

Upon a Data Governance framework which uses data categorization

Nick Smidt

June 2021

Nick Smidt

5617294

n.smidt2@students.uu.nl

Master Business Informatics – Faculty of Science – Utrecht University

First supervisor
Nico Brand
n.a.brand@uu.nl

Second supervisor
Jan-Martijn van der Werf
j.m.e.m.vanderwerf@uu.nl

external supervisor
Niels van den Broek
niels.vandenbroek@transavia.com

Abstract

Data governance is a new field of study compared to IT governance. There are few frameworks that help organizations help implement data governance. This study was performed in collaboration with Transavia. A new data governance approach and framework is designed in this study to help Transavia govern their data. The framework is derived from data governance and data classification literature. The conceptual framework is validated to find its suitability. The validation resulted in feedback and discussion points for the framework. The process of conducting this study has helped Transavia discuss and align data related objectives between stakeholders. The study is concluded with future research and improvements for the designed conceptual framework.

Contents

1	Introduction	4
1.1	Introduction	4
1.2	Research Object	5
1.3	Scope	7
1.4	Summary	7
2	Research plan	8
2.1	Problem investigation and problem statement	8
2.2	Purpose	9
2.3	Research questions	10
2.4	Approach	11
2.5	Interview approach	12
2.6	Literature review approach	12
2.6.1	Purpose	13
2.6.2	Draft Protocol	13
2.6.3	Practical screening of literature	13
2.6.4	Literature search	14
2.7	Validation	15
2.8	Planning	15
2.9	Summary	16
3	Data classification	17
3.1	data classification properties	17
3.2	Category classification	19
3.2.1	Accessibility	21
3.2.2	Access security	21
3.2.3	Availability	22
3.2.4	Ease of operation	22
3.2.5	Added Value	22
3.2.6	Relevance	22
3.2.7	Ease of understanding	22
3.2.8	Reliability	23
3.3	Classification framework	23
3.4	Summary	25
4	Data governance	26
4.1	What is data governance?	26
4.2	Data governance vs IT governance	27
4.3	Data governance frameworks	27
4.3.1	Abraham and schneider	27
4.3.2	Khatri and Brown	32
4.3.3	Ladley	34
4.3.4	Panian	35
4.4	core elements of data governance	37

4.5	closing and continuation	39
4.6	Summary	39
5	Data governance framework	41
5.1	Overarching elements and structure	41
5.2	Decision domains	42
5.3	classification of data	43
5.4	Classification of mechanisms	44
5.4.1	Structural mechanisms	44
5.4.2	procedural mechanisms	45
5.5	Category 1 low access low value	47
5.6	Category two high access low value	48
5.7	Category 3 low access high value	50
5.8	Category 4 Core data high access high value.	52
5.9	Summary	54
6	Validation	56
6.1	Tackling of validity threats	56
6.2	Research validity	56
6.3	Expert interviews	57
6.3.1	Interview one Albert Sprokholt	57
6.3.2	Interview two Peter Kromhout	59
6.4	artefact validation within the research object	61
6.4.1	Workshop results	61
6.4.2	Results and feedback	62
6.5	summary	63
7	Discussion and conclusion	64
7.1	Discussion	64
7.1.1	Subjectivity in classification of data	64
7.1.2	Complexity of the framework	65
7.1.3	Overarching structures	65
7.1.4	Decentral vs central organizations	66
7.1.5	addition of critical successfactors	66
7.1.6	interpretation of discussion	67
7.2	Conclusion	67
8	Future research and acknowledgement	70
8.1	future research	70
8.2	acknowledgement	70
A	Interview Questions	75

1 Introduction

1.1 Introduction

Data is the most valuable property of the 21st century. Perrons and Jensen state that significant strategic advantage will arise from an organization's ability to be the integrator and high-level analyzer of data within its ecosystem [Perrons and Jensen, 2015]. Many organizations have recognized the value and importance of data. Data may help organizations be more competitive, improve processes and gain knowledge. More and more organizations strive to increase the value of the data they store and process. To increase the value of data as a business asset, organizations need to establish standards, policies, and processes for the usage, development, and management of data, to create the right organizational structure, and to develop the supporting technology infrastructure [Panian, 2010]. Data governance can be a solution to provide the standards, policies, and processes for the usage, development, and management of data. Data governance is an emerging subject within the information system field. And can be described as: "The exercise of authority, control, and shared decision making (planning, monitoring and enforcement) over the management of data assets." Data governance includes High-level planning and control over data management and should be complied by every-one within an organization.[DAMA, 2009]. Data governance can provide a handhold for implementing data rules and guidelines within an organization, help an organization increase their data quality and sometimes enable new revenue opportunities. Potential benefits of a data governance program are:

