domains

Contact details
- name
- postal address (different purposes)
- electronic contact details
  - - email
  - - phone number
  - - eInvoice address
  - - EDI address
- contact person
- bonus card details

Credit information
- credit limit
- credit rating

Customer preferences
- marketing subject preferences
- customer rights (e.g. organization's right to give customer details to external parties)
- contacting channel preferences

Customer structure
- SAP: sold-to, ship-to, bill-to, payer
- several instances of same customer (e.g. branches)

Delivery information
- delivery terms
- method of delivery
- delivery locations
- delivery times
- delivery tracking & monitoring
- delivery routes
- partial delivery preferences

Identification information
- social security number
- product standards (EAN, GTIN, system Ids, etc.)
- business ID / VAT number
- name / address

Internal organizational details
- responsible buyer
- responsible sales representative
- persons making orders
- persons producing products

Legal / regulatory requirements
- hazardous information
- health & safety requirements in general
- tax requirements
- customs requirements
- recycling requirements
- traceability requirements
- reporting requirement
- component list requirements (e.g. ingredients, components and their traceability)
- quality requirements
- authorized buyers for dangerous products
- storage requirements
- classification requirements

Order information
- minimum order quantity
- small order fee

Packaging information
- allowed packaging sizes

Payment information
- allowed payment methods
- payment terms
- bank account details

Physical information
- size/dimension
- weight
- packaging variations
- packaging sizes

Risk management calculation components
- solvency
- credit risk rating
- profitability

Segmentation
- depending on business domain can be very different

Picture 8. Specification of umbrella data elements in the realigned (final) targeting matrix.

5.3 Recommendations for Further Development of Construct

This report gives a generic basis for the targeting of data quality monitoring. As mentioned earlier, the construct – that is the data quality monitoring targeting matrix – that was developed based on the research during this thesis project should be looked at critically and analysed though each separate organization's business and data environment point of view. It does not and is not pretending to offer one absolute truth.

There are some recommendations for the potential further development of the construct that were formed during the work of this thesis project. It is recommended by the author that the construct is first tested with organizations wishing to set up data quality monitoring before modifications are made. The possible developments that were highlighted already during the conclusion-making of this project are detailed below.

Developing the construct can be divided in two categories, not mutually exclusive in any way:

1) Developing the construct structure
    a. by amending the domain specifications (e.g. adding more than three each of data and business domains), or
    b. by amending the level of generality of the data elements (e.g. opening up umbrella terms)
2) Developing the construct content
    a. towards a particular industry's requirements, or
    b. towards a particular organization's requirements

For the first category of development, the recommended areas to further investigate would be the representation of analytical data in the construct, the addition of human resource/employee domains, the representation of data element relationships in the matrix and decreasing the level of coarseness of data elements.

Further investigating the representation of analytical data in the construct refers to determining whether a third dimension should be added to the matrix as suggested by one interviewee, or if analytical data can be represented by e.g. data elements in each relevant domain. Is it possible to “locate” these data elements that are “unnaturally” included in the domains without the analytics/reporting connection? It may well be that the best practise for analytical data representation in the construct will only be defined once the construct has been tested.

Employee data and the associated human resources business domain were brought up by two forward-thinking data professionals that were among the most experienced data experts in the group of interviewees. It is therefore highly recommended to consider adding these data and business domains to the matrix template, making the matrix a 5x5 construct. But in the very least, those domains would be recommended to be added particularly when approaching organizations in certain industries (public sector, for example).

The preliminary matrix of this interview specified relationships between data elements. The purpose of this was to highlight the possible need to monitor those elements for data quality in conjunction with each other for most benefit. This approach/idea was introduced to the interviewees during the field study of this project, but was not concentrated on. Interviewees did not discuss data element relationships further although agreeing to the principle when it was introduced to them. Relationships were therefore not included in the final targeting matrix. Based on testing the matrix and setting up monitoring, it might be a worthwhile subject to consider re-introducing relationships to the matrix at a later stage if found viable.

If the use of umbrella terms is found to be more deteriorating to the finding of organization-specific data quality issues than being the conversation-/idea-stimulating approach that it was thought to be, it would be recommended to open up the terms used according to the particular organization it is aimed at. This will prove a bit more work for the commissioning party but if it could result in a happier customer, it would make it worth the effort. In this approach, of course, there is always the risk of including too few or too many of the sub-terms from each umbrella term, causing confusion or even less ideas.

The second category of development could include scoping the data elements further on an industry or an organization level. These developments could, of course, not only affect the content of the matrix (e.g. the particular domains or data elements featured) but also, inevitably, the structure of the matrix (e.g. adding more than three data or business domains).

Developing the construct on an industry level would mean featuring industry-specific domains on the construct. This would be a valid suggestion especially if construct was to be used on a large scale and not much time was to be used on its "customizing" per case, as data specifications between e.g. an engineering industry organization compared to a banking and insurance industry organization differed significantly in the interviews conducted during the thesis project. However, the data specifications within these industries can be generalized fairly confidently even if a particular organization to use the construct is not known yet. It is suggested that the individual interview summaries are analysed further if deciding to develop the construct in this way.

