LUT School of Business and Management
Master's thesis, Business Administration
Business Analytics

Managing data quality with data governance - a qualitative study based on semi-structured interviews of project experts

16.6.2021
Author: Niko Kauria
Supervisors: 1. Prof. Pasi Luukka
2. Post-doc. Jyrki Savolainen

ABSTRACT

Author: Niko Kauria
Title: Managing data quality with data governance - a qualitative study based on semi-structured inter-views of project experts

Year: 2021 **Place:** Helsinki

Master's thesis
Lappeenranta-Lahti University of Technology LUT
School of Business and Management
Degree Programme in Business analytics
66 pages, 6 figures, 4 tables, and 2 appendices
Examiners: 1. Prof. Pasi Luukka, 2. Post-doc. Jyrki Savolainen

Keywords: data, data governance, data quality, data management, data ownership, IT, data management office

This thesis studies the relationship between different data governance practices and changes in data quality. This was done by researching the current literature on the topics on hand and interviewing experts on the field, whose daily work is to design and implement these processes for the big organizations in Finland. As this relationship is constructed of the concepts of data governance and data quality, study also focused to understand the major aspects of these two, and review the effect these aspects have on both concepts. Final conclusions were constructed by comparing the literature and interviews, and aiming to find similarities and differencing subjects.

The relationship between the data governance and data quality was strongly associated by both the literature and the participants. Nevertheless, by combining the findings, it was shown that the lack of metering before governance is causing the issue of authenticating the amount of data quality improvements that can be achieved by implementing data governance process. Identified issues in the governance process affecting data quality were the lack of securing sufficient resources, committing participants, and not assigning the ownership of the process to the business unit. The difference between current literature and points found out in the study was that data governance is shifting more toward separate data management offices, where the daily work on data quality and governance is more agile than presented in the literature. Although, governance responsibilities are still mainly centralized on the company level, making possible changes rather inflexible, which in a high-paced world is a challenge and needs to be changed towards comprehensively agile way of working.

TIIVISTELMÄ

Tekijä: Niko Kauria
Työn nimi: Tietojen hallinnan vaikutus datan laatuun – laadullinen tutkimus asiantuntijoiden näkemyksistä aiheeseen

Vuosi: 2021 **Paikka:** Helsinki

Pro Gradu -tutkielma
Lappeenrannan-Lahden teknillinen yliopisto LUT, Kauppatieteet
66 sivua, 6 kuvaa, 4 taulukkoa ja 2 liitettä
Tarkastajat: Professori Pasi Luukka, Tutkijatohtori

Hakusanat: Tietojen hallinta, datan hallintamalli, tiedon laatu, datan laatu, datan omistajuus, IT, datan hallinnan yksikkö

Tämän Pro Gradu-tutkielman tarkoituksena oli tutkia datan hallintamallien yhteyttä datan laadun paranemiseen. Tämä toteutettiin tutkimalla saatavilla olevaa kirjallisuusaineistoja sekä haastattelemalla kyseisen alan ammattilaisia, joiden päivittäiset työtehtävät koostuvat näiden hallintamallien suunnittelusta ja implementoinnista muun muassa suuriin yrityksiin Suomessa. Koska tätä suhdetta ei pysty selittämään ymmärtämättä hallintamalleihin ja datan laatuun liittyviä aspekteja, näiden syvällisempi tarkastelu on myös otettu osaksi tutkimusta. Lopulliset johtopäätökset muodostettiin vertaamalla kirjallisuudesta sekä haastatteluista nousseiden aiheiden eroja ja yhtäläisyyksiä.

Sekä kirjallisuudessa että haastatteluissa kävi ilmi, että datan hallintamallien ja datan laadun välille muodostettiin selkeä suhde. Kuitenkin ongelmaksi muodostui mittaroinnin puutteellisuus, joka esti numeerisen toteamisen siitä, kuinka paljon hallintamallin käyttöönotto parantaisi datan laatua. Havaitut ongelmat hallintamallien toteutuksessa olivat puutteet resurssien varmistamisessa, henkilöiden sitouttamisessa sekä prosessin omistajuuden osoittamisessa muualle kuin liiketoimintayksikön alaisuuteen. Suurimmat eroavaisuudet kirjallisuuden ja tutkimuksessa havaittujen asioiden välille muodostuivat, kun tarkasteltiin muutosta prosessin suorittamisessa: suuntauksena nykyisellään on perustaa erillisiä datan hallinnan yksikköjä, joiden sisällä työskentely on kevyttä ja ketterää, kun taas aiemmissa julkaisuissa erilaiset hierarkiat ovat olleet läsnä vahvemmin. Kuitenkin datan hallinta toteutetaan pääasiassa vieläkin kootusti koko yrityksen tasolla, jolloin mahdolliset muutokset ovat jäykkiä toteuttaa, mikä aiheuttaa nykyisessä nopea tempoisessa maailmassa haasteita, ja pitää pystyä muuttamaan tulevaisuudessa kohti kokonaisvaltaisesti ketterämpää ratkaisua.

ACKNOWLEDGEMENTS

While writing these words, I am at the same time both relieved and slightly sad. The whole process of studying and this thesis is now finalized, and the five years spent in Lappeenranta were over in a blink. First and foremost, I must thank Post-Doc. Jyrki Savolainen, whose help contributed more than a lot towards finishing this last project. I would also like to thank TietoEVRY for being a flexible employer by providing both the topic for the thesis and possibility to work part-time. Thank you also for the experts, with whom I had the pleasure to learn and discuss in the interviews.

Of course, this whole study path would not have been possible without the support of my family: parents for teaching that hard work pays off eventually, Aino for always supporting along the writing process, and friends for making everything just a little bit funnier.

Helsinki,
Niko

Table of contents

1 Introduction

This thesis focuses to study data governance and data quality and their relationship from the data governance point-of-view, where the main topic is the consequences to data quality from data governance actions. The scope also includes studying the key factors and issues in governance implementation, which affect data quality.

1.1 Background of the study

It is no surprise, that the amount of data organizations have, is significant and keeps on growing constantly as it is gathered besides with traditional forms and manual records, also with automated processes from the internet and through different monitoring systems. This 'Big Data' often consists of multiple sources and has no clear schema for the data, and it is estimated that up to 80% of organizations' data consist of unstructured data and they have no possibilities to handle, process, and protect it (Halevy, 2005, p. 53; Rizkallah, 2017). At the same time, companies are starting to understand that owning the data is not valuable per se, but the proper usage is: data is seen as a company asset, which should be invested in (Dyché and Levy, 2006, pp. 148–149; Abraham, Schneider and vom Brocke, 2019, p. 426). This means that one driver to improve the quality of that asset originates from the business purposes, while another one can be the law, such as GDPR or Data Protection Act 1998 (Al-Ruithe, Benkhelifa and Hameed, 2018, p. 18). To point out an example business case, Dyché and Levy (2006, pp. 71–72) claim that non-integrated data is frequently the cause of cost and time overruns across industries

In many organizations, it is a known fact or at least presumption, that the quality of data is bad. Mainly this is due to faulty processes in the data pipes or then more commonly, the quality is not good even when created and until it is measured at the of usage, the bad data has enough time to build up and to corrupt also everything else (Redman, 2013, p. 4). Another common case might be that data is not even collected to suit the purpose it is needed (Downing, 2003, p. 836). The list can go on and on, and to address these different kinds of data quality issues, it is needed to understand the different dimensions of the whole concept.

The old belief that data quality and governing data belongs to the IT department seems to stay still somewhat strong (Friedman, 2006). But since the issue is that even though data is technically corrected, if it doesn't suit the purpose, it is not quality data. This means that business needs to be involved more and more throughout the whole process of data governance. The ownership and accountability are seen as key components of data governance (Griffin, 2005, pp. 49–51; Khatri and Brown, 2010, p. 149; Abraham, Schneider and vom Brocke, 2019, p. 426), but there are many more, such as roles and responsibilities (Cheong and Chang, 2007, p. 1006) and master data management (Berson *et al.*, 2010, pp. 406–407; Koltay, 2016, p. 309). Data governance itself is a rather widely studied topic, but for this study, the focus is more on the relationship between data governance and data quality. There are often remarks and claims, such as Panian (2010, p. 941), that data governance and data governance framework can address issues in data quality. However, there seems to be a lack of studies, that actually measure or examine the changes in quality. And as the importance of high-quality data is rising, it is interesting and important to prove that these two do have a significant connection, and thus can be pointed out, that with proper governance, proper data quality can be achieved.

1.2 Research questions and objectives

The objectives for this study can be divided into three. Firstly, there is the relationship between the governance actions and data quality and secondly, there are the specific actions and issues in the governance process, which affect the quality. Lastly, from a business perspective, finding the important stages of the governance process will provide value for the project planning, and can be seen as one of the objectives for the study. As the topics can be divided into different aspects; the effect of data governance actions on data quality and the issues in data governance implementation affecting data quality, there is a need for two different questions. These questions are:

According to the literature, how data quality can be measured, and how organizations can improve their data quality by establishing a data governance strategy?

What kind of issues there are in the implementation of governance affecting data quality based on the expert interviews?

In addition, this study reviews how data governance and data quality are presented in the current literature, and whether they differ from the understanding of experts working in the field. The research question for this topic is:

Are there any significant trends in the data governance process or data quality concept, which are not included or presented differently in the current literature?

It should be noted that there are some biases in the question: in the first one, there is an assumption that governance will improve data quality. However, this can be reasoned through previous literature in which none of the publications implicated that the relationship would be negative towards data quality. And to follow the objective, it is needed to assume that there is a relationship.

1.3 Research methodology

The duration of the study was in total 8 months, where forming the theoretical framework took five months, conducting all the interviews one month and transcribing and analysing the results two months.

For the literature review, used sources consist of scientific articles, books, reports, and as well conference proceedings found from the school library and through search engine libraries, such as Google Scholar. As conference proceedings don't fully fulfil the requirements for scientific releases, the information presented in them is also tried to verify with multiple different sources. Methods and data used in the empirical study are further elaborated in chapter 4.

1.4 Structure of the thesis

This thesis is constructed on six different chapters with each chapter having respective subchapters. By following the guidelines for traditional research, this thesis includes both the theoretical part introducing current knowledge in the previous studies and an empirical part, which tries to present evidence on the previous literature findings and also find something additional.

Chapter 2 includes the literature definitions of data governance and data quality and presents the main concepts regarding these two topics. For chapter 3, the scope switches to review the relationship of these two themes in the current publications. This also includes a brief report of

the previous studies concerning the same topic as this thesis. After the theoretical scope is formed, the focus is shifted towards the empirical part and the used methods and participants are presented in chapter 4. It also includes a review of the results' validity, reliability, and implications. Furthermore, chapter 5 presents the findings and discusses them with topics found in the literature review. Finally, the sixth chapter gathers the findings into conclusions and presents them and possible topics for future research.

2 Literature review

While conducting the study, it was soon clear that there are several different approaches, and common standards for data, data quality, and data governance are still not fully unified in literature. At the most, this was true for data governance which has various definitions available. These are presented briefly, and most commons are selected to be used in this study. It is worth noticing that the chosen definitions are not meant to create a new standard but rather to ensure that the reader will understand what this study means with each phrase.

2.1 On the definition of data

While the Merriam-Webster dictionary states the word '*data*' to be "*factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation*" and Oxford dictionary as "*characteristics or information, usually numerical, that are collected through observation*", from more technical perspective 'data' can be seen as a data set and data units within qualitative and quantitative variables to represent the population (Australian Bureau of Statistics, 2020). This latter approach suits better the purpose of this study. In business, data is not seen as just as described above but also as an asset, that can create value to the company and should be invested in (Dyché and Levy, 2006, pp. 148–149). From the data governance perspective, many literature reviews such as Abraham, Schneider and vom Brocke (2019, p. 426) see data as a strategic enterprise asset, which should be valued and cared for.

In DIWK, data, information, wisdom, and knowledge pyramid or hierarchy (presented in figure 1) data is needed to retain information, and furthermore that information can be turned into knowledge and wisdom (Rowley, 2007, pp. 164–168; Redman, 2009). The proportion sizes represent the actual amount needed from the step below to create the step above. Distinguishing differences and connections between data and information is crucial when later discussed the properties of data quality.

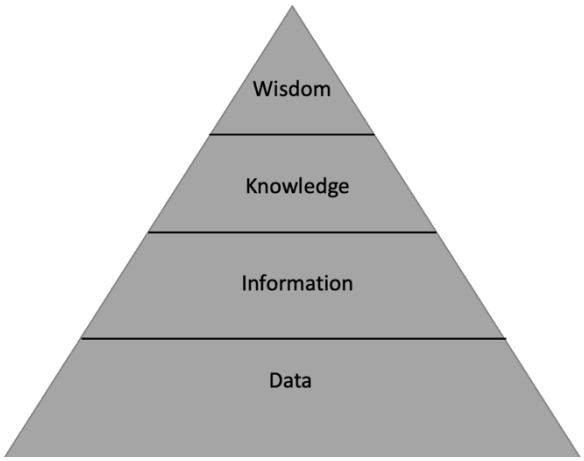

Figure 1. DIWK pyramid (adapted from Rowley, 2007. 164)

2.2 Data quality

In the reviews, data quality is often divided into smaller dimensions, where data quality dimension can be defined generally as an aspect or a feature, that is used to classify information and data quality needs by providing a way of measuring the quality (McGilvray, 2008, p. 297). In the study conducted by Jayawardene, Sadiq and Indulska (2015), data quality was divided into eight different categories: Completeness, Availability & Accessibility, Currency, Accuracy, Validity, Usability & Interpretability, Reliability & Credibility, and Consistency. Further these could be divided into 127 smaller sub-categories. Sidi, Shariat, Affendey, Jabar, Ibrahim & Mustapha (2012) categorized dimensions into 40 different groups and concluded, that to achieve high-quality data, one must study all the dimensions of data and their relations to one and another. For example, for the data to be usable, both the accuracy and the timeliness should be on a high level (Barone, Stella and Batini, 2010). Even though we can measure data quality through these dimensions, it needs to be understood that the data is not static hence the quality isn't static, and it even shouldn't be. It is also important to measure the ongoing improvements in quality over time and record changes that increased it, and replicate those tactics broader in the organization (Dyché and Levy, 2006, p. 78). As stated by both Downing (2003, p. 836) and Olson (2003, pp. 24–25), data needs to suit the purpose it is used, which needs to be known before making any changes to data.

There is no common consensus among researchers about the most important dimensions and meanings do vary between writers (Scannapieco and Catarci, 2002, pp. 1–2). To assess the data

quality later in the research, it is needed to decide which dimensions are taken into account and also define what is meant with each dimension and/or dimension group. Many of these definitions are stated already in the late 1900s and early 2000s but are still used in topic literature.

2.2.1 Accessibility and availability

Accessibility according to Stvilia, Gasser, Twidale and Smith (2007, p. 1729) can be measured with the speed and ease of locating and obtaining needed information, while availability refers to the amount of time data is accessible when needed e.g., server or source is accessible (Ranganathan, Iamnitchi and Foster, 2002). Availability is also measured during service breaks for which e.g., MS SQL server is providing node replication with the term 'High availability', where one of the nodes is always available for use.