- increased data quality
- lower data management costs
- data process standardisation
- increased data security
- increased access to needed data for analysts and business users
- improved business decision-making [Stedman and Vaughan, 2020]

An effective Data Governance program reduces costs and increases the quality of the data, which makes better decision making, thereby increasing the profitability of organizations [Oudemans,]. While achieving these goals are attractive in theory, putting data governance into actual practice requires a compelling business driver [Panian, 2010]. Implementing data governance is challenging because it affects all employees of an organization and it often involves organizational and sometimes cultural change. The benefits of a data governance program are high even though implementing it can be challenging. Many benefits of data governance are hard to measure for example improved decision. This makes demonstrating the business value of data governance also challenging [Stedman and Vaughan, 2020].

Organizations often process multiple types of data. This data they may differ in sensitivity level, the way it is processed or the time it is relevant. these different types of data may not be optimally governed with the same rules and policies. This makes data governance even more challenging since there might not be a one size fits all solution for these different data types. Nevertheless data governance can reduce data related risks and can improve an organization's strategic position if implemented well. The goal for this study is to design a data criteria based data governance framework that can be implemented into any medium to large organisation. The Design and engineering cycle is used to design and test the framework [Wieringa, 2014]. An extensive literature review and expert interviews are used to gather knowledge about data governance and design the framework. The main objective for the designed framework will be to categorize data types within an organization. The data type categories will be labeled with a color to make the categories more tangible. Policies rules and guidelines will be designed for for each individual category. The aim for this approach study is to make implementing data governance less challenging for medium to large organisation which process different data types.

1.2 Research Object

This study is performed in collaboration with Transavia. Transavia is the largest Low cost airline in the Netherlands and a wholly owned subsidiary of KLM. 2020 was a tough year for the Aviation industry, this has also affected Transavia. "The corona crisis has taken a serious blow to our strong financial position that we have carefully built up over the past few years" [Strategie 2020-2024, 2020]. Transavia has introduced a new strategy plan until 2024. The strategy consists of 3 phases. 2021 is called the reset year. In this year the product portfolio will be simplified and made manoeuvrable. In 2021, a new foundation will also be laid for a new organizational structure. 2022 will be the relaunch year for Transavia. In this year the revenue will be strengthened by scaling core activities and by introducing new product initiatives. The last year of the strategy focuses on regaining and maintaining their number 1 position of the largest low cost airline in the Netherlands. Transavia aims to do this by benefit from their varies portfolio and with sources of income within and outside their core activities. In parallel, this year is all about anchoring sustainable growth through increased flexibility, lower cost structure and integrated planning.

In the past Transavia's strategy has been to outsource costs to external parties in order to avoid them long term. Many external domain experts, architects and consultants have helped Transavia in the past. Examples have been developing new software, or moving source software to the cloud. Attracting and outsourcing these project to external experts has led to Transavia becoming the largest Low cost airline in the Netherlands. On the other hand this has led to a lack of insight into their data and software landscape. Transavia has faced data related issues / conflicts in the past. Transavia struggles to solve these data related

issues / conflicts. Examples are: that there is no clear data source ownership therefor which has led to no-one feeling responsible for fundamental source systems data, while this data was essential for day-to-day business. Transavia has started a project to optimize data processing since the introduction of the GDPR in 2018. Since 2018 Transavia's data goal is to regain insight into the data they process, where data is stored and which applications use data sources. The Architecture team of Transavia has analyzed Transavia's IT landscape in order to produce a source of truth document. The source of truth document is a formal bundle of models that describe the IT Architecture of Transavia. This document features business objects, the software applications that use these objects and the interaction between the applications. The business objects are derived from the data objects present within the IT landscape. The business objects are validated by the business-unit of Transavia. Transavia's next data goal is to use the gained insight to improve business IT alignment. Business IT alignment aims to improve the collaboration between the IT and Business departments and use IT in order to achieve business objectives. As stated before 2021 is a reset year for Transavia where new foundation will be laid for a new organizational structure. The new organizational structure introduces a new division of tasks and responsibilities for employees. Transavia would like to improve Business IT alignment and data risk management by implementing a data governance program. The goal for the data governance program will be to assign ownership to data sources and introduce standardized procedures rules and guidelines to optimally process data collaboratively and structured between teams.