Developing the construct on an organization level could mean, e.g. using the approach suggested by one of the interviews of taking into consideration the top-occurring/-performing instances of each data domain, combined with the key business processes of the organization in question, to determine the data elements that are critical. This would be an option if the construct would need to be used in an organization where not much in-house knowledge on data quality was held, and a lot of external input was needed for the analysis of data criticality. This kind of development can take fairly a lot of resources though and should be taken on with consideration.

On a separate note, the exclusion of "single-domain" data elements (only existing for one data and one business domain) could also be worthwhile. It would decrease the number of data elements in the matrix to make it more clear and correspond to what many interviewees said about the most critical data elements always being shared by several domains.

5.4 Self-Evaluation of Thesis Project and Academic Learning

Overall the thesis project ran smoothly: schedules were kept without bigger problems, the objectives of the project were reached. The small "hindrances" faced by the project (e.g. summer holiday causing delays in finishing certain parts of the report on type, resulting in the postponement of the second status meeting by a week) did not affect the overall result, and did not cause inconvenience to any parties involved.

The project started on 11.4.2014 and ended on 20.10.2014. Altogether 255 hours were used out of the planned 296 hours, both well under the recommendation of 400 hours by HAAGA-HELIA University of Applied Sciences under whose guidance the project was completed. Work load divided fairly equally between all parts of the project, although the analysing of interviews and writing of practical part of thesis portion consumed more hours than expected. This was the only work entity whose hours were underestimated in the project plan.

The thesis report was also completed as expected: the scope, methods used and content are all as planned. The scope of work supported work throughout the project and there were no major deviations from it at any point. The methods used also supported project work and objectives, and resulted in achieving the level of detail planned to be achieved for the conclusions of the project. The decision to make an additional email verification round for the findings from the interview was grounded in terms of gaining more confidence in the generality of the results, but at the same time, might have been misguided in terms of expecting the interviewees to commit time to additional, written questions. If the resources of the project had been sufficient, an additional round of calls to those interviewees who did not answer the verification email, would have probably solved the issue. After all, people tend to prefer to spend a few minutes answering questions verbally instead of needing to spend the same time (or even less) on a written "task". These calls were consciously not made, which resulted in only receiving a few answers to the verification email. This is not seen to in any way take away from the reliability or quality of the results. However, it does highlight the recommendation for the testing of the matrix with some organizations prior to launching it fully.

The content of the report is in line with what is described in the thesis background and in the author's point of view, ensures a red thread running through the report. The only "conflict" within the content was regarding defining scope to be master data but still expecting data elements not aligning with the scope to be in the resulting construct: Master data by definition are shared by many functions and systems. In the preliminary matrix, some "single-domain" data elements were given (only valid for one business domain per data domain). Was it logical to expect to find many "single-domain" data elements in the interviews considering the definition of master data being shared by many functions? It was mentioned by many interviewees that usually single-domain data elements are not as critical as cross-domain data elements. It was, however, also mentioned that even if a data element is present in many domains, it doesn't mean that its quality is as important to all of them. This leads the author to think that no absolute 'yes or no' conclusion on whether single- or cross-domain data are more critical can be made. In that sense, the scope setting and somewhat differing preliminary data quality targeting suggestion (based partially on the experiences of commissioning party representatives) was fine.

The thesis project lasted for some six months. For writing a consistent report, this is quite a long time. Some parts of the final report have been written in April/May of 2014, the last parts in October of the same year. Reading through the report regularly keep it consistent is advisable, and not done regularly enough during this thesis project. However, in the author's opinion, this did not harm the final report after final revisions.

The author of this thesis considers the learning during the thesis project to be mainly professional, and helping in the author becoming a data and data quality management professional like the persons interviewed. The subject of data management was not covered during the author's studies at HAAGA-HELIA much. Data quality consideration was left at the syntactic level of information system design and integrity constraint conformance. The biggest realization made during this project was that understanding data quality on a theoretical level is a whole other matter, than actually achieving it in practise. Discussing ways to achieve data quality in theory is interesting and fairly straight-forward, but actually putting the plan into action with the complex business, system and information environments is far from simple.

Of course, academic learning aspects were also present: writing the report acted as a reminder of how to write in an academic way. For someone like the author, who has a very personal style of expression in general, this was a learning curve indeed – something forgotten during the years in working life and later doing very practical studies.

Sources

Aiken, P. 14 January 2014. Demystifying Big Data. Webinar presentation. Data Blueprint. URL: http://www.datablueprint.com/webinars/demystifying-big-data-2/.

Aiken, P. 8 April 2014. Unlock Business Value through Data Quality Engineering. Webinar presentation. Data Blueprint. URL: http://www.datablueprint.com/webinars/data-quality-engineering/.

Alagse Consulting. 2014. Framework for Identifying Key Business Processes. URL: http://alagse.com/pm/p4.php. Read: 22.4.2014.

Butel, L. Curtis, T. McIntyre, J. Pearce, J. Rainbow, S. Smith, D. Swales, C. 2005. Business Functions: An Active Learning Approach. The Open Learning Foundation.

Eckerson, W. 2002. Data Quality and the Bottom Line. Achieving Business Success through a Commitment to High Quality Data. The Data Warehousing Institute (TDWI) Report Series. The Data Warehousing Institute. Chatsworth.