Issues in these dimensions can be one of the triggers in the data governance process: constantly growing volume of data may trigger performance issues causing accessibility of data to decrease. Or in the opposite scenario, data is too easily accessed from unauthorised use. In case of issues in the system or for example user errors, it is necessary to have backups of the data available and accessible.

2.2.2 Coverage and completeness

Data is a way to model real-world phenomenon in a measurable format. For data to be valid, it needs to cover all of it but on the other hand, it needs to maintain the purpose to have a reasonable amount of data to work with (Price and Shanks, 2005, p. 10; Eppler, 2006, p. 83). After one can cover the whole entity, next it is needed to make sure that the recorded data is complete. Completeness can be defined to have all the values in the data I.e., no values are missing (Kimball and Caserta, 2004, p. 115; Gatling, Stefani and Weigel, 2012, p. 345). Furthermore, completeness can be viewed as a unit measure, where the level of completeness can be given a value (Redman, 1997, p. 257; Scannapieco and Catarci, 2002, p. 11). It is worth noticing that 'null' -values can either improve or deteriorate completeness depending on the attribute style (Redman, 1997, p. 262; Loshin, 2001, p. 218).

2.2.3 Accuracy

Data accuracy can be measured by comparing the data values to the identified source of correct values and calculate the ratio of this (Loshin, 2001, p. 217; Olson, 2003, pp. 24–25; McGilvray,

2008, pp. 31–32). However, in the real world identifying and finding the real values usually requires huge manual steps in an otherwise automated process (Loshin, 2001, p. 217).

To measure accuracy correctly, one must understand the context: Olson (2003, pp. 24–25) highlights that a database with records of physicians in the area of Texas with 85% of accuracy, is poor for informing of new law changes but would be excellent for a technical manufacturer, who is looking for potential customers. What Olson is trying to point out is that even though the accuracy level remains the same, the usage can define whether it is poor or excellent.

In addition to accuracy at one point in time, it needs to retain the same level of accuracy across all the records regarding time perspective. This can be described either by data integrity (Zviran and Glezer, 2000) or with the common definition of reliability (Trochim, 2020). This dimension is highly linked with consistency which is presented in chapter 2.2.7.

2.2.4 Currency and timeliness

Askham *et al.* (2013, p. 10) see timeliness as the difference between the reality of a specific time and the 'reality' data represents of that time, where Loshin (2001, pp. 115–116) describes this with the term 'currency'. In addition, both Loshin (2001, pp. 115–116) and Fan, Geerts & Wijsen, (2012, p. 71) use 'currency' to measure, how correct the data is despite of the changes in time. This can be a major issue in today's high-paced rhythm, where information and data are tracked constantly, and there can be multiple records from a same situation or a same event but from different time periods, which have all become obsolete (Fan, Geerts and Wijsen, 2012, p. 71).

The meaning of timeliness is identified differently in literature sources. While both Loshin (2001, pp. 115–116) and Fan, Geerts and Wijsen (2012, p. 71) determinate timeliness as the measure between the event happening and data record available, Batini & Scannapieca (2006, p. 29-30) and Heinrich, Kaiser & Klier (2007, p. 5) measure timelines with how current and up-to-date the value still is. To prevent misunderstanding later in this research, Loshin's and Fan et al.'s definitions will be used, where 'currency' measures how to correct data is despite changes in time, and 'timeliness' is valued based on the time difference between incident and record available.

Both the timeliness and currency are dependent on the maintainability of the system: how difficult or resource-consuming it is to organize and update data on an ongoing basis (Eppler, 2006, p. 84). The ability to import new data is also highly linked to these dimensions.

2.2.5 Validity

Validity in general is measured by the probability of how a certain statement represents the real world (Rich *et al.*, 2011, p. 105). From the data point of view, this interferes with accuracy-dimension on some level but differentiating different data quality dimensions is difficult, and not even meaningful. Although, in regard to data quality, validity is defined slightly differently: validity in data is defined on multiple levels: at dataset level validity is measured whether the data is collected in regards to the meant purpose at the right point of time (Downing, 2003, p. 836). At the data element level, the record needs to be within a valid range of values and calculated values must be derived from correct formulas or derivation rules (English, 2009, p. 123). Furthermore lower, absence or poor quality of metadata can negatively affect the validity and quality of data itself (Price and Shanks, 2005, pp. 8–9).

2.2.6 Interpretability

Data is presented in an intangible manner when there is no possibility of misunderstanding recorded values, whether the interpreter is a person or machine (Price and Shanks, 2005, p. 3). Machine learning and other algorithms are more and more used and it puts pressure on interpretability: if values are linguistic, they need to meaningful and clearly understandable but linguistic can be better in terms of interpretability versus numeric i.e., performance valuation with values of ‘poor’, ’good’ and ’excellent’ have slighter changes of being misunderstood than values from one to three (Redman, 1997, p. 262; Guillaume, 2001, p. 32).

2.2.7 Consistency

Data can be called consistent when two or more things do not conflict with each other i.e., there cannot be other values for level 4 employer salaries than between $40 000 and $60 000 (Redman, 1997, p. 259; Gatling, Stefani and Weigel, 2012, p. 32). In extended perspective, data attributes need to follow unique principles, where both index field and other unique fields only include unique values, meaning also that there cannot be any duplicate records (Loshin, 2001, p. 443; McGilvray, 2008, pp. 128–133; Askham *et al.*, 2013, p. 9). Data needs to be standardised

with regards to naming and structure of data elements (Bisbal *et al.*, 1999, p. 11). Representational consistency indicates that all entries for an attribute need to be in the same format (Scannapieco and Catarci, 2002, p. 11) for which the most known example is NASA's $125 million Mars Climate Orbiter, which was lost in space due using both imperial and metric units in preparation (Oberg, 1999). In addition to using the metric and imperial system, there are other cases where issues can arise, for example in date timestamp formats, where MM/DD/YYYY-format is used in the US, whereas in contrary Europe uses DD/MM/YYYY-format. There are ways, such as ISO 8601 format YYYY/MM/DD, which tries to harmonize these situations and ensure consistency.

2.3 Data governance

While practical implementation of data governance is out of the scope of this study, to study the relationship between data quality and data governance, the study must present also implementation process briefly. The need for implementing data governance derives from old belief that data quality still belongs to the IT department (Friedman, 2006), while it should be managed with corporate-wide practises from both business and IT with clear definitions of roles and responsibilities (Wende and Otto, 2007, pp. 1–2). In addition, combined and centralized data governance can benefit the economics of scale (Brown and Grant, 2005, p. 700), and Koltay (2016, p. 305) takes this even further by stating that data governance shouldn't be optional but rather precondition for repeatable and compliant practices. Data governance can be described by saying that it creates organization-wide standards and guidelines for data quality management (Wende and Otto, 2007, p. 2) or as a '*service that is based on standardized, repeatable processes and is designed to enable the transparency of data related processes'* (Koltay, 2016, p. 309). Furthermore, governance can be seen as a way to ensure that certain goals and objectives are assigned and resources are used in an efficient manner (Rau, 2004, p. 35). There are multiple similar but slightly different definitions and as Abraham, Schneider and vom Brocke (2019, pp. 425–426) and Al-Ruithe, Benkhelifa and Hameed (2019, p. 7) noticed while researching definitions for data governance, there yet doesn't seem to be any universally accepted standard. Abraham, Schneider and vom Brocke (2019, pp. 425–426) concluded their findings with their own definition: '*Data governance specifies a cross-functional framework for managing data as a strategic enterprise asset. In doing so, data governance specifies decision rights and accountabilities for an organization's decision-making about its*

data. Furthermore, data governance formalizes data policies, standards, and procedures and monitors compliances.'

There are some regulations affecting data governance protocols in relation to each industry fields, such as Data Protection Act 1998 (and GDPR in Europe), which drive the governance process to at least on a certain level (Al-Ruithe, Benkhelifa and Hameed, 2018, p. 18). There can be other drivers to start governance process, such as strategic, organisational, system-related, and cultural factors, which can be seen more as internal motivators and are pushed through to gain competitive advantage, while regulatory governance is a must (Abraham, Schneider and vom Brocke, 2019, p. 432). After establishing the key reasons, Abraham, Schneider and vom Brocke's (2019, p. 426) research introduce the following requirements for the governance process: it is a cross-functional project, which enables collaboration across all levels. It needs a framework, which is the base for structured and standardised management for the data and another one for the data itself, which includes policies, standards, and procedures. It determines the decision rights and accountabilities and also all possible actions made for data-related quests.

Implementation of a data governance plan is usually dependent on various things and is highly organization/case-specific. As Wende and Otto (2007, p. 8) present this in figure 2, there are multiple different factors affecting the possible option that should be taken. The process starts from contingency factors, such as firm size and structure, and advances into design parameters, and finally to model configuration.

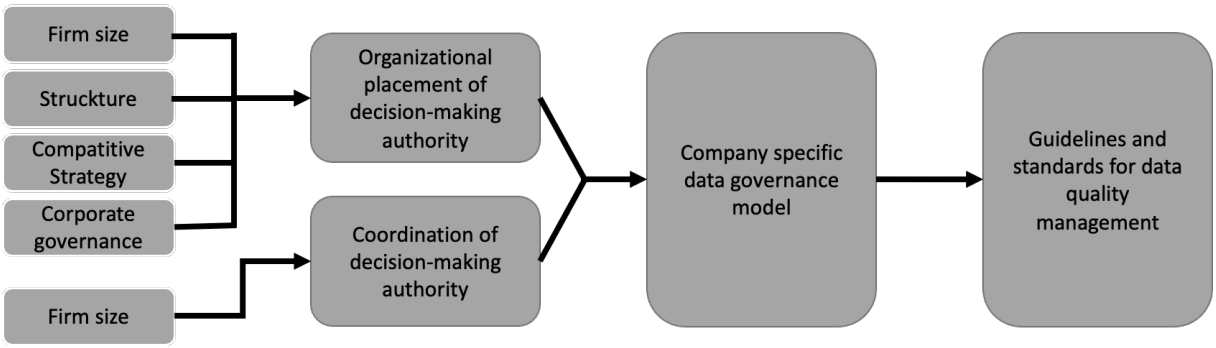

Figure 2. Governance case example (adapted from Wende and Otto, 2007, p. 8)

Soares (2013, p. 29) divides enterprise data governance policies under eight different sections, which include topics such as data ownership, master data management. This highlights that the universal standard is also missing for the composition of data governance.

2.3.1 Governance mechanisms

Governance mechanisms are presented in many literature sources, and e.g. Abraham, Schneider and vom Brocke (2019, pp. 427–428) conclude that they are used to plan and control data management activities and connect them with business and IT. Stripped down, these can be further categorized under three different mechanisms: procedural, structural, and relational (Borgman *et al.*, 2016, p. 4903).

Structural mechanisms consist of reporting structures, accountabilities, and governance landscapes including mainly discussion about the responsibilities and decision-making authorities (Bowen, Cheung and Rohde, 2007, p. 192; Borgman *et al.*, 2016, p. 4903). Roles in governance are discussed more widely in chapter 2.3.3.

Relational mechanisms present the collaboration of all the stakeholders and its importance is crucial especially at the beginning of the process (Van Grembergen, De Haes and Guldentops, 2004, p. 21; de Haes and van Grembergen, 2009, p. 135). By communicating with all the stakeholders about the importance and benefits of quality data and emphasizing overall awareness, one critical factor is achieved (Cheong and Chang, 2007, p. 1002). However, alone pure communication doesn't necessarily satisfy this, since users might not see the logic behind each policy, meaning there is a need for constant training procedures to ensure everyone's data competencies (Tallon, Short and Harkins, 2013, p. 196; Alhassan, Sammon and Daly, 2019, pp. 190–191).

Procedural mechanisms include manures that the data is held "securely and confidently, obtained fairly and efficiently, recorded accurately and reliably, used effectively and ethically and shared lawfully and appropriately", and are the same as Donaldson and Walker (2004, p. 281) list as NHS's goals for their governance program. These are also discussed with 'data processes' or 'data policies', as Alhassan, Sammon and Daly (2019, pp. 195–196) present and divide them into defining data regulations and access rights, implement them within the business systems

and finally monitor their compliance with both internal and external regulations. These mechanisms should also include metric solutions for long-term monitoring (Watson, Kraemer and Thorn, 2009, p. 438).

2.3.2 Governance scopes

Data governance programs can be executed on many different levels depending on the purpose, and as the name implies, organisational scope represents the extent of it and whether it's intra-organisational or inter-organisational (Abraham, Schneider and vom Brocke, 2019, p. 430). Tiwana, Konsynski and Venkatraman (2013, pp. 9–11) present their framework for governance scope with three simple questions: *Who, what, and how is governed,* and further illustrate these dimensions with a cube in figure 3 but they do want to emphasise that this should be only handled as starting point for theoretical discussion and not as an absolute theory.

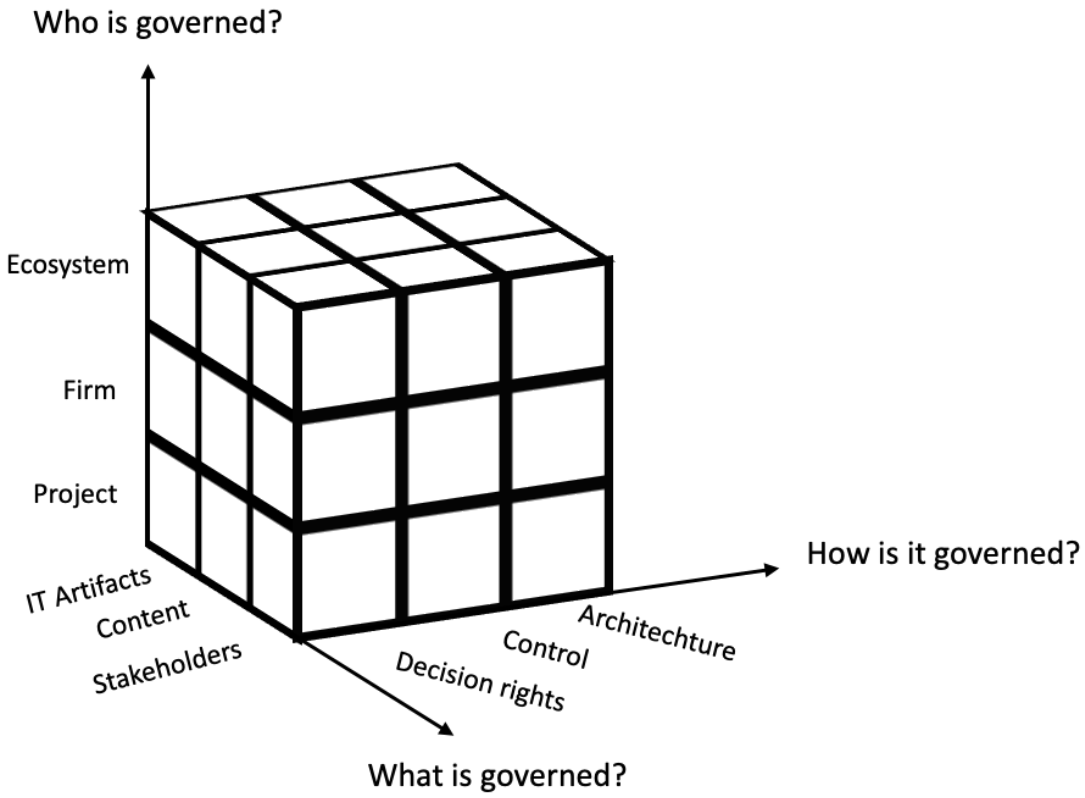

Figure 3. Governance Cube (adapted from Tiwana, Konsynski and Venkatraman 2013, pp. 9-11)

The first dimension and *Who* corresponds rather similar as the Abraham, Schneider and vom Brocke's (2019, p. 430) organisational scope where scale can start from a single project and

extent to ecosystem-level, such as hundreds of thousands of firms in Apple's IOS ecosystem. If we take a deeper look at it, in intra-organisational scope the process is conducted within the organization and in inter-organisational governance is shared between companies or even the ecosystem of firms (Abraham, Schneider and vom Brocke, 2019, pp. 430–431). While this might create information exchange issues, it can benefit with competitive advantage in overall (Rasouli *et al.*, 2016, p. 466). Though there are several scholars on the topic, such as Tallon, Short and Harkins (2013, p. 196) who see that to avoid misunderstandings in data policies, governance program should be a corporate-wide function from the start, as bottom-up approaches where business units develop their own policies tend to create complexity and inconsistency.