This study will design a general data governance framework that can be implemented into any medium to large sized organization. This study will help Transavia govern data by optimizing the designed a data governance framework to their requirements. The optimization process of the framework for Transavia will start the conversation between the business and IT departments of Transavia. This will optimally result in improving the business IT alignment. The data governance rules guidelines and policies will be derived from the requirement given by both departments. The optimized data governance framework has a requirement that it should be easy to implement within the Transavia culture to ensure adoption of the commercial domain. Designing such a model and setting up a program is easier said than done. Data governance programs almost always involves cultural and behavioral change within an organization. Cultural and behavioral change makes adoption by employees harder since they might counter-work changing their behaviour. An organization and its employees should have the right mindset and capacity for changing and optimizing its behavior and culture to comply to the data governance program. Yet Transavia is building a new organizational structure in 2021, which means many employees already have to alter their working behaviour. This provides an excellent opportunity to introduce data governance within Transavia and teach the employees how to better process and handle data using to the data governance framework.

1.3 Scope

The objective of this study is to design a data governance framework based on data categorisation. The framework is intended for organizations that process multiple types of data that can not be governed the same. The framework will be designed to be applied to any medium to large sized organization. Small organisations process and store less data compared to medium and large organization and therefore often do not have the requirement to implement data governance. The data governance framework is therefore intended for medium to large organizations that do not know how to set up data governance. A medium sized organization is defined as an organization which employs between 100 and 500 employees. An organization with less than 100 employees is considered a small sized organization and an organization which employees more than 500 people is considered large. The framework will not be domain specific but the second part of this study will focus on the aviation domain in order to implement and optimize the framework to the research object. The framework will be made applicable to the aviation domain in order to ease and improve validation and adoption of the research object. The aviation domain is a fully competitive market where the flight portfolio and flight prices are fundamental for success. The data sensitivity is very diverse the flight portfolio is freely accessible to anyone but the passenger and booking data is strictly confidential. The varying data sensitivity make data governance more challenging to implement because some data sources need to be more governed than others. Elaborating and optimizing the data governance framework to the aviation domain and research object will provide other domains information for how to make the framework applicable to their domain if necessity. The data governance framework will be tested to the commercial domain of the IT landscape of the research object. The commercial domain of the research object is defined as all teams, processes and data that support the flight booking process. The data governance framework will be tested with data associated with booking a flight for example the fare price, date and, passenger data of a flight.

1.4 Summary

This chapter emphasized how valuable data can be as well as the importance of Data Governance. Data is essential for organizations to be competitive. Organizations often process many different types of data that can not be governed the same. This chapter demonstrated the benefits of data governance, it's challenges and lack of data governance frameworks. Data governance can help prevent business related problems in advance and improve and organization's business IT alignment. Next the research object of the study was introduced. The research object of the study is Transavia. Transavia is the largest low-cost airline in the Netherlands and have the desire to implement data governance within their organization. The chapter was concluded with setting the scope of the study. The scope of this study will be to design a new data governance framework for any medium to large sized organization that does not know how

to start implementing data governance.

2 Research plan

This chapter outlines the approach for conducting this study. The problems of data governance are stated and investigated. Next a solution to solve these data governance related problems is researched and proposed. Lastly the approach is validated using a validity matrix.

2.1 Problem investigation and problem statement

many organizations have become aware of the increasing importance of governing their data, to ensure confidentiality, integrity, quality, and availability of customers' data [Groß and Schill, 2012]. Good data governance can help organizations to create a clear mission, achieve clarity, increase confidence of using the organizational data, establish accountabilities, maintain scope and focus, and define measurable successes.[Begg and Caira, 2012] Moreover A well implemented data governance program can help organizations remain competitive and agile [Panian, 2010] Since 2018 all organization in Europe have to comply to the GDPR. The General Data Protection Regulation (GDPR) is a legal framework that sets guidelines for the collection and processing of personal information from individuals who live in the European Union (EU) [GDP,]. The introduction of the GDPR provided a legal incentive for organizations to improve processing, storage and governance of data. Yet, the lack of literature and the lack of an industry standard has led many organizations to grope in the dark.