English, L. 1999. Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. 1st ed. John Wiley & Sons, Inc. New York.

Eppler, M. 2006. Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes. 2nd ed. Springer. Berlin.

Failte Ireland. 2013. Managing Business Processes. Business Tools resource guides and templates series. Failte Ireland. Dublin.

From, R. 12 May 2014. Towards Better Analytics. Industry Manager. Google Finland. Seminar presentation. Big Data and the Financial Services Sector Seminar. HAAGA-HELIA University of Applied Sciences. Helsinki.

Funk, J., Lee, Y., Pipino, L. & Wang, R. 2006. Journey to Data Quality. MIT Press. Cambridge.

Gulati, R. 2007. The Four Cs of Customer-Focused Solutions. Harvard Business Review. URL: http://hbr.org/web/special-collections/insight/customers/silo-busting-how-to-execute-on-the-promise-of-customer-focus. Read: 26.6.2014.

HAAGA-HELIA University of Applied Sciences Thesis Guidelines Working Group. 2014. Writing Your Bachelor's Thesis: Contents and Methods. Moodle Learning Platform. BITe_Thesis. BITe Thesis Course Content. URL: http://hhmoodle.haaga-helia.fi/. Accessed: 11.6.2014.

Harris, J. 2009. The Nine Circles of Data Quality Hell. OCDQ Blog. 26.5.2009. URL: http://www.ocdqblog.com/home/the-nine-circles-of-data-quality-hell.html. Read: 24.6.2014.

Harris, J. 2011. The Role of Data Quality Monitoring In Data Governance: Aligning Data Quality Metrics with Business Insight. Dashboard Insight. 31.3.2011. URL: http://www.dashboardinsight.com/articles/business-verticals/the-role-of-data-quality-monitoring-in-data-governance.aspx?page=2. Read: 23.6.2014.

Haselden K. & Wolter, R. 2006. The What, Why and How of Master Data Management. Microsoft Developer Network (MSDN) library. URL: http://msdn.microsoft.com/en-us/library/bb190163.aspx. Read: 26.6.2014.

Hristova, M., Soft, M. & Tjurkmen, H. 2013. Data QualYtl – Do You Trust Your Data? Innovations in Software Test Automation (ISTA) Conference 2013.

Informatica. 2014. Start Monitoring Data Quality from Day One. URL: http://www.informatica.com/us/products/complex-event-processing/proactive-monitoring/proactive-monitoring-data-quality/. Read: 18.5.2014.

International Association for Information and Data Quality. 2014. IQ/DQ glossary. URL: http://iaidq.org/main/glossary.shtml. Read: 11.5.2014.

Investopedia. 2014. Supply Chain. URL: http://www.investopedia.com/terms/s/supplychain.asp. Read: 11.8.2014.

McKinsey&Company. 2014. Accelerating the digitization of business processes. URL: http://www.mckinsey.com/insights/business_technology/accelerating_the_digitization_of_business_processes. Read: 11.6.2014.

Kolehmainen, A. 2011. Master datan sekavuus oli luonnonlaki. Tietoviikko 14.4.2011. URL: http://www.tietoviikko.fi/cio/master+datan+sekavuus+oli+luonnonlaki/a611351?articlepage=1.

Kontra, K. 2014. Owner, Leading Advisor. Datpro. Interviews. April 2014. Helsinki.

Kontra, K. 2010a. Business Value of High Quality Data. Presentation. 2010 Information Quality Industry Symposium. July 2010. Massachusetts Institute of Technology (MIT).

Kontra K. 2010b. Data are facts, information is data in context - well... only sort of. Datalifeuniverse. Blog. 17.11.2010. URL: http://datalifeuniverse.blogspot.fi/2010/11/data-are-facts-information-data-in.html.

Oracle. 2009. 10 Understanding Data Quality Management. Oracle Warehouse Builder User's Guide 10g Release 2 (10.2.0.2). URL: http://docs.oracle.com/cd/B31080_01/doc/owb.102/b28223/concept_data_quality.htm#BGBBCBHA. Read: 23.6.2014.

Oxford Dictionaries. 2014. Definition of Monitor in English. Oxford University Press. URL: http://www.oxforddictionaries.com/definition/english/monitor. Read: 24.6.2014.

Pelkonen, P. 2006. Strukturoidun informaation laadun vaikutus yrityksen liiketoimintaan. Diplomityö/Thesis. Teknillinen korkeakoulu. Helsinki.

Porter, M. 1985. Competitive Advantage. 1st Free Press Export Edition (2004). Free Press, New York.

Price, R. & Shanks, G. 2005. A Semiotic Information Quality Framework. URL: http://www.researchgate.net/publication/228795537_A_semiotic_information_quality_framework. Read: 15.5.2014.

Redman, T. 2013. Data's Credibility Problem. Harvard Business Review. December 2013. URL: https://archive.harvardbusiness.org/cla/web/pl/product.seam?c=29814&i=29816&cs=8cd87434b8cc76d931d428673c7a3ece&mkt_tok=3RkMMJWWfF9wsRoluaXKZKX-onjHpfsX56eooWKO1lMI%2F0ER3fOvrPUfGjI4DScVkI%2BSLDwEYGJlv6SgFT-LXDMbdtzbgEWhk%3D.