What -question is three-dimensional since it amplifies whether the governed topic is IT artifacts (i.e., hardware and software), the content (such as data), or the stakeholders involved in those. According to Tiwana, Konsynski and Venkatraman (2013, p. 10), discussion in the literature was mainly focused on the IT artifacts and stakeholders at the time, and they predicted that the focus will shift more on the data as big data and data quantities enlarge. Abraham, Schneider and vom Brocke (2019, p. 431) divide the data scope into traditional data and big data, where traditional data contains master data, transactional data and reference data, and big data as data with high enough variety, velocity, and volume (Loshin, 2008, p. 6; Ward and Barker, 2013, p. 1). For traditional data, governance measures mainly consist of data policies and processes (Loshin, 2008, p. 68) while for big data besides measuring and monitoring, the goal is also to find solutions for proper data storage, optimization, computing, communication, and data management (Al-Badi, Tarhini and Khan, 2018, p. 275). In addition, as the data amounts increase rapidly and most of it is machine-generated, identifying the sensitive data and establishing policies of its use as well as data retention and deletion planning will play a significant role (Morabito, 2015, p. 89).

Finally, the *How* argues which mechanisms are in use in the governance process and if it is more focused on the decision rights, control mechanisms, or an overall architecture renewal. In the architectural approach, requirements such as data retention, granulation, scale, and unified definitions for the information, and data warehouse modelling are examined and the governance process is built on top of these (Watson, Kraemer and Thorn, 2009, p. 437). By data warehouse modelling it is meant, that users are both able and allowed to run efficient queries across subject

areas (Watson, Kraemer and Thorn, 2009, p. 437). Tiwana, Konsynski and Venkatraman (2013, p. 10) see that a more traditional way of governance includes mainly control mechanisms and architecture is overlooked and ponder if the future will make a difference.

2.3.3 Roles

People inside organization work in different roles and have different aspects on the data and its use. All of the people involved in the process need to collaborate closely to ensure the key trade-offs between data and information quality criteria (Eppler, 2006, p. 340). Cheong and Chang (2007, p. 1006) found out in their study that a lack of clear roles and responsibilities among stakeholders leads to an ineffective data governance process.

A key aspect of data governance is the accountability; who is entitled to make decisions, who is responsible for them, and to whom correct roles, such as data governance steward and data ownership groups, are appointed (Griffin, 2005, pp. 49–51; Khatri and Brown, 2010, p. 149; Abraham, Schneider and vom Brocke, 2019, p. 426). There are different approaches to distribution of accountability: Borgman *et al.* (2016, p. 4903) see that it can be centralized, where decision-making responsibility is managed company-wide, federation, with both focused company-level control and also business unit level control, or decentralized, where business units are responsible for their own governance. Where centralized control benefits of increased coordination and control and suffers from added bureaucracy and more stiffer reaction to local demands, decentralized tackles the issue of inflexibility but doesn't benefit from standardization gains (Borgman *et al.*, 2016, p. 4903). On the other hand and as mentioned earlier, Brown and Grant (2005, p. 700) believe strongly to only centralized data governance and its economics of scale.

The RACI chart is a commonly used way of assigning responsibilities in any activity (Smith and Erwin, 2005). From the perspective of data governance, according to Wende and Otto (2007, p. 7) and Soares (2013, p. 33) the roles can be assigned as follows:

1. R - Responsible: a role who is responsible for executing a particular data quality management activity
2. A -Accountable: a role who is eventually responsible for authorizing a particular activity
3. C – Consulted: a role whose input and/or support is needed before the activity should be carried out, where there is two-way communication

4. I – Informed: a role that is notified about the activity, where there is only one-way communication

Wende (2007, pp. 419–421) and Weber, Otto and Österle (2009, p. 11) categorises these roles in their own paper further in the following roles as presented in table 1 and are opened more in this chapter with *cursive* text. Funding, support, and overall sponsorship from top-level management can be seen as the *executive sponsor* role's critical advantages to the success of the initiative. He is also responsible for the day-to-day management of data governance (Loshin, 2008, p. 83). Koltay (2016, p. 305) emphasizes that data governance needs to have clear definitions of its objectives, processes, and metrics. *Data quality board* is responsible for defining strategic goals and defines corporate-wide standards and policies to ensure uniformity on all levels. While data quality board is more accountable for the planning phase of the process, different stewards handle the practical implementation. *Chief Steward* should take the practical lead and/or support role in the process by having the necessary skill set of IT and understanding of business, whereas *business data* and *technical data stewards* provide their capabilities and expertise on more specific topics and help unify those on company level. Both of the latter roles are necessary and cannot replace one and another, since technical data experts usually work with file formats, access permissions, interfaces etc., and have the understanding of the backend but lack the understanding the business understanding, whereas business users have this and understand why and to which purpose the data is collected (Morris, 2006, pp. 32–33). Finding the right people for these roles can be challenging because it might not be suitable to choose the senior enterprise manager for some role since their calendar are more often highly booked, so ad-hoc meetings are not an option, but roles shouldn't be filled with a junior person who doesn't have the necessary understanding of the systems (Morris, 2012, pp. 95–96).

Table 1. Data governance roles (adapted from Wende (2007, pp. 419–421) and Weber, Otto and Österle (2009, p. 11))

Role	**Description**	**Organizational assignment**
Executive Sponsor	Provides sponsorship, strategic direction, funding, advocacy, and oversight for DQM	Executive or senior manager
Data Quality Board	Defines the data governance framework for the whole enterprise and controls its implementation	Committee, chaired by chief steward, members are business unit and IT leaders as well as data stewards
Chief Steward	Puts the board's decisions into practice, enforces the adoption of standards, helps establish DQ metrics and targets	Senior manager with a data management background
Business Data Steward	Details the corporate-wide DQ standards and policies for his area of responsibility from a business perspective	Professional form business unit or functional department
Technical Data Steward	Provides standardized data element definitions and formats, profiles and explains source system details and data flows between systems	Professional from IT department

Korhonen *et al.* (2013, p. 16) see that this listing isn't sufficient enough to form a well-balanced data governance model. According to them, this listing lacks roles pertaining the efficiency and effectiveness aspect at the strategic, tactical, and operational levels, as well as roles dealing with day-to-day activities. They do add that their conclusions are based on secondary sources and lack empirical evidence, which this study will furthermore provide.

To combine these roles and earlier presented RACI standard, Wende (2007, p. 420) presents following solution in the table 2. Later in the study, this theoretical perspective will be compared to actual findings.

Table 2. Example RACI-responsibilities based on Wende (2007, p. 420)

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

Of course there will be other roles associated with the process but it is worth noticing that one shouldn't include anyone, who doesn't have a real contribution to the process and whose in-

volvement may infer with their actual competencies (Morris, 2006, pp. 36–37). However, enlisting those who are needed in the process is extremely valuable but also difficult since, besides their time, they might need to use their own projects' money and resources to companywide governance process (Dyché and Levy, 2006, p. 73). Because governance projects require a lot of manual work due its uniqueness, human errors, misunderstandings, and misjudgements can play a major role. Halevy (2005, pp. 54–55) describes a scenario where people on senior-level database course designed completely different solutions to a single page instructed database purpose. It just highlights the fact that unifying different sources with different people can be difficult if the requirements are not clearly stated.

2.3.4 Master data management

In a relation to data governance, also master data and master data management is highlighted in various sources, such as Koltay (2016, p. 309) and (Berson *et al.*, 2010, pp. 406–407). Significant issues in data quality are also seen as a result of badly organized master data (Cleven and Wortmann, 2010, p. 1). For the definition of master data management, White *et al.* (2006, p. 2) present that '*master data is the consistent and uniform set of identifiers and extended attributes that describe the core entities of the enterprise'*. To elaborate this slightly more, master data management in practice means creating and maintaining a single '*authoritative, reliable, sustainable, accurate and secure data environment'*, which is accepted throughout the organization by all possible users (Berson and Dubov, 2007, p. 11; Das and Mishra, 2011, p. 131). In addition, master data management isn't just a technological problem, but in many cases, changes in business processes require clean master data and these issues can more political than technical (Das and Mishra, 2011, p. 131). Often also metadata is associated with master data. The difference between master and metadata is that metadata is information on the properties of the data unit, for example, length of field in database (Sen, 2004, p. 151).

3 Data governance and data quality in literature

The following chapters present a more focused view on the relationship between data governance and data quality. It also highlights the most common issues in data quality from the governance point of view.

3.1 Relationship

The terms data governance and data management can be mixed in general discussion but in literature, the difference between them is based on the aspect they take on the data: data governance states who is accountable for the decision making and deciding the standards while data management focus more on the metrics employed for data quality and implementing the decisions (Dyché and Levy, 2006, p. 150; Khatri and Brown, 2010, p. 148; Otto, 2013, p. 96). Since many of the activities of data governance and management aimed at data quality are invoked eventually by the same individuals or groups, it furthermore distinguishes the line between these terms (Pierce, Dismute and Yonke, 2008, p. 11). If we continue to compare data governance and data quality, a commonly used example is the water supply system where the system and maintenance protocols and personnel are used to describe governance whereas the water and its purity refers to data quality. To further elaborate the differences and linkages between these terms, Otto (2011, p. 48) has present the following figure 4:

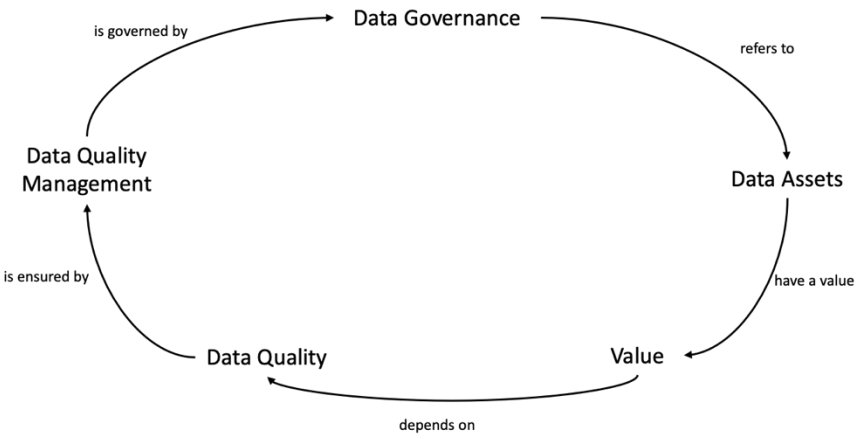

Figure 4. Data governance relationships (adapted from Otto 2011, p.48)

Data governance is the foundation for data management, and it provides answers to e.g. availability and access possibilities, provenance, meaning, and trustworthiness of the data (Koltay,

2016, pp. 305–306). It is also indispensable for managing data quality (also data literature) (Koltay, 2016, p. 309). Improving data quality with the means of governance can be derived into the following decision areas and key tasks:

A data quality strategy is needed to steer all the activities to be in line with the selected business strategies and goals. Typical tasks for this include developing the data quality strategy, defining a portfolio of the data quality initiatives, formulate business cases and carry out status quo assessments and establish review processes (Wende and Otto, 2007, p. 7). After establishing the strategy, designing an operational plan, which includes roles and responsibility defining, determining metrics and standards, and designing data processes, are the next logical steps (Wende and Otto, 2007, p. 8). In order to comprise all the possible information together, there is a need for data quality architecture, which ultimately ensures the consistent understanding of data by, for example, developing a common information object model, creating a business data dictionary, and defining information systems support including data quality tools (Wende and Otto, 2007, p. 7).

Data's accessibility aspect can be viewed through data governance actions: the target of the governance process is to ensure that the business process will have high-quality data accessible at the right place and at the right time (Korhonen *et al.*, 2013, p. 15). Besides aspects already listed, Watson, Fuller and Ariyachandra (2004, p. 437) also add that after the governance, data from multiple source systems should be so well integrated that it could be also accessed through one single endpoint. In their big data framework study, Al-Badi, Tarhini and Khan (2018, p. 275) present that the final goal for data governance in big data framework is to establish solutions for i.e. storage and optimization, and eventually improve data quality. What all these have in common is that already in the planning phase, data governance is designed to eventually improve data quality and that they have a strong relationship.

3.2 Data quality issues

The quality of data is stated at the moment it is created but it only measured at the time of use leading to a situation where there are tons of poor data in the system (Redman, 2013, p. 4).

Dyché and Levy (2006, pp. 71–72) claim that non-integrated data is frequently the cause of cost and time overruns across industries. In the following chapters are presented previous literature reviews of data quality issues in any sort of transformation process.

Morris (2006, p. 9) lists usually problematic into following four topics: '*Underestimating*' happens when the scale of all activities that need to be undertaken is failed measure in advance. The amount of data preparation can be difficult to predict in advance. With '*techno-centricity*' he means that the process is seen solely as a technical problem where data selection, quality etc. is seen as such a high priority that the actual business needs, ownership, and historical understanding of the data are forgotten. Also, Howard (2011, p. 12) found out that in most of the successful migrations, business engagement was ranked as the highest priority, and Halevy (2005, p. 54) highlights the business understanding when designing the schema for the solution. '*Lack of specialist skills*' is causing problems if the experts coordinating don't understand the business needs or lack technical skills and hence fail to communicate with technical colleagues. Their expertise is needed also if the project is heading towards '*Uncontrolled recursion*' by which Morris means a situation where problems accumulate and are tossed between the project and the business.

3.2.1 Data quality before official governance actions

Variance in the data quality can be caused for example name anomalies, like nicknames, missing data fields, misspelled addresses, or lack of standards in data value insertion where middle initial versus middle name is used (Dyché and Levy, 2006, p. 98). For some companies, the issue in data quality is the ignorance of some manager to admit that their data isn't good enough or their inability to fix the issue of poor-quality data causing this data to be significantly better in some department than in others when single managers aren't able to push their targets further in the organization (Redman, 2013, p. 8). Sometimes managers are also scared to admit to others that their data is, in fact, bad but they have kept using it (Dyché and Levy, 2006, p. 177).

If the data is missing clear ownership, there is no one taking responsibility for its quality and hence creating the 'pride of ownership'. Although this doesn't come without problems because if data is crafted by someone personally, it means that the details, rules, and derivations can be extremely complicated meaning that transferring this knowledge is difficult (Dyché and Levy, 2006, pp. 177–178).