Data governance is an actual field of study. The roots of data governance research can be traced back to early 1980s; however, the first efforts to propose a framework for data governance that were published in 2007 [Niemi, 2013].The great interest in data governance by industry is not matched by the academic community. The academics currently working in this area do appear to agree that there is no single approach to the implementation of data governance in all enterprises [Begg and Caira, 2012]. Fu et al state a lack or even the absence of data governance frameworks [FU et al., 2011]. "Despite its highly recognized importance, the area of data governance is still under developed and under researched. There are only a few frameworks for data governance mainly developed by industry associations such as DAMA, DGI and IBM" [Al-Ruithe et al., 2019]. Other data governance frameworks proposed in literature are mostly conceptual. There are not many articles that implement their proposed framework as a case study. Often data governance frameworks will focus on a single type of architecture for example data governance for cloud. These frameworks can be implemented easily within small organization with only a single or few data sources stored with a single architecture. The same does not hold true for larger organizations. Many large organizations have an hybrid architecture where data is partially stored in-house and in the cloud. Therefore most proposed frameworks

are only applicable for a subset of data within the organization. This makes implementing data governance become more complex for large organizations since different rules and guidelines may apply between different architectures. As stated earlier data governance often involves organizational and sometimes cultural change within an organisation. It can be hard to implement change because employees might not want to change the way they work. Everybody within an organisation should be on board in order to effectively implement change.

The lack of data governance frameworks and the challenges it comprises makes it a new interesting field of study. The objective for this study is to design a framework for data governance which splits the governance for categorized data sources based on criteria. The goal for this study is to overcome some data governance related challenges, ease implementing data governance and reduce the lack / absence of data governance frameworks. The designed framework should give an high-level overview and approach for organizations how to design data governance using a data categorized approach. The framework should be easy for an organisation to implement. Secondly, The Framework should be simple, transparent and unambiguous to ensure adoption, compliance and adaptations to industry standards.

2.2 Purpose

The purpose of this study is to design a new data governance framework. The framework will provide a new approach to data governance. Data governance is understudied but a very important and relevant topic for industry. "Despite its highly recognized importance, the area of data governance is still under developed and under researched. There are only a few frameworks for data governance mainly developed by industry associations such as DAMA, DGI and IBM" [Al-Ruithe et al., 2019]. Data governance being an understudied topic is unfortunate for industry but provides an excellent opportunity to propose a new framework and fill a gap in literature. Testing the framework with the research object will add a data governance case studies to literature as well. The social contribution of this study emphasize the importance of data governance, and help organizations assigning responsibilities and procedures for and to data. The objective for this study is to make it easier for an organization to implement data governance and provide a handhold to start data governance.

2.3 Research questions

The Main research question for this study is:

MRQ What would be a data governance framework that is based upon data classification?

The sub questions for this study are:

RSQ1 What is data governance and what is the benefit of data governance?

RSQ2 What are the current solutions and frameworks for data governance?

RSQ3 what are the core elements of data governance?

RSQ4 What data properties are most suitable for differentiating data categories?

RSQ5 What is a suitable data classification framework?

RSQ6 What is a suitable draft framework?

RSQ7 Can the draft framework be validated as a data governance framework based on data classification?

Research question (RSQ)	Approach	Chapter/Section
How can we design a data governance framework that uses data classification?	Literature review, expert interviews, answers of sub questions	3,4 and 5
1. What is data governance and what is the benefit of data governance?	Literature review	4.1
2. What are the current solutions and frameworks for data governance?	Literature review	4.3
3. What are the core elements of data governance?	Literature review, Expert interview	4.4
4. What data properties are most suitable for differentiating data categories?	Literature review	3.1 and 3.2
5. What is a suitable data classification framework?	Literature review	3.3
6. What is a suitable draft framework?	Literature review, expert interviews, answer of sub questions	5
7. Can the draft framework be validated as a data governance framework based on data classification?	Expert interviews workshop	6