Reh, F. 2014. You Can't Manage What You Don't Measure. URL: http://management.about.com/od/metrics/a/Measure2Manage.htm. Read: 18.5.2014.

Roebuck, K. 2011. Data Quality: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Emereo Publishing.

Ryan, L. 2014. 'If You Can't Measure It, You Can't Manage It': Not True. 10.2.2014. Forbes Online. URL: http://www.forbes.com/sites/lizryan/2014/02/10/if-you-cant-measure-it-you-cant-manage-it-is-bs/. Read: 18.5.2014.

University of Cambridge Institute for Manufacturing. 2014. Porter's Value Chain. URL: http://www.ifm.eng.cam.ac.uk/research/dstools/value-chain-/. Read: 26.6.2014.

Uniserv. 2014. Data Governance – Everything in View. URL: http://www.uniserv.com/en/products/data-quality-service-hub/data-governance/. Read: 18.5.2014.

Wikipedia. 2014. Business process. URL: http://en.wikipedia.org/wiki/Business_process. Read: 18.5.2014.

Attachments

Attachment 1. Interview introductory e-mail

The below e-mail message was sent to interviewees latest one week prior to their agreed interview date. All interviews were conducted in Finnish and so the e-mail was sent in Finnish. Below is also a translation for the email.

”Hei XXXX

Kiitos vielä kerran kun löysit aikaa opinnäytetyöni auttamiseksi!

Työni siis koskee datan laadun seurannan kohdentamista: työn tuloksena tuottaisin pragmaattisen, mahdollisimman geneerisen viitekehyksen sille, mitä liiketoimintakriittisen datan (master datan) osajoukkoa tulisi seurata laadun osalta, jotta saataisiin mahdollisimman konkreettisia ja hyödyllisiä tuloksia.

Kuten lupasin, tässä tulee hieman lisätietoa työn ja haastattelun sisällöstä. Liitteenä on työni alustava johdantokappale ja hieman taustatietoa työstä. Opiskelen englanninkielisellä linjalla AMK:ssa ja täten työni on englanniksi. Haastattelu voidaan pitää kummalla kielellä tahansa.

Haastattelun runko olisi seuraavanlainen (tosin keskustellaan aiheesta hyvin vapaamuotoisesti):
1. Pyytäisin sinua **tutustumaan lähettämääni materiaaliin** (liitteenä oleva johdantokappale työstäni) ja **aihealueisiin** (listattuna alla) etukäteen muutaman hetken, jos mahdollista.
2. Haastattelun alussa **näytän alustavaa ideaa** siitä, minkälaista tulosta työllä haetaan / työlle ollaan ajateltu – tämä on vain keskustelun avaamiseksi.
3. **Keskustelemme** aiheesta ja näkemyksistäsi.
4. **Näytän pidemmälle viedyn version** tuloksesta – Datpron asiantuntijoiden antaman tiedon ja työni teoriaosuuden perusteella rakennetun viitekehyksen.
5. **Keskustelemme** tuosta datan osajoukosta, sen todenmukaisuudesta mielestäsi ja sen mahdollisesti herättämistä lisäajatuksista.

Aihealueita pohdittavaksesi ennen haastattelua:

- **2-3 tärkeimmän liiketoiminta-alueen määrittäminen** (työssä nimellä ”business domain”, jolla voidaan tarkoittaa prosessia, funktiota, jne.)
- **2-3 tärkeimmän data-alueen määrittäminen** (työssä nimellä ”data domain”, jolla tarkoitetaan datakokonaisuuksia kuten asiakasta, tuotetta, jne.)
- Yllä mainittujen liiketoiminta- ja data-alueiden sisällä **tärkeimmät data-attribuutit tai attribuuttikombinaatiot**, jotka olisivat kriittisiä keskisuurten ja suurten yritysten datan laadun seurannan kohteeksi.
- Ja **perusteluja**, minkä takia juuri nuo attribuutit ovat mielestäsi tärkeimmät tähän viitekehykseen.

Jos sinulla herää kysymyksiä tästä, otathan yhteyttä ja vastaan mielellään niihin!”

The same in English:

“Hello XXX,

Thank you once again for finding the time to help with my thesis!

To re-cap, my work is about the targeting of data quality montoring: the deliverable of the work would be a pragmatic, as generic as possible, framework for what subset of business-critical data (master data) should be monitored for data quality to achieve as concrete and beneficial results as possible.

As promised, here is some further information on the content of the thesis and the interview. Attached please find the preliminary introduction of my thesis and some background information on the work. I study in English at a Univerity of Applied Sciences and therefore my thesis is in English. The interview can be conducted in either language.

The structure of the interview would be as follows (although we will discuss the issue quite freely):

1. I would kindly ask you to **familiarize yourself with the materials I provided** (the introductory chapter and background information attached) **and the topics for discussion** (as listed below) in advance for a moment, if possible.
2. At the beginning of the interview **I will show a preliminary idea** of what kind of result is sought after / has been considered for the work – this is to act as a conversation starter only.
3. **We will discuss** the matter and your view.
4. **I will show a further developed version** of the re- a framework built based on information provided by Datpro experts and the theoretical part of my thesis.
5. **We will discuss** the subset of data presented above, verifying it and discussing any new ideas that the framework brought up.