Poor data can be a reflection of a faulty process, misrepresenting the actual world (Dyché and Levy, 2006, p. 77). Dealing with poor data quality is usually the users' issue and they can either fix the faulty data or decide to ignore it. In a longer run, this practice is not optimal in opposed to getting the data collector and data processor to communicate about the underlying issues where even small dialogue can make major quality improvements (Redman, 2013, p. 4). Or there can be issues in data validation, as Dyché and Levy (2006, p. 89) point out a situation, where a 'null' value was replaced in system by the default value of 1.1.1900 causing significant distortion in data.

In relation to default values, the overall standardization process is needed before further actions, where data consistency is enforced. Dyché and Levy (2006, pp. 96–98) introduce steps, such as parsing and semantic reconciliation, which are part of standardization. In parsing, an example value of '157 Wisteria Lane' is broken into different components such as street number, name, zip code etc. With semantic reconciliation, they mean that words with the same semantic meaning, such as tyre and tire, would be combined. This will help to build a logically consistent database.

Once data quality issue is acknowledged, it can lead to a no-value-adding process of cleansing the old data rather than to identify the underlying issues and root causes and focus on getting new data right from the start (Redman, 2013, p. 5). IT department cannot do much for the validity of the data if the quality is fixed at creation and measures are not set properly by business, IT doesn't have the understanding to correct the data (Redman, 2013, p. 6). In the governance process, moving bad data to another location is just a waste of resources which could be prevented with constant data profiling already in the source, where the data is studied and compared to its native source and made sure its accuracy is on a good level before importing it into migration process (Dyché and Levy, 2006, p. 95).

It's estimated that 80% of data in the organizations is unstructured and they have no means to handle and protect it (Rizkallah, 2017). One reason for this is the nature of unstructured data: most often systems with unstructured data don't have a clear schema for data and this data is shared to multiple systems causing more variation (Halevy, 2005, p. 53). This data ends up in

systems through automated processes from the internet, emails, and other unreliable or not up-to-date sources, which are not monitored closely (Dong, Halevy and Yu, 2009, p. 471).

3.2.2 Data quality issues while implementing governance

Poor understanding of legacy systems will lead to incorrect specification requirements for the target system which will eventually lead to failure (Bisbal *et al.*, 1999, p. 10). One solution to this is to ensure that the project team has Wende's (2007, pp. 419–421) and Weber, Otto and Österle's (2009, p. 11) *Chief steward* and *technical data steward* -roles. Besides the specification documentation, these experts need to define the relationships between the legacy system and other systems which remain in use, and ensure that the new target system will have the capabilities to replicate those relationships (Bisbal *et al.*, 1999, p. 11). Old, internal systems can be developed specifically for a certain business purpose and integrating data from them is difficult (Halevy, 2005, p. 53).

Besides understanding their data, companies may fail to identify the locations or sources of critical customer data. In legacy systems, the customer data may be buried in many different source systems and end up doing a prolonged data sourcing in order to create decent inventory for the purpose of migration (Dyché and Levy, 2006, p. 94). To map these relationships, usually two different terms are associated: ‘logical data model’, which refers to relationships of the data elements in business terms and hence reflect actual data requirements, and the ‘physical schema’ which refers to database tables as they are reflected in and processed by the process (Dyché and Levy, 2006, p. 67). When the data is buried in different sources, the owners of these sources might not be willing to share the data they own and want to keep it in their own system (Dyché and Levy, 2006, p. 74). Because of the companies’ uniqueness, standardized governance processes won’t probably work, and every company needs their own solution based on, for example, their data growth speed (Dyché and Levy, 2006, p. 69).

3.2.3 Data quality issues after governance actions

Dyché and Levy (2006, p. 156) see that most of the times it is better to have one data management unit, who is responsible for all the data in the organization. But after implementing all the data from multiple sources to one solution, the data might lose its quality, which has created the need for the shift from structural to semantic integration of the data (Dong, Halevy and Yu, 2009, p. 8). Besides combining sources, implementing new functionalities in the system, the

process takes a risk of missing if something has changed since the two systems are now not comparable (Bisbal *et al.*, 1999, p. 12).

If the real-life semantics are not fully specified and multiple sources and independent developers' work are combined, the data can have semantic heterogeneity i.e., different terms are used to describe the same event (Halevy, 2005, p. 50; Dong, Halevy and Yu, 2009, p. 9). If this issue hasn't realized already in source legacy systems, when business needs have shifted and data is shared between internal organizations, it will usually occur in mergers and acquisitions where data is migrated (Halevy, 2005, p. 52). In order to cope with this, *semantic mappings* can be used to specify how to translate the data from one source to another while maintaining the true semantics of the data. The problem here is that this is a manual labour intensive step (Halevy, 2005, p. 52).

Even after governance implementation, new (and old) data quality issues i.e., new subject areas, arise and changes made will affect later levels further in the system (Watson, Kraemer and Thorn, 2009, p. 438). This issue usually is related to data warehouses, where a solution could be the use concept of data ownership where the data owner of the current branch is responsible for its contents, such as correct modelling, documentation and quality control, methods for ETL-procedures and development (Winter and Meyer, 2001, p. 3). However, this isn't only related to data warehousing, since governance programs must have ways to monitor and control compliance in all aspects also in the future to be fully successful (Abraham, Schneider and vom Brocke, 2019, p. 426).

When discussing the security dimension of data quality, the focus has shifted towards cloud security, the physical security is often ignored although in the '80s and '90s it was the prior form of a security issue (Barker, 2016, p. 222). At the same time, the price of flash storage has decreased, and the sizes have increased, making data theft potentially easier.

3.3 Previous literature reviews for governance affecting data quality

In Barker's (2016, p. 165-166) study it was clear that organizations with a high focus on data governance and high level of sponsorship were able to enhance the quality of their data throughout the process, and even firms with lesser effort were also able to achieve improvements through scorecard- and monitoring systems.

Berson *et al.* (2010, pp. 406–407) present a combination term master data governance which includes master data management and data governance policies to specifically improve the data quality. In other words, this basically highlights the importance of master data in the governance process and puts it as a priority across enterprises. Although master data entities only consist of less than 10% of the enterprise's total data model, by tackling issues in this data helps to solve 60-80% of the most critical and difficult-to-fix issues in data quality (Berson *et al.*, 2010, pp. 406–407).

A hidden effect of concluded data governance program is that the organization gains a deeper understanding of their data, and this establishes a base where they are able to plan and tackle issues beforehand while organizations with less understanding have to focus on solving urgent crisis meaning less time running and developing the business processes. Naturally, understanding of the data further empowers their efficient usage of data (Barker, 2016, p. 165).

4 Data and methodology

In this chapter, used research methods, interviewees, and other relevant aspects regarding the study will be presented to provide evidence on how the study was executed. In addition, there will be discussion on why certain questions, methods and people were selected, and how it might affect the outcome of the study.

4.1 Research methods

Barker's (2016) study, *Data Governance: The Missing Approach to Improving Data Quality,* was the only one that had previously studied this issue with a case study approach as he failed to acknowledge enough material for a quantitative method study, and this is also one reason this study was conducted with interviews and as qualitative research. Used research method was a semi-structured interview, where the interviewee has the possibility to present their opinions freely and flexibly on open-ended questions. However, by repeating the same questions to each interviewee, the outcomes are comparable to one and another and the topic of the interview remains on the intend through everyone (Galletta and Cross, 2013, p. 47).

All the interviews were held online as a Teams-call, and they were recorded with the permission of participants to ensure that anything important was not missed. The interviews were held in both Finnish and English, and the transcribed and translated answers were collected on Excel-sheets first question by question and then further dividing them into common topics. From the answers, customer names and other possible identifying details were left out, as well as the case company's name and the name of the participants.

Before interviewing it was necessary to present the overall subject and research perspective to the interviewees firstly, to ensure the consensus of the used terms between participants and secondly, to guide the flow of conversation to revolve around data quality. This was done by providing short descriptions of the most used terminology as well as overall questions in advance before the interview. This is attached as appendix 1.

4.2 Participants

Participants for this study were selected from the case organization based on their experience on the studied issue. In total there were six professionals from Finland and abroad. Based on their own explanations and previous working history, these six people were divided into three different groups: technical (3), managing (1), and both (2). This information is also presented in table 3, which includes participants' overall title (direct titles could compromise anonymity) and their own definition for their role. Later these groups and persons in them will be referred like T3, M1, or B2, where the first letter refers to the correct group and the number to the specific person in that group. In addition, people in the technical group will be later referenced also as technicians, which is not most correct in the literal sense but is used to ease the referencing. For some of the interviewees, the background was more visible in the answers than for others i.e., someone was extremely skilled in his own narrow technical segment of expertise while being less interested in governance on a higher level. This created an interesting and truthful combination that likely would exists also in a real governance process.

Table 3 Interview participants

Title	**With own words**	**Group**	**Reference**
Head of Data Unit	Both manager and technical expert	Both	B1
Head of Data Unit	Background heavily on the technical side	Both	B2
Data management consultant	Mainly managing and planning	Managing	M1
Solution consultant	Some management responsibilities but focus on technical work	Technical	T1
Data architect	Purely technical expert	Technical	T2
Tech lead	Purely technical expert	Technical	T3

4.3 Interview questions

The interview questions (included in appendix 2) were written in a way that would produce insights to answer the original research questions and to follow the topics of the theoretical framework, where there are three top-level concepts: data quality, data governance, and the relationship between them. However, questions should have the emphasis on the relationship and leave the deeper phenomena of data governance and data quality out of the scope. This turned out to be rather difficult as they are highly linked together, and to understand and study the relationship, one needs to understand the individual aspects behind both data governance and data quality. The difficulty was to ask enough about background without drifting the study towards wrong direction, but then on the other hand the relationship exists because of these background factors and could not be explained without understanding them.

The lack of previous studies affected shaping and defining the research questions, as having almost no other study to compare and analyse possible difficulties in questions' layout and results' outcome, setting the question was challenging. On the contrary, with only limited earlier hypotheses, the possible outcome of the survey wouldn't be affected by any possible hidden bias in the questions.

The first part of the questions was aimed to study the understanding of data quality and its challenges in modern IT company environment. As it was clearly visible from the literature review, there are multiple different definitions for data quality and dimensions and as the theoretical meaning isn't at the scope of this research, data definitions of quality dimensions were presented to the interviewees, when asked to rate most crucial dimensions. The purpose of this was to see whether there are similarities in the selection and furthermore tighten the scope to find out which actions in the governance process are aimed to improve selected quality dimensions. There is also question about the most common data quality issues before and after the governance process.

Second part of the questionary revolved around the governance process and challenges in it affecting the data quality. As the interviewees represent different roles in the case organization, they probably provide different point-of-views for the same questions, meaning that the interview should be started by asking about their previous cases and roles they have played in the governance process. This question acted also as a conversation starter towards deeper inside

the casting which was widely discussed in chapter 2.3.3. Besides roles, in literature reviews best practices and technical tools very highlighted to play a significant role in the succeeded governance process, so a question of these was also included.

The relationship between the governance process and data quality was constantly carried along in both of the sections with questions like '*What are the changes in data quality after governance?* ' and '*With which data governance actions data quality in these dimensions can be secured?*'. It was believed that by asking directly about the relationship between governance and quality, the answers would have either been left blank or been irrelevant. Also, this way the conversation wouldn't drift away from the purposed topic.

When compared the planned interview questions to the original research problem and questions, some overlapping could be seen, meaning that valid answers could be expected. One issue that could arise from the questions is that because of the variety in the interviewees' positions inside the organization, other answers can go deep into technical challenges while others focus more on a higher level. This can broaden the aspect and enlighten the phenomena as a whole but there can be difficulties to get similar answers from multiple persons simply due to the limited number of interviews. This was considered while designing the questions by trying to minimize the scope and by asking the interviewees about their background.

4.4 Results' validity, reliability, and implications

For qualitative research, validity cannot be completely fulfilled, for example, due to social construction, where all of the participants have their own social and professional backgrounds, and they interpret the topic through these viewpoints. However, the topic can be described more as professional than as personal, meaning that all the participants should at least have a similar understanding of the topic, but from their own career perspective. In most cases, this should mean better reliability for the answers since the interviewees' answers are not related closely to their personal life. One compromising factor is that while some of the topics were presented by multiple interviews, some topics were brought up only by one interviewee. The goal for this type of study is to gather new information as long as there is no more new information obtainable (saturation), meaning the scope of this research did not fully exceed this point. However, aspects brought up by the participants can be understood as their subjective interpretation of

the topic and as stated earlier, due to their professional background and due the fact that they have no motivation to twist their responses, a reasoned hypothesis can be made that the reliability is retained.

Results and conclusions stated in this study were done in an objective way using good scientific methods which adds validity of the study. Questions were sent in advance and also presented on-screen during interview. This ensured that all the topics were discussed, and recording ensured that later on these were all analysed. There were some jumping between the questions, but since the topic was about the relationship between main themes, this does not endanger reliability.

The scale of study is a factor that must be considered when assessing about the possible implications. Participants are all from a single IT consultancy company, which might narrow the perspective, but many of the participants had worked previously on different organizations, some of them very recently, thus enlarging the perspective.

5 Results, analysis, and aiscussion

In the following chapters, the results of the semi-structured interviews will be presented alongside with a comparison and discussion with the topics already presented in the earlier literature. The presentation of the results follows slightly the original pattern of the study questions but are also grouped under three main topics of data governance, data quality, and their relationship. Some smaller entities are grouped under the last subchapter, 5.4.

5.1 Data governance

As literature presented, the topic of data governance can be wide and include many things, and to find out how each interviewee sees this, they were asked to present it in their own words. B2 and T2 had similar ways to describe data governance when B2 said that '*governance is basically knowing and managing your data at rest and in motion*' and T2 rationalized through an example: '*if you have for example a data lake with terabytes of data, it needs to be managed in order to find anything'*. Another similarity in the answers was that data governance was named as a framework or a management model, which sets the rules and guidelines for processes. While T1 presented that these rules should ensure that data is both imported into the system and also processed correctly, B1 left this on a higher level by stating that guidelines define the processes and connect them together from every area of the organization. Also, M1 backed up these views by concluding that data governance creates transparency and manageability for the data, and data management as a process implements these actions into practice. According to T1 and T2, one of the key aspects of data governance are user-rights management and responsibilities, which they included already in their definitions.

As also stated by Abraham, Schneider and vom Brocke (2019, pp. 425–426) and Al-Ruithe, Benkhelifa and Hameed (2019, p. 7), there does not seem to be a universal definition for data governance between the participants either. Neither there was a clear difference depending on your career path, but all the participants presented more or less the same themes. When compared these themes to the ones found from literature, we can see that there are repeating keywords e.g., *framework and guidelines, managing data, user- and access-rights, organization-wide, roles and responsibilities.* However, later in the interviews, there were some implications that the scope of data governance can be dependent on your career path: a more technical person

is more interested in his own specific field, while managers adapt the overall picture. But all of them are capable of defining the data governance and know what is meant with it, even though they have their own way of describing it.