Table 1: Research question and approach table

Answering the sub questions will ease the design process of the data governance framework. The answers to the sub questions gained from performing

a literature review and conducting expert interviews. The first sub question is used to gain knowledge about what data governance entail, to learn about its history and to learn what can be achieved with it. The second sub-question is used to gain information about currently existing data governance frameworks and will also be answered by the Literature review. the third sub question is used to compare the existing frameworks for data governance and to track down the the fundamental elements of data governance. the forth and fifth frame sub questions are about data categorisation. These last 2 sub questions are used to gain knowledge about data classification in order to use this knowledge in the designed framework proposed by this study.

2.4 Approach

The data governance framework is designed by using the design and engineering cycle of Wierenga. The artefact will be designed to be applicable and easily implemented for any medium to large size organisation. The data governance artefact is tested against the research object which will be Transavia's data landscape. The design and engineering cycle consists of the following tasks:

- Problem investigation: What phenomena must be improved? Why?
- Treatment design: Design one or more artifacts that could treat the problem.
- Treatment validation: Would these designs treat the problem?
- Treatment implementation: Treat the problem with one of the designed artifacts.
- Implementation evaluation: How successful has the treatment been? This may be the start of a new iteration through the engineering cycle. [Wieringa, 2014]

Section 2.1 emphasized the importance of data governance and concluded it being an understudied subject. Only a few data governance frameworks exist. Designing a new data governance framework might enable new organizations to use it. The artefact designed in the treatment design phase will be a new data governance framework that aims to solve the data governance problems stated in Section 2.1. The artefact will be deduced from literature and expert interviews. The Literature review and expert interviews will provide information to answer the research questions. RSQ1 and RSQ2 provide input for RSQ3. RSQ4 provides input for RSQ5. The inswers to all reseach question will provide input for the initial artefact design. Next the designed artefact will be tested within the commercial domain of the research object. This study will be ended with evaluating the treatment implementation. this is done by reviewing feedback from employees of the research object. A schematic overview of the study approach can be found in figure 1.

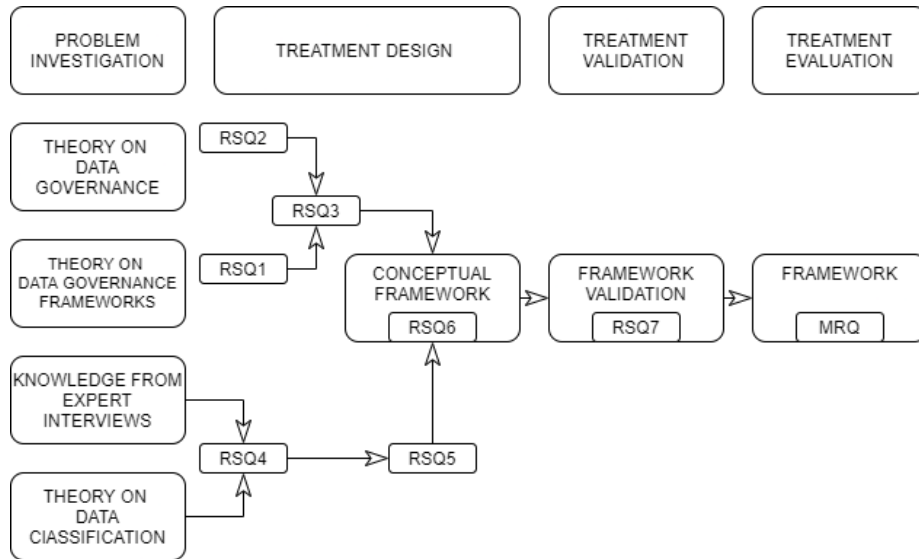


Figure 1: Schematic presentation of the research

2.5 Interview approach

Expert interviews are one of the two information sources used in this study. The expert interviews will be semi structured interviews with IT consultants and stakeholders of the research object. A semi-structured interview is a meeting in which the interviewer does not strictly follow a formalized list of questions. Instead, they will ask more open-ended questions, allowing for a discussion with the interviewee rather than a straightforward question and answer format [Doyle, 2020]. A list of initial questions is made to guide the interview towards the topic of data governance. A list of initial questions will be included into the appendix of the research. The interview will be semi structured to enable the interviewer too zoom in on to certain topics within the interview. This will be improve the information and knowledge gained by conducting the interview. Secondly semi structured interviews are used in this study to encourage two way communication between the interviewer and the interviewee. Data governance has a human aspect. It often involves cultural and behavioral change of stakeholders. The semi structured interviews are used as a tool to improve and emphasize the importance of data governance which will improve the overall outcome of the study.