The topics for you to consider prior to the interview:

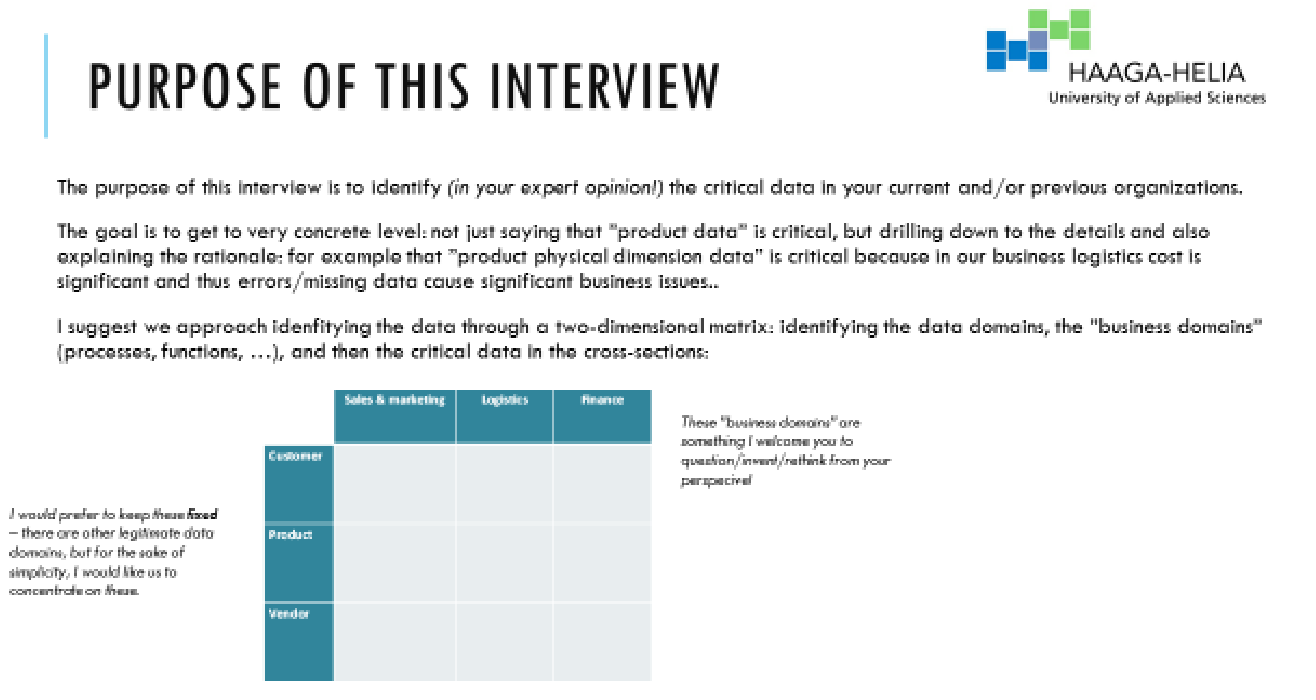
- **Identification of 2-3 key business domains** (functions, processes, etc)
- **Identification of 2-3 key data domains** (customer, vendor, etc.)
- **Key data attributes or set of attributes** within those business and data domains that are critical for a medium-large to large organization to monitor for data quality
- **Reasons** why those data elements would be the most beneficial to monitor

If you have any questions regarding this, feel free to contact me and I will be happy to answer them!”

Attachment 2. Interview template / Target matrix for data quality monitoring

A short PowerPoint presentation was prepared for basis of discussion for the semi-structured interviews conducted as part of the field study. Below are the slides prepared on the target matrix shown to interviewees. The first slide explains the approach in general, the second gives a few examples of from which angles the data attributes or sets of attributes can be considered critical for monitoring and the third features a further developed model of the target dataset based on Datpro expert experiences.

Slide 1



PURPOSE OF THIS INTERVIEW

HAAGA-HELIA
University of Applied Sciences

The purpose of this interview is to identify *(in your expert opinion!)* the critical data in your current and/or previous organizations.

The goal is to get to very concrete level: not just saying that "product data" is critical, but drilling down to the details and also explaining the rationale: for example that "product physical dimension data" is critical because in our business logistics cost is significant and thus errors/missing data cause significant business issues..