5.1.1 Why govern data?

While conducting the interviewees, this topic seemed to be one with the most variance in the answers. This was somewhat due to question setting: afterwards it would have been better to ask for both preconditions and goals since both were mentioned by multiple interviewees in relation to this question. However, after reviewing the answers, there were few themes that were discussed by multiple interviewees.

There is a need for a common driver in the organization that will be pushed from the top-down, included in the strategy, and where data quality and governance is communicated from the business to the IT (M1, T1). Even though this sounds more like a precondition than the goal, it can be seen also as the desired outcome: earlier the initiative for governance process was some sort of need, like a demand for better data, but now the situation has shifted, and organizations have some sections of governance process implemented but the execution is lacking the driving force and common goals (B1). Furthermore, the common understanding and objectives between business and IT is a crucial milestone for the governance, because even with technically correct solutions, IT cannot produce correct numbers from the business point of view, if there is no mutual understanding (T1, B2). Luckily, according to T1, this has been understood already rather widely in organizations.

'Ideal situation would be to see how completely new organization would start: if they would go with the framework first or with "do and fix later"-mentality' -B1

On the other hand, the goal for the governance can be derived rather directly from its definition: the goal is to create a management model (M1), increase the awareness of one's data (B2, T1), and so on. Sometimes implementing governance is besides profitable but also mandatory, like T2 points out, that different regulations, such as GDPR, can be driving force and goal to achieve. Even if there are regulations for data, the optimal situation would be such that data quality would be included into the strategic goals and performance targets (B1). Sadly, M1 experiences that this is often forgotten, and its importance is not understood.

While analysing the results and comparing them to the previously written literature, it was quite surprising that close to no one talked about the drivers of the governance process. From the used literature sources, only Al-Ruithe, Benkhelifa and Hameed (2018, p. 18) mentioned regulation drivers such as GDPR, and Abraham, Schneider and vom Brocke (2019, p. 432), who briefly mentioned that there can be strategic, organisational, system-related and cultural factors which can act as a driver. Even after re-searching from Google Scholar and university library archive with search words 'drivers for data governance', the number of publications handling topic was slight. Also previously cited Weber, Otto and Österle's (2009) research mentioned drivers few times in passing, while Ladley's (2013, pp. 97–99) book on data governance implementation did focus more thoroughly on the issue and defining the differences between the driver and goal of governance process: '*Goals are refinements of drivers, expressing the general trend in terms that indicates desired accomplishments within a timeframe'*. While this did explain why participants linked these two terms together, it still didn't highlight the importance of the drivers as much as it was emphasised in the interviews. Since most of the participants answered and highlighted the importance of finding the inner drivers for the process, even when the initial question was asking about the goals of the process, it seems safe to say that finding the inner driver inside the corporation was seen as more important than it is presented in the literature.

> *'Often the driver is forgotten to communicate: in banking, it's the regulations but in the industry field, the data quality issue should be included into the organization strategy'* - M1

5.1.2 Expectations

The difference between expectations and goals for governance is that expectations are set from the customer side and goals can be decided also from the vendor side. Mainly this topic was included to understand whether the expectations from the customers are in line with the possibilities and goals of governance. Answers stated that there are still significant differences in expectations and actual possible outcomes.

The main reason for the unrealistic expectations is that the business, who is paying for the governance process, doesn't know nor understand the initial goal (B1). This also leads to a situation which was raised in the interviewees by M1, TI and T2, where business is expecting

something new, like sources or huge data quality improvements right away. The problem is that depending on the state of the governance before, there is often huge amounts of work needed to ensure the future governance steps, and for business, this shows as a nonprofitable and expensive cost compared to the profit gained. The future steps will not be funded which would have been more efficient per euro used since the process is scalable, and a properly designed base can be extended to all organizations inside the company. However, when the process is advanced to a certain point, close to the end, the cost of fixing the final percentages will rise once again. According to T1, it is common that once the process is started, it is then overworked and tuned to the point where improvements made do not anymore improve the quality with the same efficiency as earlier. With this information, there can be an assumption made, that the cost of improving the data quality follows an S-shaped function, where the x-axis presents the cost and the y-axis the improved quality. An example is presented in figure 5 below, where data quality starts from 0% and ends up till 100%, which is not the most realistic scenario but represents the issue.

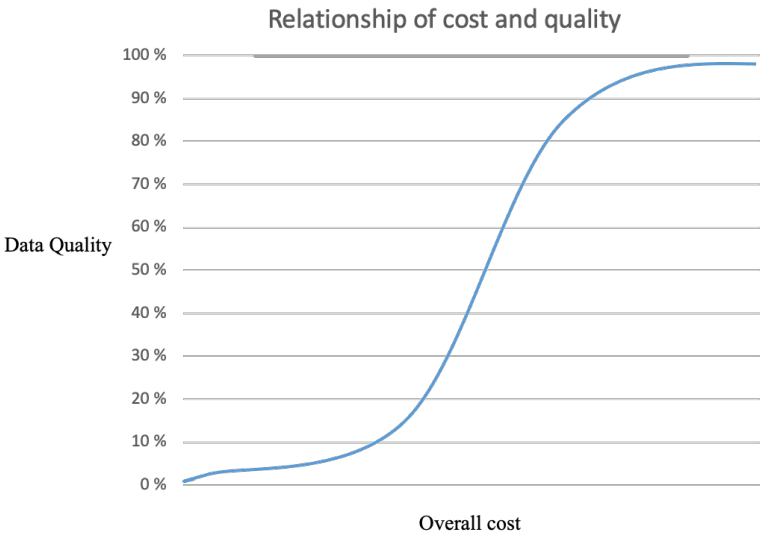

Figure 5. S-shaped relation of cost and quality

5.1.3 People in the governance process

Topics of roles and casting were highly focused in the literature indicating that it is important to include. For most of the interviewees, their own roles have been mainly on the vendor side

but there were also some, like B1, who had acted as a customer. However, this nor other background didn't seem to affect the answers, but all the interviewees highlighted similar topics.

From the literature sources, like (Griffin, 2005, pp. 49–51; Khatri and Brown, 2010, p. 149; Abraham, Schneider and vom Brocke, 2019, p. 426), the topic of ownership and accountability were mentioned as key aspects, and this assumption was supported by multiple interviewees in various relationships. Besides, topics of responsibility and ownership were mentioned while defining the data governance, indicating their importance. For B1, one caricature example of clear ownership is, that the person responsible for a mistake in data can be found and addressed easily. M1 did follow this example by describing that assigning the ownership to someone means, that they are then responsible and can be held accountable later on. In the interviews, the discussions were held on a more overall level, while the literature presented different standards, such as the RACI (Wende, 2007, p. 420). Probably this is mainly due to the question setting and the fact that this was just one small topic among others. Even though there was no mentioning of the RACI standard, by interpreting the answers and example of T2, the *Responsible* was the technical expert (possibly external expert from vendor) and *Accountable* and *Consulted* was the business unit owning the data. For the last, *Informed*, there were no mentions, but could mean, for example, any end-user using the data. A separate data management office was mentioned in the interviews, which would include both the business owners and it, and finding the difference between the *Accountable* and *Consulted* inside this one unit is probably challenging and circles up to between individual people. From one point of view, it seems that the RACI standard or similar is still valid to a certain extent, but the organizations have switched towards a more agile and flat structure, meaning that the roles of *Responsible, Accountable and Consulted* do exists but are all working together, and the communication is continuous. Korhonen *et al.* (2013, p. 16) did mention that this RACI -standard was lacking efficiency and effectiveness aspects at the strategic, tactical, and operational levels, as well as roles dealing with day-to-day activities, but this statement was lacking empirical evidence. This study does support these views and provides some evidence.

Another strongly highlighted aspect was that people working for governance projects need to have the time allocated. M1 saw that '*the challenge is that often people are running governance-related tasks besides their own day jobs*' and T1 agreed with similar views that '*management is responsible to allocate enough time for people*'. Partly this is also linked to what is

already discussed earlier and what T2 also brought up, that the governance process needs to get acceptance from high-level management to secure adequacy of resources. Besides time, people working on it need to understand the benefit they are gaining and/or producing to commit fully to the process as data quality management is often seen as boring and non-beneficial (B2, M1).

One bottleneck in the casting process might be finding the person with high enough position, who is also willing to adopt and invest in the governance process (B1, T1). T1 extended the importance of a good sponsor with the thought, that these people have often good connections and networks in the company which simplifies the sometimes-challenging process of finding correct people. On the complete contrary, M1 suggested that if capable people are not found easily from the organization, there is a possibility of buying the process and the experts as continuous service from a third party. '*Why invest heavily own resources into governance when it is not your key competence'*, he summed up. At least at the beginning, this approach is easier, but once the process has gone further, there can and should be founded new positions specifically for governance-related tasks, such as data management officers (M1).

Both the literature and interviews stayed on the same topics, such as capabilities, responsibilities, and different roles. In the interviews, committing the participants and improving their motivation towards the project was more discussed, and in fact, seen as the key factor to success. This wasn't completely left out, as Morris (2006, pp. 36–37) did mention the importance of everyone's involvement, but the literature, like Wende (2007, pp. 419–421) and Weber, Otto and Österle (2009, p. 11), did focus more on listing the correct capabilities for each role, and not so much of keeping them engaged. In addition, listing the requirements doesn't necessarily help finding the correct person, if you don't know where to look at. T1's example of using one's network inside the company was an interesting observation since '*networking'* in general is a trending topic, but mainly outside one's own company. This aspect on the other hand requires building contacts inside the large companies and over the borders of the tribes, which is always desirable but challenging. Most of these connections were probably previously made during smoking or coffee breaks, but nowadays as remote work has come as an inseparable part of the working life, how are these networks built. It doesn't happen over a Teams call, does it?

5.1.4 Tools and practices

The overall concept of data governance is tightly bonded with frameworks that dictate the best practices and management structures. The downside of these is that although that there are multiple different ones, finding and adapting one to fit one's organization's needs can be challenging (M1). Even different people can interpret the same framework slightly differently, like T1 pointed out a case, where a big organization needed multiple architects who all had their own way of designing, and without a clear leader, the result was a mix of different styles.

The frameworks usually consist of two elements: defining the ownership and the practices. For this topic, there was a clear difference between technical and managing people since technicians highlighted generalized guidelines as best practices while managers leaned more towards people and building data management offices. In this context, the guidelines refer to the structured way of working inside the whole organization meaning that the same steps are repeated in similar activities. While this does ensure the same level of quality, finding the balance between enough and too many guidelines is necessary so the organization doesn't end up in a situation such as in T2's example of PowerPoint presentation with more than 100 pages of how-tos. From the management point-of-view, the key takeaway was that there needs to be clear leadership inside the data office, who is capable to do the decisions (B2, T1). This data office should also own all the relevant data so there isn't the possibility to form any individual tribes such as developers and data support office (T1). However, gathering necessary resources, meaning mainly funding, for a large data office might be a challenge (M1).

Available software has one major weakness: they tend to leave the topic on a theoretical level and tell the capabilities of the software but the guidance for practical implementation of the tools and structures is left out (T1). Some solutions do provide on-call support and ready-made processes, but the price tag tends to make this available only for larger companies. Opposing solutions are open-source solutions, which are less expensive to use but require more experience from implementers (M1). Finding the best one that suits your needs isn't necessarily clear but in choosing process, B1 saw that you should go with the one that's most leanest way to your organization, while T2 preferred to have tools that can be interconnected, like Microsoft Teams and Azure DevOps.

One of the new presented solutions is called Data Catalogues, which was especially highlighted by T1 and T3, but also mentioned from others. More interestingly, there were no mentions in the literature, which can be due to fact that these are rather new solutions, for example, Azure presented data catalogues in 2016 (Microsoft, 2016). In briefly, data catalogues are a collection of all of the organization's data assets and it includes the metadata of each source, making browsing and finding information that already exists easier and more manageable (Microsoft, 2016). Due to big data and the constantly growing amount of data, finding the information you need is getting more and more time-consuming, where this type of solution is valuable.

5.2 Data quality

For the question of data quality, there were few different kinds of answers that arise. T1 described the technical qualities and data quality dimensions of the data, such as completeness, and emphasized the success of achieving these types of goals, while many of the interviewees connected it to data governance. Statements like '*That (data quality) also is practically dependant on how good data governance you actually have'* from B2 and '*it can be used to meter the functionality of data governance*' from B1, prove that the concept of data quality is strongly associated with data governance. On the other hand, metering and testing the data was mentioned by B1 and M1 as a way to ensure and define data governance. A reason why data quality definitions were highly related to governance might be a result of the overall topic, but there was a difference that managing roles saw quality and governance more as together than technical people. For technicians, the data quality was more linked to quality dimensions and measurable values.

5.2.1 Key data quality dimensions

The participants were asked to list the top three data quality dimensions according to them from the below list, which was constructed based on previous literature. The purpose of this was to find if some significantly more important dimensions would rise from the answers, but this didn't quite happen. There was also a follow-up question on whether there were ways to tackle issues on these specific dimensions, which is discussed later in the context of governance actions on data quality.

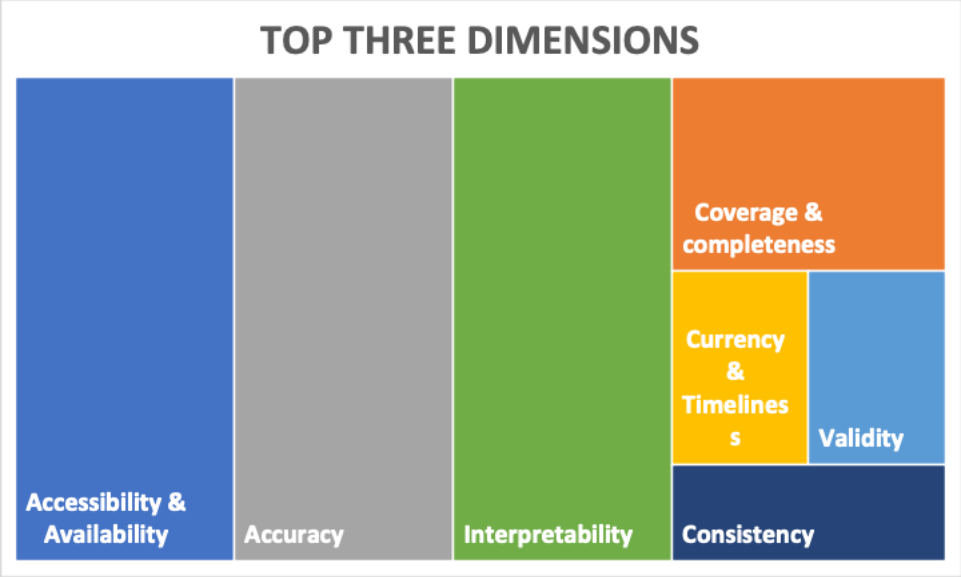

Figure 6. Data quality dimensions based on their importance to participants

Naming the top three dimensions was thought to be a difficult question in advance and the results did follow this assumption. While few participants could name the top three, there were also many who mentioned more or less than the asked three. In the figure above (6), the answers are listed based on their popularity. Accessibility & availability, accuracy, and interpretability were mentioned most of the times, while currency & timeliness, validity, and consistency were mentioned only once. Many of the participants mentioned that naming these depends on various aspects, like from whose point of view we are discussing and the number of existing issues in them: T1: '*For end-user, interpretability and accessibility & availability are more underlined features - - for technical people, it's coverage & completeness'* , M1: '*In an ideal world, all of these would be fixed but there seldom is an organization which has issues in all of them so we can focus on the dimensions with most issues'*. Based on these comments, it seems unwise to say that some dimensions are seen more important than the others, even though some dimensions were presented slightly more often. In addition, there was a huge number of different dimensions, according to Jayawardene, Sadiq and Indulska's (2015) review total of 127, and grouping them to 8 most likely leads to some overlapping, which was experienced by the interviewees. Dividing data quality into such small dimensions might not be even that beneficial: it can help identifying and naming every quality issue that is occurred, but at the same time it seems difficult to talk about the dimensions as individual entities as they are all connected to each other and thus also need all to be fixed to gain the desired outcome.