2.6 Literature review approach

This section will explains the approach, literature collection, screening and reviewing method for this study. This study will use 2 information gaining methods to answer the research questions. Expert interviews are used to gain prac-

tical knowledge about data governance. A literature review is used to gain theoretical knowledge about data governance. The eight step approach proposed by Okoli is used to perform the literature review [Okoli, 2015]. This framework presents a step-by-step approach to carrying out the rigorous, scientific methodology of a systematic literature review. The approach is very general and is applicable to many fields of study however the guide focuses specific on the information systems field. The eight steps for conducting a systematic literature are: 1. **Identify the purpose:** the First step identifies the review's purpose and intended goals in order for it to be explicit to its readers 2. **Draft protocol and train the team:** the second step only applicable if the review is performed by more than one reviewer. In this step all reviewers agree to the procedure followed during the review. This ensures consistency in the execution of the review. 3. **Apply practical screen:** The third step is inclusion screening. In this step explicit literature requirements for are stated in order to be included in the study. 4. **Search for literature:** in the fourth step details of the literature search are described in order to assure the search's comprehensiveness 5. **Extract data:** This step extracts the applicable information from each study found in step 5 that meet the requirements step in step 4 6. **Appraise quality:** in the sixth step the quality of the found literature is reviewed and studies of insufficient quality are excluded. 7. **Synthesize studies:** the seventh step is to analyze the facts extracted from the literature 8. **Write the review:** The final step is to write a review about the gained knowledge found in the literature.

2.6.1 Purpose

The goal for the Literature review is to gain knowledge about existing data governance frameworks and . The framework will be founded on Transavia's requirements but can be implemented within any organisation. The purpose of the Literature review is to gain knowledge and understanding about data governance and its existing frameworks. Having knowledge and understanding will make designing the data governance framework much easier. The existing frameworks will compared with each other in order to find similarities and differences.

2.6.2 Draft Protocol

This study will not feature a draft protocol since the literature review is performed by a single reviewer.

2.6.3 Practical screening of literature

This section will describe the selection and screening protocol for the literature. Firstly a selection of search engines, data sources, keywords and criteria are determined. Standardizing these terms will ensure repeat-ability of the study and provide trace-ability for found literature. The search engines used for finding

literature are Google Scholar and Mendeley. A List of several key words are used to query these search engines. A potential article should at least feature one of keywords listed in order to be relevant for this study. The key words used to query the search engines are: "data governance", "corporate governance", "data governance framework", "data asset management", "data classification" The quality requirements for the literature are: **1** be freely available with UU library access. **2** not be older than 10 years. **3** Must appear on the first page of the search engine's result. Quality appraisal criteria and requirements are used to ensure the found literature is aligned with this study. A Paper may be excluded in this study if it does not meet the quality appraisal criteria and requirements. A paper should meet one of the quality appraisal criteria and requirements in order to be listed below: A Paper should **1** Propose a framework or methodology that solves a problem which resembles the problem of this study. **2** Review literature that answers one of this study's research questions. **3** Be written in English. Forward and backward snowballing [Wohlin, 2014] of the selected articles is used to gain insight and provide bigger context to this study. Backward snowballing is checking the references cited by an article, which means going back in time. Forward snowballing is checking if the article is cited in other articles, which means going forward in time.

2.6.4 Literature search

The results for Querying Mendeley and Google scholar resulted in 10 papers per search term per search engine. There was a small overlap in literature. Therefore the literature search resulted in 80 potential literature artefacts to be potentially included in this study. The 80 papers found in literature are checked for quality appraisal criteria and requirements as proposed in earlier. Papers that meet the criteria and requirements are included in this study. The results of the quality appraisal criteria and requirements can be seen in the table below.