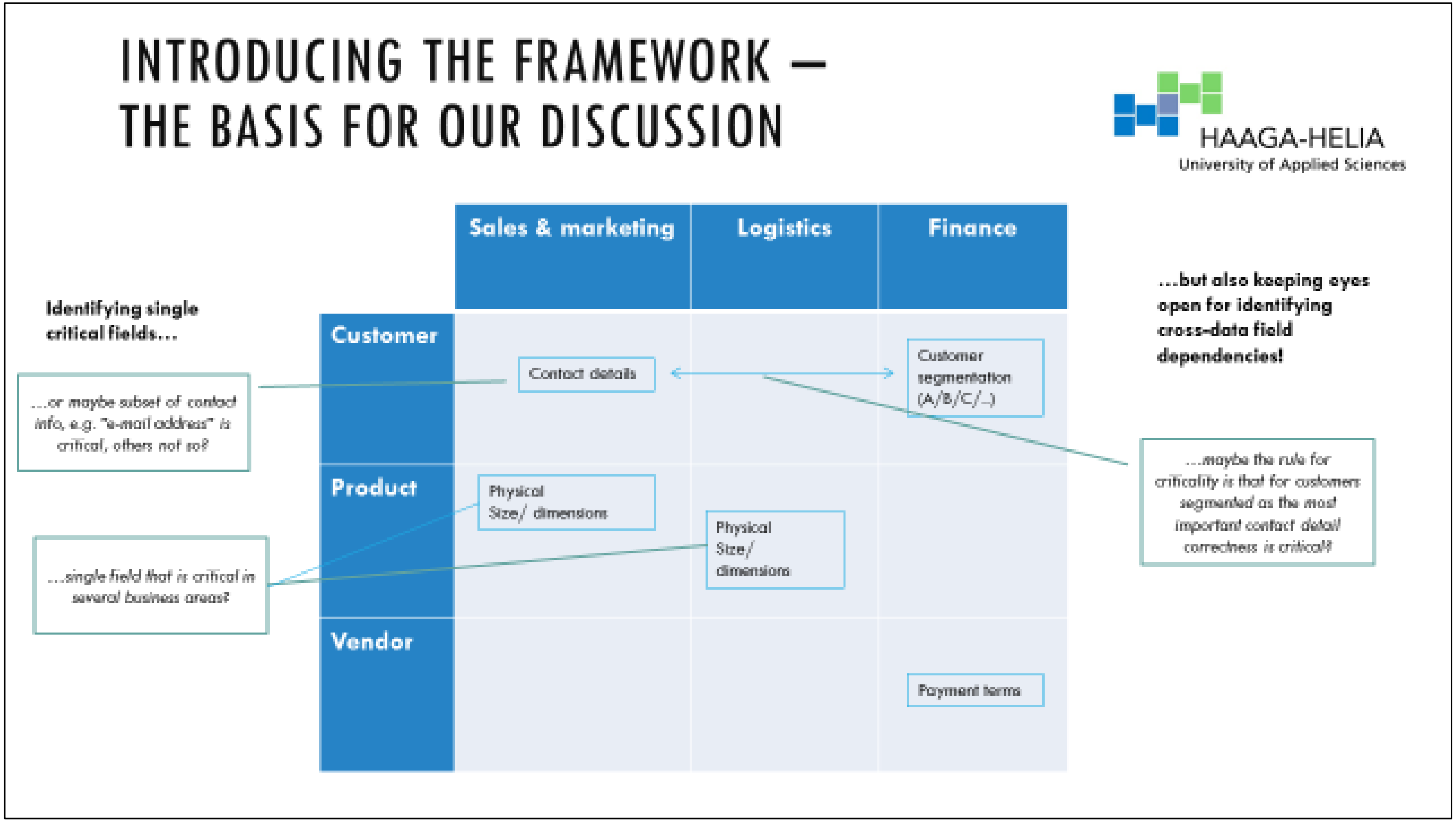
I suggest we approach idenfitying the data through a two-dimensional matrix: identifying the data domains, the "business domains" (processes, functions, ...), and then the critical data in the cross-sections:

	Sales & marketing	Logistics	Finance
Customer			
Product			
Vendor			

These "business domains" are something I welcome you to question/invent/rethink from your perspecive!

*I would prefer to keep these **fixed** – there are other legitimate data domains, but for the sake of simplicity, I would like us to concentrate on these.*

Slide 2



INTRODUCING THE FRAMEWORK – THE BASIS FOR OUR DISCUSSION

HAAGA-HELIA
University of Applied Sciences

Identifying single critical fields…

…or maybe subset of contact info, e.g. "e-mail address" is critical, others not so?