One differencing topic was raised by T3, that is, in a sense, linked to data quality but also to used tools. In his example, there are trackers placed on a website, that track customer behaviour and this data is then stored in a database. But when a business unit is looking at the data of sold items and tries to find patterns, this data won't ever raise to those reports, since there is no actual transaction, if the tracked customer didn't make the purchase. Is this behaviour issue of data quality or is it an incapability of used software of poor design of the overall system? For T3, the solution could be the proper use of data catalogues (presented in chapter 5.1.4), but still requires a proper understanding of the data from the users, which is then a completely new aspect to the issue.

What the above example and previous chapters highlight, discussing purely on data quality as an isolated entity is not possible. Obviously, there are aspects that can be studied individually but for the most part, they are interconnected with many things. With these conclusions, there is no point trying to analyse these dimensions at this point but rather focus them later alongside related topics.

5.3 Relationship between data governance and data quality

The relationship was already clearly visible from both definitions of data governance and data quality even though it was also listed as a separate question. For this question, some of the participants copied or adapted their previous answers, like B1 who said that the relationship is direct: '*data quality can be used to meter implementation of governance and governance is used to define the quality*'. Even though this means the same as 'data governance is needed to ensure data quality' (e.g. Koltay, 2016, p. 309), all of the previous reviews discussed this phenomenon from this perspective and not the other way round. The difference comes when we are in a situation, that we have bad data quality despite implementing data governance, which hypothetically couldn't be possible as implemented data governance should ensure the data quality. But these situations do occur, and by switching the mindset around, we can start looking critically at the implemented governance procedures and how are those failing. In addition, M1 presented an interesting aspect to data quality: it can be measured with different KPIs and such, but the governance is the linkage to the goals of the business side, and this is why it also improves the quality of the data. On one hand, T1 also pointed this out already while defining data

quality: '*the data that is accessible, needs to fit for the purpose it's needed'*. More importantly, this supports the angle, that data governance can be measured through data quality: end-users will eventually determine the quality of the data, and even with technically correctly governed numbers, the quality for business users can bad, indicating that the governance is not correctly done.

There were also answers, which followed the same 'governance ensures data quality'-way of thinking: for B2, the relationship was really straightforward: '*From point of origin onwards when the data moves to another application or else in the organization, if the governance is not good, the quality will get bad eventually*'. This corresponded to his answer for the governance, where he saw that through governance, one is able to tell the whole lifeline of data: from its point of origin to consumption. Similar thought was also presented by T2, whose key takeaway was that with proper governance, you are able to pinpoint and fix the issues in data quality because one knows where '*data comes from, who can access it and everything is documented'*. T1's definitions of governance and quality also overlapped quite much but for her the linkage was especially present through liability issues, which would be present in data quality without the governance. Quality responsibility was also focused on in M1 answer, where he saw that still too often data quality is forgotten while implementing governance program, or it is not taken under the governance program but rather put up as separate program and responsibilities and ownership of the data are not assigned correctly.

By no means, neither the 'governance ensures quality' or 'data quality measures the governance'-way of thinking is more incorrect. But the interesting part is that, even though they mean the same, together they enlarge the viewpoints of why data quality is bad in relation to data governance. And furthermore, this helps identifying the mistakes in both.

5.3.1 Common data quality issues from governance's perspective

Most organizations start with different levels of governance already implemented, and in bigger organizations there can be significant differentiation between units of how governance is handled (B1). While the topic of tribe behaviour is a completely different issue, it does seem to also affect governance but not with the traditional 'reluctant to share information' -way but in a way that each unit have their own Excel sheet or system, which they believe is the master for the whole company. Then information is shared across the organization, and everyone interprets

this information slightly different (B1, M1, T3). Issues could be solved by establishing the responsibilities and ownership so that every party knows who holds the master record and is authorized to update it (B1, B2). According to Redman (2013, p. 8), managers have significantly different capabilities of managing the data, and by assigning the ownership incorrectly, it can affect the outcome. However, there doesn't seem to be any issues with managers not understanding or admitting in the quality of their data, like mentioned in the literature (Dyché and Levy, 2006, p. 177; Redman, 2013, p. 8), since many of the participants mentioned that bad quality of the data is commonly known, but the root causes, however, are not. Probably this will lessen the chances that wrong managers are chosen or volunteering to be involved in the governance process.

To define the ownership, organizations need to know where data is originating and its complete supply chain. M1 presented that every data pipe needs to be described so the organization can identify its whole lifeline. Similar thoughts were also given by B2 and T2, who presented this through an example concerning the difference between null and zero-values. If the whole pipe is not known, there is a great chance that different programs and tools process these values differently and end-user is not able to tell the difference between the reported zero and a missing value. In addition, it is necessary to focus the attention on fixing the root cause and not the issues in hand, but since described processes are missing, organizations don't understand nor have capabilities to address the cause of the issues (M1, T1).

The buzzing word 'big data' was mentioned both in the literature and in the interviewees. The excess amount of data and its growth speed was mentioned, for example, by Rizkallah (2017) and Dong, Halevy and Yu (2009, p. 471). B1 brought up the same issue, while talking about the validity-dimension and the huge amounts of unnecessary data in the systems. For T1 and T2, the amount of data corresponded directly to the availability and performance of the systems. Although this was seen as a problem, there weren't any clear statements that this would be covered by governance. Even though storage is becoming less and less expensive and processing power improves, this will still likely create issues in the future, meaning that it would be good to include in governance scope.

Already briefly touched on in chapter 5.1.1, one of the founding misconceptions is that data quality and governance are fixed by the IT department and with technical solutions. The problematic in this was highlighted in multiple contexts by multiple people and was brought to concreate life with T1's example: '*there is the issue of capability: IT cannot be held responsible for sales posting if the numbers are not entered correctly to begin with*'. In addition to this, B1 stated that it's not uncommon that finance, product, customer data etc. are divided into separate entities and under different units to manage although they are connected to each other: e.g., product data shares some features of finance data, which's understanding is in the finance unit. These findings follow closely the same mindset which was presented also by many researchers in the previous literature, for example, Friedman (2006) and Wende and Otto (2007, pp. 1–2). More interestingly, this issue is recognized in the literature already 15 years ago and it seems that organizations still struggle to understand it fully. Luckily, there seems to be some light at the end of the tunnel, since participants mentioned that nowadays more and more organizations are forming separate data management offices, which are constructed of technical specialists but also of business experts, whose whole work description is dedicated to data governance-related tasks. To establish a data management office, the organization needs to recognise that a data governance program is not a one-time improvement but an on-going process, which is needed just like an HR program (M1). This does circle back to the problem of resources, where the first steps are funded by business and are expensive compared to gained profit.

5.3.2 Changing the habits of data quality metering

A common factor in the answers was that with bad data quality comes also bad metering of it. In fact, in many organizations data quality is missing metering completely or it's implemented at wrong place, and the starting point is a 'feeling' of bad quality, but actual evidence is missing. In this topic, wrongly implemented meters can be technical meters, which can for example monitor the duration of ETL load times providing maybe information about the technical solution, but not about the actual quality of the data inside the solution (T1). While that is true for some parts, there are also contrary statements supporting technical meters, like data growth and duplicate row counts, which do provide some evidence on the quality of the data (M1). 'The feeling' of data quality on the other hand leads to a situation where certain issues are recognized and fixed but nothing is documented, nor root cause is fixed (T1).

A better way to approach the issue is to define the meters to address the quality itself, which in the end is not even difficult, once there is consensus and understanding of which numbers are correct. For example, product data can have several attributes for a single product and by comparing values at the end of the data pipeline to the original, verified values, the ratio of correct values is rather straightforward to calculate (B1). With historical repetition trends of development can be then found and with proper governance and ownership, correct people can be addressed (B1). This also creates the possibility to meter the progress of the governance process (M1).

Telling the business that changes are happening without concrete evidence can be challenging, and the problem is realized when there is no data prior governance making the comparison impossible. Even with some improvements in some technical meters, convincing business to keep paying for the process was felt difficult (M1). The best way to cope with this issue is by trying to define the meters in a way that either gained cost-savings or business advantage can be presented but creating up measurable values can be challenging. There were some examples provided by M1 and B2 concerning mainly created incidents: There might be cases where wrongly entered product code can be shipped out from the warehouse, but the billing department will then need to start a time-consuming process of settling what was actually shipped out. In this example, the amount of time wasted due to the wrong number in the system can be metered and then presented to the business.

The topic of metering was one of the most interesting ones in the study because in a sense it was one of the initial reasons for the whole study. There was very limited amount of previous information about the changes in data quality after the governance process: Barker's (2016, p. 165-166) study did point out that high-level data governance enhanced data quality, but for the rest of the studies, it was mainly ‘a given fact’, that quality would improve alongside the process. The results, that metering is not implemented, proved that this topic is missing from the previous releases because there isn’t evidence to draw conclusions. The best indication is the hunch, that quality was bad, and it improved. The proven fact is that after implementing the governance and metering, addressing issues is easier and corrections are done on a constant basis, meaning that at this point changes in data quality can be factually presented. For the organisation, and more precise for the persons planning the governance process, implementing

metering before starting the process could be beneficial in the long run. As there is always the issue of funding, this type of evidence can define the continuation of the governance process.

5.4 Other relevant findings

In the table 4 below, the most important findings of this study are gatherer together. Some of these were already reflected independently in the chapters above, but in the following chapter these are discussed together with each other and compared to literature.

Table 4. Differences and similarities between literature and interviews' answers

Theme of data governance	Findings from interviews	Literature
Definitions of the topics	Common themes were carried in all of them	-
Starting point	Finding the inner driver and communicating it top-down	Inner drivers were not highlighted, some outer drivers such as GDPR were mentioned
Expectations	Business is expecting something new, and improved quality of old data might not be enough	The linkage between expectations and resources was not discussed
Resources	Understanding the amount of needed work can be vital for governance to succeed	The linkage between expectations and resources was not discussed
Ownership	With ownership comes responsibility and with responsibility comes quality	One of the key aspects, and too often forgotten
Scope	Companywide, to include everything	There were multiple different approaches and levels on how widely governance could be done
IT vs Business	Business is responsible for the data, not IT	There were multiple different approaches and levels on how widely governance could be done
People	Securing enough time to invest	Top-down structure, RACI-standard
After governance	Continuation, never-ending process as data management offices	No significant discussion
Data Quality Dimensions	No one more important than the other	All the dimensions need to be taken into account and there is no consensus between the researchers of the most important dimensions. While conducted already in 1900 and 2000 are still valid
Metering	There is no metering before governance	No significant discussion
Differences depending on the role	Wasn’t clearly visible from the answers	-

Based on the highlights in Table 4, the main discrepancies between the theoretical literature and expert interviews seemed to relate to the scope of the governance, the way people are organized in the governance process, and how significant role resource planning plays. Next, the interview findings are elaborated in more detail.

5.4.1 Definitions of the topics

As somewhat predicted, the definitions raised in the interview followed quite closely the same as presented in the literature. However, some differences were detectable, such as the connectivity between data governance and data quality was more present in the interview answers than it was in the literature. Most likely this was affected by the overall topic of the research, and the participants connected the topics slightly more than usually. Also as discussed more in chapter 5.4.6, there was no significant difference between technicians and managers, but both of them were able to elaborate the terms.

5.4.2 It always comes down to funding

The topic of resourcing and more precisely funding was extremely highlighted in the participants' answers, and on completely contrary it was only mentioned slightly by Dyché and Levy (2006, p. 73). The main concern among the participants was the continuation of funding to keep the governance process running, while the first steps are cost-intensive, and the results might not be visible at that point. Another issue is created once the governance process reaches the point, where a constant process of data governance is needed to establish. Both the issues can be solved prematurely by following the rule presented both in the literature (Friedman, 2006; Wende and Otto, 2007, pp. 1–2) and in the interviews, that the main responsibility for the governance needs to be hosted by business and not IT. The literature reviews discussed this topic more through the technical versus business objective perspective, and on securing the funding by involving the business in the process.

Also involving the business in the project, the possible unrealistic expectations can be discussed in advance. According to interviews, the business doesn't fully understand the scale of data governance program and it may cut the funding before the process is completed to the point of constant governing. This further hopefully helps the business to understand that the governance process doesn't necessarily mean 'something new', as they tend to hope, but it is a necessity for data assets to be worth something.

5.4.3 From data quality dimensions to master data

As already discussed in chapter 5.2.1, there were no significant differences between the quality dimensions regarding their importance. However, the topic of master data was mentioned by multiple participants as one of the reasons for bad data quality, and with comparison to the literature, these findings do support each other. The controversial issue here is that, while in both of them this was seemed as important, but it was still passed rather quickly. For some part, this could be explained because even though being a major issue, it can be fixed rather simply by forcing the one master to everyone in the company. But as stated in the literature by e.g., Das and Mishra (2011, p. 131), the issue can be multi-layered with a political topic, like who owns the data and so on.

5.4.4 Scope of the governance

In the literature, there were different levels for the scope of the governance: for example in Tiwana, Konsynski and Venkatraman's (2013, pp. 9–11) governance cube, these scopes were discussed through *who, what, and how* questions, whereas in the interviews the consensus was that everything should be governed. Like B1 pointed out, there often is differentiation in the level of governance between different organizations inside the whole company, but this means that the governance should be fixed on all levels. A similar type of pattern was seen in the differences regarding the distribution of responsibilities: e.g., Borgman et al. (2016, p. 4903) presented that it can be either centralized or decentralized, and both of them have advantages and disadvantages over the other. The experts in the interviews believed strongly only on centralized distribution and its standardization gains and the economic of scales. Although, Borgman et al.'s (2016, p. 4903) point of how slow and inflexible centralized model is compared to decentralized does seem to put a challenge towards modern world’s agile organization culture. Then again, many participants mentioned that the key issue is master data management and different units having their own master data versions. Maybe this means that the future way would be more towards a decentralized model, but it is only possible once the standards are implemented on the company level, and then it can be distributed correctly to separate organizations. Skipping this step might be harmful in the long run.

5.4.5 Way of working

Differences between the ways of working were not necessarily in the scope of the governance and data quality, but since it was presented in the literature and also mentioned in the interviews, there should be few words also included. The highlighted difference was that in the interviews the way of working was agile and separate data management offices were used to govern all the data in the organizations. However, for example, Wende (2007, pp. 419–421) and Weber, Otto and Österle (2009, p. 11) presented their roles of the *executive sponsor, data quality board, chief steward, business data steward, and technical data steward*, and while they don't specifically mention any hierarchical structures, they are still visible. Korhonen *et al.* (2013, p. 16) did recognise that this type of system is not sufficient enough to tackle day-to-day issues. From the interviews point of view, the data management offices would be suitable also for this.

The difference itself might be explainable by the overall shift towards a more agile and lean way of working, but before there can be drawn any larger conclusions, it is needed to remember that the study consists only of experts in one organization. There can be major differences between organizations regarding the learned ways of working, which can lead to polarized results. Nevertheless, these data management offices are pushed also to the customers of this case-organization, meaning that they are also accepted to their working culture.

In addition, while literature discussed about governance mechanisms (e.g. Borgman *et al.*, 2016, p. 4903; Abraham, Schneider and vom Brocke, 2019, p. 29), these were not specifically mentioned in the interviews while asked about the best practices. However, mentioned practices can be grouped under the *structural*, *relational,* and *procedural* mechanisms. Procedural mechanisms were highlighted especially by T2, whose emphasis was on the role-level security and holding the data securely and confidently. One mention related to this is that as discussed by Barker (2016, p. 222), the attention also in these interviews, was more on the cloud security and not on physical security, which represented the main threat in the '80s and '90s. This is rather alarming as the development towards cloud-based solutions has not necessarily eliminated this threat. For relational mechanisms, such as collaboration between stakeholders (IT and Business) was mentioned multiple times. As so were structural mechanisms while talking about the responsibilities. So, while the mechanisms were not mentioned, it doesn't necessarily mean that participants have not adopted them, but they don't at least present the actions they perform through this framework. Probably for the managing or sales persons it would be suitable to

learn and understand the entities, and also to present them through this mechanism in order for also a listener to understand the relationship better between actions.

5.4.6 Difference between technicians and managers

There were not many noticeable differences between the groups regarding their answers. The main observation was that for the technicians, the overall concept of data governance might be slightly unclear. Data governance tasks are executed, and data quality is improved but the entity how things revolve and are connected beyond those single actions was not necessarily on their mind. On the contrary, for the managers, the actual execution of steps might be somewhat unknown, but on the other hand, it isn't their competence. More deep down, this wasn't necessarily visible from their own answers, but from the technicists' answers, which were more on a detailed level. This might be due to the fact that besides managing the team, managers need to sell and explain processes to the customer, meaning that sounding convincing is something they have rehearsed earlier while technicians have not.

More important is that what does this finding actually means. The importance of communicating the goal and driver for everyone in the process was highlighted in the study, also from technicians, but it seems that there are still some issues in the information pipeline. Whether the reason is the complexity of systems or overly long PowerPoint presentations or that people's interest consists of only their own narrow field, encouraging everyone involved to truly understand the entity would be beneficial.

'It was year 2011 and major organization in Finland (name left out) was missing data governance principles and official structures although governance activities were carried out. Eventually they started to form towards an official framework' – T1

6 Conclusions

The goal for the study was to discuss the relationship between data governance and data quality, and more specifically how does an implemented governance process affect the data quality in the organization. The first research question, '*According to the literature, how data quality can be measured and how organizations can improve their data quality by establishing a data governance strategy?*' was designed to enlighten this issue. This question is better to be broken

into two parts, and the review can be start with the first part including the measurement of data quality. For this, the literature introduced mainly the framework, which can used to assess the quality and properties of data. Jayawardene, Sadiq and Indulska's (2015) comprehensive study of the 127 different dimensions of data quality presents this framework comprehensively, but the issue raised in the interviews was that it is quite impossible to discuss about this many dimensions. Surely, they can be used to identify the specific problem in the data but understanding the overall picture from these might be challenging as they seem individual dimensions but are still greatly interconnected. On the contrary, while the literature provided the framework, it did not enlighten the issue of how to measure them. Based on the interviews, this is one of the challenging parts as there might not be any processes metering the quality to begin with, and eventually creating the suitable meters is even more difficult, since coming up with the suitable cases to meter business benefits was experienced as a major challenge. Nonetheless, measuring was mentioned as the key driver as it secures the funding for the process.

For the latter part of the question, the fundamental issue is that in the previous publications, there was a scarcity of evidence of how much the data quality improves. In most of them, it was mentioned that the quality will shift towards better. This was confirmed through the interviews, as multiple participants mentioned that there is a common consensus inside organisations that the quality of data is bad at the start. However, the most crucial information gained in the interviews was, that there is usually no metering in use before the governance process. This explains the missing data on the previous literature, but for the research question, it also puts some challenges to answer it directly. Luckily, there were few examples of previously used meters and their comparison to the situation afterwards, such as the number of incidents/tickets raised directly and indirectly due to the bad data. The challenge is to point which incidents were truly raised due to data quality, but there was evidence of a reduction in these, indicating that quality does improve. Another case mentioned was the technical meters, which do include measuring the accessibility and availability dimensions of data quality, and in which the quality changes are easily tracked: the duration of ETL-jobs will shorten with better data management and by removing unnecessary records.

Another aspect for this question was that according to interviews, almost every company nowadays has already implemented some sort of governance program and different units inside the organisation can be in completely different levels in terms of data quality and data governance.

This poses a challenge, since the initial situation of the data has been completely different for different units and combining them into a one governed system will make it impractical to compare the quality to the situations beforehand. However, there were mentions that after implemented governance, metering has shown improvements in data quality, and furthermore, if it has shown degrading, pointing the root cause is now possible due correctly implemented data governance.

To conclude these findings to answer the research question, it can said that by establishing a data governance program organisations are able to improve the data quality by making it transparent due implementing metering and defining the ownership, thus making it possible to find the ones who can improve the quality. Properly done, the scope will be also extended to concern the whole company, not just single units further improving the quality overall. To prove any measurable units of how much governance affects quality is besides impossible due to lack of metering but also highly case-dependant.

The second research question for the study was '*What kind of issues there are in the implementation of governance affecting data quality based on the expert interviews?*', but as pointed out in the previous chapter, it cannot be answered directly without the numerical evidence. However, the biggest challenges in the data governance process can be highlighted and their effect on the quality reflected. The first and foremost issue was the adequacy of resources with two separate issues: involving business in the process in order for them to understand the scale of the process from the beginning and securing the continuation of the process with a separate data management office. Without the involvement of the business, they often have some sort of misleading expectations that the data governance process would present something fundamentally new, like data sources, to the system while the actual goal is to improve data quality and clarify data management habits. Because they don't get anything new and the impact on the quality usually takes time to be visible, the business might not start funding the process at all or end it prematurely. This also means that the continuation of the process is in stake. Best way to ensure that the governance stays as a permanent process would be a separate data management office including both the business owners and technical specialists.

To highlight other important aspects affecting data quality, it was found out that the concept of ownership is highly related to succeeding in both data governance and thus data quality. The

topic of ownership is not purely just a problem by its name and doesn't revolve around the issue of who owns the data, but good ownership needs the owner's interest and time. For the governance process, this means that the right people who have the motivation and skills to become data owners, need also to have the time and other resources allocated sufficiently so they can succeed in this area. Furthermore, this circles back to the topic of the adequacy of funding. However, even with unlimited financial funding, data governance needs to have the inner drivers and goals to be an actual working unit.

For the third question, "Are there any significant trends in the data governance process or data quality concept, which are not included or presented differently in the current literature?", the biggest difference was that in the literature, the RACI-standard (Smith and Erwin, 2005) and more structured hierarchy (Wende, 2007, pp. 419–421; Weber, Otto and Österle, 2009) were presented, while the experts all agreed on the idea of a separate data management office, making it more agile and lean. This does not exclude the roles of RACI, but they are all included in the management office and this office is then supported from the company level. The idea that there would be some high-level sponsor actively participating in the process is somewhat naive, and it was also seen by Korhonen et al. (2013, p. 16) whose view was that the presented roles are not sufficient enough to deal with day-to-day activities. Even though a centralized governance process was seen as the best solution to go, and there were not yet direct implications towards decentralized governance, with data management offices the transition towards decentralization seems probable as the shift from 10- to 20-year-old literature is already visible. The main issue to tackle here if this happens, is how to prevent the tribe behaviour and recreation of multiple master records.

Another rather surprising aspect was the possibility to view the relationship of data governance and data quality from both points of view: in the literature, such as Koltay (2016, p. 309), saw that by implementing data governance one can ensure good data quality, but in the interviews, the experts saw this also from a complete different way. For them, data quality itself can tell if there are issues in data governance actions. This does mean pretty much the same, but it was not presented by previous reviews, that the state of governance can be actually told by looking at the data quality it creates, and implementing governance steps doesn't necessarily mean better quality.

6.1 Limitations of the study

The scope of the study presents a challenge for the study's implications as the interview group was formed only from one organization's employees and the total number was as limited as six. However, many of them had worked quite recently also in other companies extending the scope quite much. Furthermore, the participants were also constructed from different nationalities, locations, and professional backgrounds, which also further extends the scope.

Another aspect is that there were some topics which can fulfil the saturation requirement of the qualitative study, where new information is no more presented. But there were still some topics which certainly didn't fulfil this requirement as some notions were only mentioned once. This can raise a question of whether a group interview would have been a better choice, where participants could have interacted together and maybe this way complete each other others answers. But then again this could have left some answers unmentioned as participants would have not been as free to talk as in individual interviews.

With these limitations considered, it still seems rather safe to say that this study does provide a quite wide preview on the topic on hand at current time. This doesn't exclude the possibility of something missing from these conclusions but stated findings can be extended to cover the issue.

6.2 Suggestion for future research

Even though the initial research goal was to review these topics from a broader perspective, for some part this study was not focused enough: the topic of data governance and especially the challenges in implementation were somewhat wider than expected. Mainly this was because of the interconnectivity between various data-related subjects. Breaking down some of the biggest issues, like funding and casting, into more tighter topics could be possible research themes from the data governance point of view. This might be more profitable than studying technical issues related to governance while they are more case-by-case situations, where research results wouldn't be generalized.

From another point of view, for the data quality, the topic for further study can be derived directly from the issue with metering: to study the factual changes in data quality, the researcher needs to have numerical data to support the hypothesis. Probably this would be then done as a

case-study, but challenge in this is the duration of the study, since already the planning and first phases of governance process will take a long period to finish.

7 Bibliography

Abraham, R., Schneider, J. and vom Brocke, J. (2019) 'Data governance: A conceptual framework, structured review, and research agenda', International Journal of Information Management. Elsevier Ltd, pp. 424–438. doi: 10.1016/j.ijinfomgt.2019.07.008.

Al-Badi, A., Tarhini, A. and Khan, A. I. (2018) 'Exploring big data governance frameworks', in Procedia Computer Science. Elsevier B.V., pp. 271–277. doi: 10.1016/j.procs.2018.10.181.

Al-Ruithe, M., Benkhelifa, E. and Hameed, K. (2018) 'Data governance taxonomy: Cloud versus non-cloud', Sustainability (Switzerland), 10(1). doi: 10.3390/su10010095.

Al-Ruithe, M., Benkhelifa, E. and Hameed, K. (no date) 'A systematic literature review of data governance and cloud data governance Biologically-Inspired Highly Resilient Distributed Systems View project Software Defined Systems View project'. doi: 10.1007/s00779-017-1104-3.

Alhassan, I., Sammon, D. and Daly, M. (2019) 'Critical Success Factors for Data Governance: A Theory Building Approach', Information Systems Management, 36(2), pp. 98–110. doi: 10.1080/10580530.2019.1589670.

Askham, N. et al. (2013) The Six Primary Dimensions for Data Quality Assessment, Group, DAMA UK Working. Available at: https://www.dqglobal.com/wp-content/uploads/2013/11/DAMA-UK-DQ-Dimensions-White-Paper-R37.pdf (Accessed: 22 January 2021).

Australian Bureau of Statistics (2020) Statistical Language. Available at: https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Language?OpenDocument (Accessed: 10 December 2020).

Barker, J. M. (2016) 'Data governance: The missing approach to improving data quality', ProQuest Dissertations and Theses, p. 293. Available at: https://login.pal-

las2.tcl.sc.edu/login?url=https://search.proquest.com/docview/1862110139?accountid=13965%0Ahttp://resolver.ebscohost.com/openurl?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/ProQuest+Dissertations+%26+Theses+Global&rft_v (Accessed: 14 March 2021).

Barone, D., Stella, F. and Batini, C. (2010) 'Dependency discovery in data quality', in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Berlin, Heidelberg, pp. 53–67. doi: 10.1007/978-3-642-13094-6_6.

Batini, C. and Scannapieca, M. (2006) Data Quality. Berlin: Springer Berlin Heidelberg (Data-Centric Systems and Applications). doi: 10.1007/3-540-33173-5.

Berson, A. et al. (2010) Master data management and data governance, second edition. McGraw-Hill. Available at: http://carletonu.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwY2AwNtIz0EUrE0xTLYxSgJVbomkisIRkmZpDDpY3NwoBWiCuaUZaPtzuLNRQKhRsCNo7zps3h2ylAOxIqgI6BdQLxHLmYkwRaC2abGRSYU5bMYVso23uMA-zgm2CQyYbkEb3pktDED5wNs5AnywAsRw2KkjoFLVTZCBNRW050CIgSk1T5iBA7YYXYTBy (Accessed: 15 March 2021).

Berson and Dubov (2007) Master data management and customer data integration for a global enterprise, 感染症誌. Available at: https://www.adlibris.com/fi/kirja/master-data-management-and-customer-data-integration-for-a-global-enterprise-9780072263497 (Accessed: 25 May 2021).

Bisbal, J. et al. (1999) Legacy Information System Migration: A Brief Review of Problems, Solutions and Research Issues.

Borgman, H. et al. (2016) 'Dotting the i and crossing (out) the T in IT governance: New challenges for information governance', in Proceedings of the Annual Hawaii International Conference on System Sciences, pp. 4901–4909. doi: 10.1109/HICSS.2016.608.

Bowen, P. L., Cheung, M. Y. D. and Rohde, F. H. (2007) 'Enhancing IT governance practices: A model and case study of an organization's efforts', International Journal of Accounting Information Systems, 8(3), pp. 191–221. doi: 10.1016/j.accinf.2007.07.002.

Brown, A. E. and Grant, G. G. (2005) 'Framing the Frameworks: A Review of IT Governance Research', Communications of the Association for Information Systems, 15, pp. 696–712. doi: 10.17705/1CAIS.01538.

Cheong, L. and Chang, V. (2007) 'Association for Information Systems AIS Electronic Library (AISeL) The Need for Data Governance: A Case Study Recommended Citation "The Need for Data Governance: A Case Study" The Need for Data Governance: A Case Study', Association for Information Systems, p. 999. Available at: http://aisel.aisnet.org/acis2007/100 (Accessed: 3 March 2021).

Cleven, A. and Wortmann, F. (2010) 'Uncovering Four Strategies to Approach Master Data Management Quartierstrom (local peer-to-peer energy market) View project Blockchain-Initial Coin Offerings View project Uncovering four strategies to approach master data management'. doi: 10.1109/HICSS.2010.488.

Das, T. K. and Mishra, M. R. (2011) 'A Study on Challenges and Opportunities in Master Data Management', International Journal of Database Management Systems, 3(2), pp. 129–139. doi: 10.5121/ijdms.2011.3209.

Donaldson, A. and Walker, P. (2004) 'Information governance-a view from the NHS', International Journal of Medical Informatics, 73, pp. 281–284. doi: 10.1016/j.ijmedinf.2003.11.009.