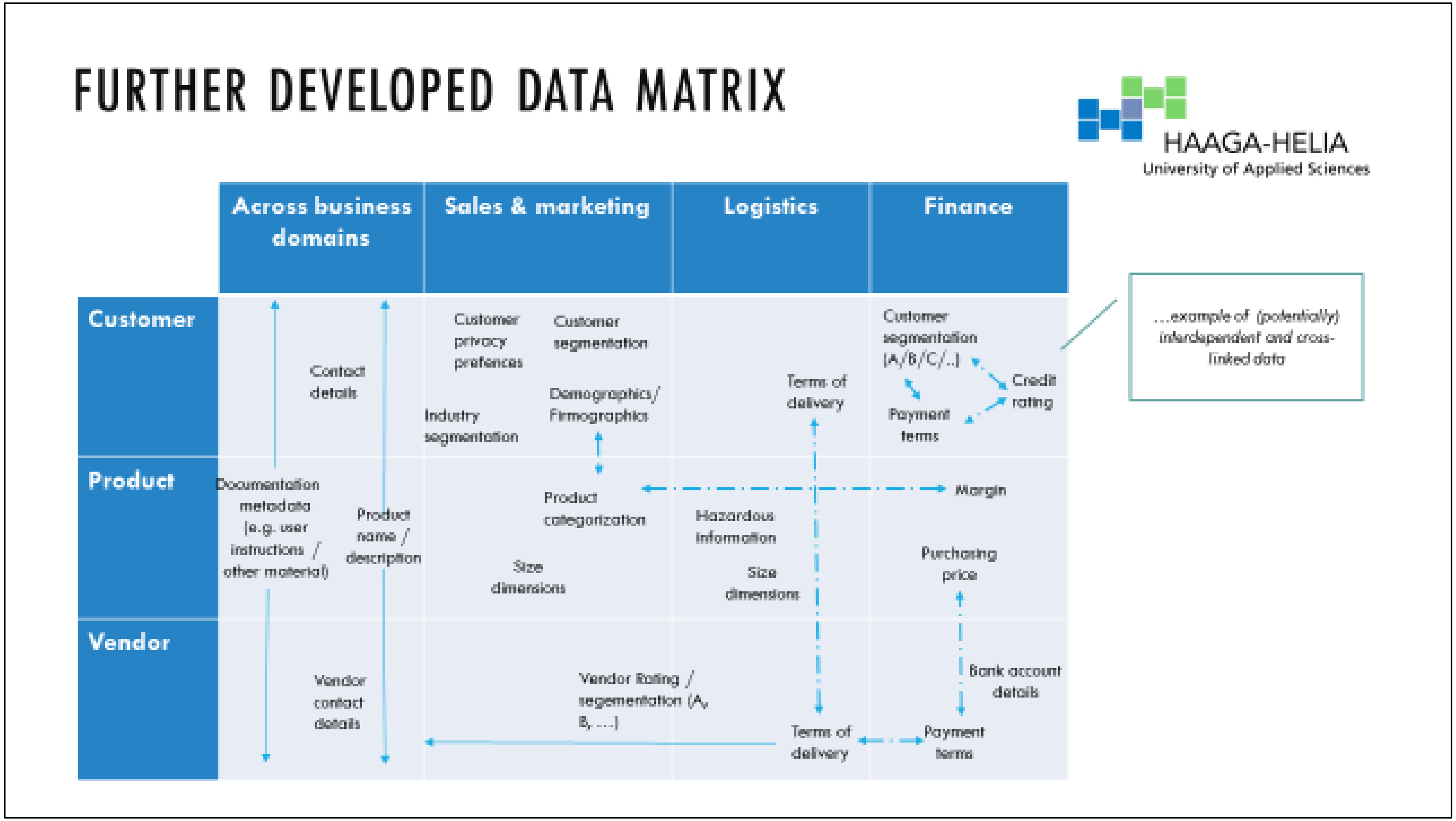
…single field that is critical in several business areas?

	Sales & marketing	Logistics	Finance
Customer	Contact details		Customer segmentation (A/B/C/…)
Product	Physical Size/ dimensions	Physical Size/ dimensions	
Vendor			Payment terms

…but also keeping eyes open for identifying cross-data field dependencies!

…maybe the rule for criticality is that for customers segmented as the most important contact detail correctness is critical?

Slide 3



FURTHER DEVELOPED DATA MATRIX

HAAGA-HELIA
University of Applied Sciences

	Across business domains	Sales & marketing	Logistics	Finance
Customer	Contact details	Customer privacy prefences Customer segmentation Industry segmentation Demographics/ Firmographics	Terms of delivery	Customer segmentation (A/B/C/..) Credit rating Payment terms
Product	Documentation metadata (e.g. user instructions / other material) Product name / description	Product categorization Size dimensions	Hazardous information Size dimensions	Margin Purchasing price
Vendor	Vendor contact details	Vendor Rating / segementation (A, B, ...)	Terms of delivery	Bank account details Payment terms

...example of (potentially) interdependent and cross-linked data

Attachment 3. Interviewee profiles

For the purposes of establishing the expertise of the interviewees, a small introduction of each of the 10 interviewees is collected here. This is to verify that they are, as mentioned in the thesis, knowledgeable in the field of data and data quality, and qualified to make educated assumptions on the subject of this thesis.

The commissioning party representative that has been the most active in providing information for this thesis is introduced below. The rest of the interviewees that are external to Datpro, are introduced in a similar way by current and previous titles as well as data-relevant experience, but kept anonymous. All interviewees are currently employed by well-known medium-large to large Finnish-owned but internationally operating organizations.

Kimmo Kontra

- Currently: Lead Advisor, Owner / Datpro Oy
- Previously: Manager / Accenture
- Data-relevant experience: Kimmo has more than 15 years of data governance and data management experience in a wide variety of industries and projects, has given speeches on the subject at MIT (Massachusetts Institute of Technology) and is the Vice Chairman of the Board at DAMA Finland, an international data management association's Finland division. Has Master's Degree from Helsinki University of Technology.

Interviewee	**Date and time of interview**	**Current position of interviewee**	**Previously held positions**	**Description/comments of interviewee**
1	24.6.2014, 10:00-11:00	Consultant at an IT consulting company		Interviewee is most probably among the first ones in Finland who has concentrated in Data Quality in his Master Thesis (2006) where he worked under one of the pioneers in quality management discipline in Finland, Paul Lillrank. Has worked as a consultant in various architecture and process re-engineering roles in multiple projects where data has been a strong influence. Project background is in public administrative, financial and industrial & construction industries. Has a Master's Degree from Helsinki University of Technology.
2	25.6.2014, 10:00-11:00	IT Development Manager at a retailing conglomerate	Enterprise Architect & Master Data Architect at retailing conglomerate, Solution Architect & Senior Specialist Information Architect at communications and IT corporation	Interviewee has been involved in renewing and reinventing the whole solution base for Finland's biggest retailing conglomerate, heavily concentrating on master data management and governance. Has worked in master data since the beginning of the 21st century, dealing with de-

				mand-supply planning, data modelling, data warehousing, etc. Worked as a project/programme manager in several data projects in many major Finnish companies. Has two Bachelor's Degrees in Information Technology and Computer Science.
3a	25.6.2014, 14:00-15:00	Business Solutions Director at a major state-owned company	Supply Chain Planning Director, Development & Planning Director, Logistics Director, Development Manager Material and Selection Management at major state-owned company	Interviewee's main responsibilities have been for years in the domains of Supply Chain Management and Logistics, but he has demonstrated an interest in the role of data management in the company's business. He and his team built a data management approach for the company when it decided to move to a single ERP in 2007. Afterward, interviewee has been building a practice where traditional Master Data and Data Quality, ERP's data support, and analytics/business intelligence side of Data Management converge. Has a Master's degree from Helsinki School of Economics and Business Administration.

3b	25.6.2014, 14:00-15:00	Global Master Data Team at a state-owned company	Master Data Concept Developer, Conversion Team Lead at a building and home improvement corporation, SAP Consultant at a major consulting company, Data Coordinator at a global chemical corporation	Interviewee works as part of the Global Master Data team and has years of experience as a master data consultant in various projects. Has a Master's Degree at Tampere University of Technology.
4	26.6.2014, 9:00-10:00	Director of SAP Management at an engineering and service company	Enterprise Architecture Director at a metal and consumer brands company, Retail Processes Senior Manager and Solutions Architect at a communications and IT corporation	Interviewee has a strong background in enterprise architecture in three major Finnish listed corporations. He is familiar with major corporate applications and their role in overall architecture, business and IT alike. He has served in major business units' management teams. His main interests have always included information architecture, data quality, master data and master data processes. He has participated in setting up comprehensive data management concept and practice in one of his employers. Has a Master's degree from Helsinki University of Technology.

5	30.6.2014, 13:00-14:00	Master Data Architect at a municipal co-operative network	Master Data Manager at a metal and consumer brands company, Data Manager, Master Data Solution Architect & Metadata Concept Owner at a communications and IT corporation, Senior Specialist at a major consulting company	Interviewee has over 10 years of experience in data management, having concentrated on master data management, MDM processes and governance in the last 5 years. Has started from hands-on data work with meta and reference data at a time when not much was known about master data, moving up towards total data governance work. Has a Master's degree from University of Helsinki.
6	1.7.2014, 10:00-11:00	Master Data Manager/Enterprise Architecture in a stainless steel corporation	Various information management positions at a stainless steel corporation	Interviewee has been awarded as a pioneer of master data management in Finland by the Data Management Association Finland (DAMA Finland). Started working on master data in 1999 when no one knew about master data and there certainly was no documentation. Between 1999-2004 created processes and ways of working for master data that are still in use today for the organization in question. Retired this year after 15 years' career in data management and governance.

7	2.7.2014, 10:00-11:00	Master Data Management Manager at a pharmaceuticals and health care products company	Information Management Senior Expert at a data management consulting company, Information Architecture Manager & Master Data Concept Owner at a global chemical corporation	Interviewee has a solid background in Master Data, having worked solely in data and information management since 2007. Has experience from hands-on manual data work to designing and delivering data governance in organizations. Was involved in data quality software development project where designed data quality processes from scratch on top of plenty of data quality analysis work previously. Will lead global master data strategy development in current organization. Has a Master's degree from Helsinki University of Technology.
8	7.7.2014, 10:00-11:00	Development Manager at a major banking and insurance company	Management Consultant at a consulting companies, Contract Administrator at a telecommunications company	Interviewee has a background in management consultancy and in banking & finance industries. He has strong background in managing certain core banking applications from business perspective, though in practice acting as a bridge between business and IT. In addition, he has a key role in enterprise-wide customer master data maintenance (both consumers and corporate customers). Lately he has concentrated in "digital enablement", building digital capabilities that the enterprise must cross in

				order to be able to serve the customers in new way, involving data heavily. Has a Master's degree from Helsinki University of Technology and is completing a Master's degree at Helsinki School of Economics and Business Administration.
9	9.7.2014, 10:00-11:00	Head of Division (Healthcare, laboratory) at multi-industry conglomerate	Regional Operations Director, Facility Administrator	Interviewee has a strong background in sales and marketing of multi-national healthcare industry. Has been involved in strategic planning and competitive analysis endeavours at very demanding companies. Has a Master’s degree from Helsinki School of Economics.