Dong, X. L., Halevy, A. and Yu, C. (2009) 'Data integration with uncertainty', VLDB Journal, 18(2), pp. 469–500. doi: 10.1007/s00778-008-0119-9.

Downing, S. M. (2003) 'Validity: On the meaningful interpretation of assessment data', Medical Education, pp. 830–837. doi: 10.1046/j.1365-2923.2003.01594.x.

Dyché, J. and Levy, E. (2006) Customer Data Integration: Reaching a Single Version of the Truth (SAS Institute Inc.). Available at: http://web.a.ebsco-host.com.ezproxy.cc.lut.fi/ehost/detail/detail?vid=2&sid=87ba10e9-507d-4223-ae48-04490c5181e1%40sdc-v-sess-mgr03&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3D%3D#AN=165127&db=e000xww (Accessed: 25 January 2021).

English, L. P. (2009) Information quality applied: Best practices for improving business information, processes and systems, Best Practices for Improving Business Information …. Wiley Publishing Inc.

Eppler, M. J. (2006) Managing information quality: Increasing the value of information in knowledge-intensive products and processes, Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes. doi: 10.1007/3-540-32225-6.

Fan, W., Geerts, F. and Wijsen, J. (2012) 'Determining the currency of data', in ACM Transactions on Database Systems. ACM PUB27 New York, NY, USA, pp. 1–46. doi: 10.1145/2389241.2389244.

Friedman, T. (2006) 'Gartner Study on Data Quality Shows That IT Still Bears the Burden IT Is Still Perceived as the Owner of Data Quality', (February). Available at: https://www.gartner.com/en/documents/489562/gartner-study-on-data-quality-shows-that-it-still-bears-t (Accessed: 24 February 2021).

Galletta, A. and Cross, W. E. (2013) Mastering the semi-structured interview and beyond: From research design to analysis and publication, Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication. doi: 10.5860/choice.51-2430.

Gatling, G., Stefani, H. and Weigel, G. (2012) Enterprise Information Management with SAP.

Van Grembergen, W., De Haes, S. and Guldentops, E. (2004) Strategies for Information Technology Governance, Strategies for Information Technology Governance. doi: 10.4018/978-1-59140-140-7.

Griffin, J. (2005) 'Data Governance: A Strategy for Success', DM Review, 15(6), p. 49. Available at: https://search-proquest-com.ezproxy.cc.lut.fi/docview/214678744/fulltext/C5BB7410FCA84D2EPQ/1?accountid=27292 (Accessed: 1 March 2021).

Guillaume, S. (2001) 'Designing fuzzy inference systems from data: An interpretability-oriented review', IEEE Transactions on Fuzzy Systems, 9(3), pp. 426–443. doi: 10.1109/91.928739.

de Haes, S. and van Grembergen, W. (2009) 'An Exploratory Study into IT Governance Implementations and its Impact on Business/IT Alignment', Information Systems Management, 26(2), pp. 123–137. doi: 10.1080/10580530902794786.

Halevy, A. (2005) 'Why Your Data Won't Mix', Queue, 3(8), pp. 50–58. doi: 10.1145/1103822.1103836.

Heinrich, B., Kaiser, M. and Klier, M. (2007) 'How to measure data quality? - A metric-based approach', in ICIS 2007 Proceedings - Twenty Eighth International Conference on Information Systems. Available at: https://www.researchgate.net/publication/200047424_How_to_measure_data_quality_-_A_metric_based_approach (Accessed: 22 January 2021).

Howard, P. (2011) 'Data Migration', in SpringerReference. Berlin/Heidelberg: Springer-Verlag, p. 17. doi: 10.1007/springerreference_64701.

Jayawardene, V., Sadiq, S. and Indulska, M. (2015) An Analysis of Data Quality Dimensions. Available at: https://espace.library.uq.edu.au/data/UQ_312314/UQ312314_UPDATED_2015_02.pdf?dsi_version=ef3ceaacb5d3741215cbbe5d5fed2af2&Ex-

pires=1607099785&Key-Pair-Id=APKAJKNBJ4MJBJNC6NLQ&Signature=VxDQbCqA2v8UTczXSrvFHCbGYiZ~RjtGyhq-wVWrpSDUf97wvOVkUqkOSGxYWWcUAaBh5b5 (Accessed: 4 December 2020).

Khatri, V. and Brown, C. V. (2010) 'Designing data governance', Communications of the ACM, 53(1), pp. 148–152. doi: 10.1145/1629175.1629210.

Kimball, R. and Caserta, J. (2004) The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning Conforming, and Delivering Data, Wiley. Available at: http://au.wiley.com/WileyCDA/Section/index.html%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Data+Warehouse+ETL+Toolkit+Practical+Techniques+for+Extracting,+Cleaning,+Conforming,+and+Delivering+Data#1%5Cnhttp://scholar.google.com/s.

Koltay, T. (2016) 'Data governance, data literacy and the management of data quality', IFLA Journal, 42(4), pp. 303–312. doi: 10.1177/0340035216672238.

Korhonen, J. J. et al. (2013) 'Designing data governance structure: an organizational perspective', GSTF Journal on Computing, 2(4), pp. 11–17. Available at: https://search.proquest.com/openview/27266e14fdc66859e34abda0877d07b4/1.pdf?pq-origsite=gscholar&cbl=1036337 (Accessed: 1 March 2021).

Ladley, J. (2013) Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program, Journal of Chemical Information and Modeling.

Loshin, D. (2001) Enterprise knowledge management: The data quality approach. Morgan Kau.

Loshin, D. (2008) Master Data Management, Master Data Management. doi: 10.1016/B978-0-12-374225-4.X0001-X.

McGilvray, D. (2008) Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. 1st edn. Morgan Kaufmann. Available at: https://www.elsevier.com/books/executing-data-quality-projects/mcgilvray/978-0-12-374369-5 (Accessed: 19 January 2021).

Microsoft (2016) General availability: Azure Data Catalog . Available at: https://azure.microsoft.com/en-us/updates/general-availability-azure-data-catalog/ (Accessed: 13 May 2021).

Morabito, V. (2015) Big data and analytics: Strategic and organizational impacts, Big Data and Analytics: Strategic and Organizational Impacts. Springer International Publishing. doi: 10.1007/978-3-319-10665-6.

Morris, J. (2006) 'Practical Data Migration', Kybernetes. 2nd edn, 35(9), p. 248. doi: 10.1108/k.2006.06735iae.008.

Morris, J. (2012) Practical Data Migration. Second, Kybernetes. Second. doi: 10.1108/k.2006.06735iae.008.

Oberg, J. (1999) Why the Mars probe went off course, IEEE Spectrum. doi: 10.1109/6.809121.

Olson, J. E. (2003) Data Quality: The Accuracy Dimension, Data Quality: The Accuracy Dimension. doi: 10.1016/B978-1-55860-891-7.X5000-8.

Otto, B. (2011) 'Organizing Data Governance: Findings from the telecommunications industry and consequences for large service providers', Communications of the Association for Information Systems, 29(1), pp. 45–66. doi: 10.17705/1cais.02903.

Otto, B. (2013) 'On the evolution of data governance in firms: The case of Johnson & Johnson consumer products North America', in Handbook of Data Quality: Research and Practice. Springer Berlin Heidelberg, pp. 93–118. doi: 10.1007/978-3-642-36257-6_5.

Panian, Z. (2010) ‘Some practical experiences in data governance’, World Academy of Science, Engineering and Technology, 38, pp. 150–157.

Pierce, E., Dismute, W. S. and Yonke, C. L. (2008) ‘The State of Information and Data Governance - Understanding How Organizations Govern Their Information and Data Assets’, (April), pp. 1–39.

Price, R. J. and Shanks, G. (2005) ‘Empirical refinement of a semiotic information quality framework’, in Proceedings of the Annual Hawaii International Conference on System Sciences, p. 216. doi: 10.1109/hicss.2005.233.

Ranganathan, K., Iamnitchi, A. and Foster, I. (2002) ‘Improving data availability through dynamic model-driven replication in large peer-to-peer communities’, in 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, CCGrid 2002. doi: 10.1109/CCGRID.2002.1017164.

Rasouli, M. R. et al. (2016) ‘Information Governance as a Dynamic Capability in Service Oriented Business Networking’. doi: 10.1007/978-3-319-45390-3_39.

Rau, K. G. (2004) ‘Effective governance of it: Design objectives, roles, and relationships’, Information Systems Management, 21(4), pp. 35–42. doi: 10.1201/1078/44705.21.4.20040901/84185.4.

Redman, T. C. (1997) Data Quality for the Information Age. Artech House, Inc.

Redman, T. C. (2009) ‘Data driven: profiting from your most important business asset’, Choice Reviews Online, 46(06), pp. 46-3345-46–3345. doi: 10.5860/choice.46-3345.

Redman, T. C. (2013) ‘Data’s credibility problem’, Harvard Business Review, (DEC).

Rich, R. C. et al. (2011) Empirical Political Analysis: Quantitative and Qualitative Research Methods, Routledge.

Rizkallah, J. (2017) ‘Council Post: The Big (Unstructured) Data Problem’, Forbes. Available at: https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/?sh=1452fd67493a (Accessed: 28 January 2021).

Rowley, J. (2007) ‘The wisdom hierarchy: representations of the DIKW hierarchy’, Journal of Information Science, 33(2), pp. 163–180. doi: 10.1177/0165551506070706.

Scannapieco, M. and Catarci, T. (2002) Data Quality under the Computer Science perspective, Computer Engineering. Available at: https://www.researchgate.net/profile/Tiziana_Catarci2/publication/228597426_Data_quality_under_a_computer_science_perspective/links/0fcfd51169a156b61a000000.pdf (Accessed: 21 January 2021).

Sen, A. (2004) ‘Metadata management: Past, present and future’, Decision Support Systems, 37(1), pp. 151–173. doi: 10.1016/S0167-9236(02)00208-7.

Smith, B. M. L. and Erwin, J. (2005) ‘Role & Responsibility Charting (RACI)’, Smith, Michael L Erwin, James, pp. 1–14.

Soares, S. (2013) ‘Big Data Governance’. Available at: www.information-asset.com (Accessed: 2 March 2021).

Stvilia, B. et al. (2007) ‘A framework for information quality assessment’, Journal of the American Society for Information Science and Technology, 58(12), pp. 1720–1733. doi: 10.1002/asi.20652.

Tallon, P. P., Short, J. E. and Harkins, M. W. (2013) ‘The evolution of information governance at intel’, MIS Quarterly Executive, 12(4), pp. 189–198.

Tiwana, A., Konsynski, B. and Venkatraman, N. (2013) ‘Special issue: Information technology and organizational governance: The IT governance cube’, Journal of Management Information Systems, 30(3), pp. 7–12. doi: 10.2753/MIS0742-1222300301.

Trochim, W. M. . (2020) Research Methods Knowledge Base. Available at: https://conjointly.com/kb/ (Accessed: 23 January 2021).

Ward, J. S. and Barker, A. (2013) Undefined By Data: A Survey of Big Data Definitions. Available at: http://bigdatawg.nist.gov/home.php. (Accessed: 11 March 2021).

Watson, H. J., Fuller, C. and Ariyachandra, T. (2004) 'Data warehouse governance: Best practices at Blue Cross and Blue Shield of North Carolina', Decision Support Systems, 38(3), pp. 435–450. doi: 10.1016/j.dss.2003.06.001.

Watson, J., Kraemer, S. and Thorn, C. (2009) 'Data Quality Essentials Guide to Implementation: Resources for Applied Practice.', in.

Weber, K., Otto, B. and Österle, H. (2009) 'Ones size does not fit all -A contingency approach to data governance', Journal of Data and Information Quality, 1(1), pp. 1–27. doi: 10.1145/1515693.1515696.

Wende, K. (2007) Association for Information Systems AIS Electronic Library (AISeL) A Model for Data Governance-Organising Accountabilities for Data Quality Management Recommended Citation Wende, Kristin, "A Model for Data Governance-Organising Accountabilities for Data Q. Available at: http://aisel.aisnet.org/acis2007/80 (Accessed: 1 March 2021).

Wende, K. and Otto, B. (2007) 'A contingency approach to data governance', in Proceedings of the 2007 International Conference on Information Quality, ICIQ 2007.

White, A. et al. (2006) 'Mastering Master Data Management'.

Winter, R. and Meyer, M. (2001) 'Organization of Data Warehousing in Large Service Companies - A Matrix Approach Based on Data Ownership and Competence Centers', Journal Of Data Warehousing, Vol. 6(No. 4), pp. 23–29.

Zviran, M. and Glezer, C. (2000) 'Towards generating a data integrity standard', Data and Knowledge Engineering, 32(3), pp. 291–313. doi: 10.1016/S0169-023X(99)00042-7.

8 Appendices

Appendix 1 – Introduction to the topic for participants

Introduction to the topic:

This study is conducted to find out the missing linkage between the data governance and data quality as quality improvements are mentioned many times as one of the results of succeeded governance process, while the evidence of this is less presented. To ensure that both me and You are discussing with actual same terms, I have concluded key terms and how I refer to them:

- Data governance: framework for managing data as an enterprise asset. Furthermore, data governance formalizes data policies, standards, and procedures and monitors compliances and decision rights
- Data quality: In tightest definition data quality can be described by concluding that it is a state when data is usable for its planned use. It can be measured through different dimensions:
    - *Accessibility & Availability:* speed and ease of locating and obtaining needed information & amount of time data is accessible when needed
    - *Coverage & completeness:* covers whole phenomenon & no values are missing
    - *Accuracy:* ratio of data values and identified correct data values
    - *Currency & timeliness:* how correct data is in spite of change in time & time difference between incident and record available
    - *Validity:* how a certain statement/record represents the real world
    - *Interpretability:* values leave no possibility for misunderstanding
    - *Consistency:* there are no conflicting values in datasets

There might be other definitions for data quality and there is probably more deeper levels in data governance process than questions will cover. However, the point of this study is to find connection between concluded governance processes and changes in data quality, so please consider this will thinking answers. In the interview, I hope that we can discuss freely on the questions and answers, so I am not expecting direct, pre-thought answers.

Appendix 2 – interview questions

Questions:

Could You really briefly describe Your role and main tasks?

(This is just to divide interviewees into groups, like technical, managing etc.)

1) Could You define briefly data governance, quality and explain their relationship in Your own words?
2) From Your point of view, which are the goals for governance process?
3) What has been Your role in governance process? As a vendor or customer in the project You have attended?
4) Which do You prefer to be the key areas/dimension to take into account while thinking data quality?
    a) For example, if You had to choose top three from the dimensions listed below?
        i) Accessibility & Availability
        ii) Coverage & completeness
        iii) Accuracy
        iv) Currency & Timeliness
        v) Validity
        vi) Interpretability
        vii) Consistency
        viii) Other?
5) Based on Your experience, what are the most common issues with data?
    a) Before governance?
    b) After governance?
6) How are the changes in data quality after governance?
    a) Do these changes (/improvements) correspond to expectations from vendor's/customer's side?
    b) Are there any metrics etc. used?
7) Which tools are used in the processes and is there any difference between them in i.e. usability?
8) Are there any best practices that organization should adopt after governance to ensure data quality in future?

9) From my point of view, these are the questions I like to include. Do You have further questions for me or is there any aspects You like to point out that is missing from this?