Search term	literature artefacts included in this study
"data governance "	12
"data governance framework"	12
"corporate governance"	4
"data asset management"	10
"data classification "	...

Table 2: results of quality appraisal

Forward and backward snowballing resulted in an addition of 7 artefacts to be included in this study.

2.7 Validation

The study will be validated with doing a Case study with the research object Transavia. The designed framework will be optimized for the research objects in order to properly test and review it. The framework will be validated with feedback given from the

This section will discuss the validity and validity threats of this study. The validity matrix of Verhoeven is used to discuss the validity threats and how they are tackled. The internal validity can be described as the extent to which accurate conclusions can be drawn from the results. The reliability of the study ensures the operations of a study can be repeated with the same results. Construct validity is the extent to which you are measuring what you want to measure. Lastly, external validity is showing whether and how a case study's findings can be generalised. [Verhoeven, 2019]

Threat	Risk	Mitigate
Internal validity	Expert provide false information in an interview.	only include information that is confirmed in multiple sources
	False negative or false positive feedback in evaluation provided by stakeholders	Draw conclusion based on a large sample of feedback
Reliability	Literature study is not reproducible	Use a framework for the literature review and state all step taking precisely in the report
Construct validity		
External validity	Research object might be too specific to generalize findings	Do not generalize findings only based on the case study alone

Table 3: Validity table

2.8 Planning

The long proposal is set to be completed mid June this will end the problem investigation phase of the study. The framework will be designed in the next 6 weeks. The designed conceptual framework is optimized for the research object in the next 4 weeks and will be completed near the end of August. This study will end with evaluating the optimized framework tested with the research object.

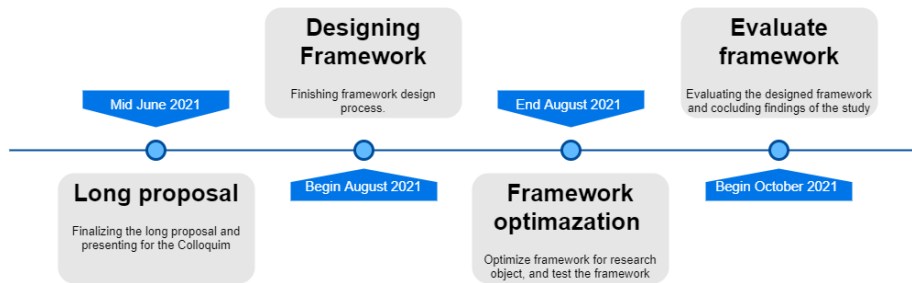


Figure 2: Planning for this study

2.9 Summary

This chapter provided the research plan of this study. Additionally this chapter investigated the problem and provided a problem statement. A lack of data governance frameworks was highlighted. The problem investigation also stated that data governance is an understudied field. Next the purpose of study was explained. The purpose of this study is to make it easier for an organization to implement data governance and provide a handhold to start data governance. The main research question is "What would be a data governance framework that is based upon data classification?". The approach to solve the research questions was given next. The study will use a literature review as input to design a new data governance framework that uses data classification. This framework will be validated with expert interviews and workshops. Next validity threats of the study were discussed and solutions were stated to tackle them. The chapter was concluded with a general planning of the remainder of the study.

3 Data classification

This chapter provides an extended literature review of data classification. The knowledge gained in the literature review will provide input into the design process of the data governance framework. Secondly, this chapter is used to answer research question 4 and 5. The chapter is structured in sections in which some sections of this chapter aim to answer one of the research questions. Data classification is a term for categorizing data based on its properties. The goal for data classification in this study is to create data categories with different data governance requirements. Additionally a classification framework is designed to classify data into these categories. Later chapters will extend the classification framework by adding suitable data governance mechanisms to the categories in order to propose a new data governance framework.

3.1 data classification properties

This section will be used to find data classification properties most suitable for differentiating and creating data categories. These data properties are used to build a data classification framework suitable for conceptual data governance framework. The result of this section will be the answers to the fourth and fifth research question of this study. The data landscape of an organization can be very diverse and not all data within the landscape is processed the same way. The difference in properties and requirements of data make data management and governance very complex. On the other hand, distinct governance for each individual data entity creates an overwhelming amount of work as well as an unstructured foundation for data governance. An in-between solution may be to categorize coherent data entities and govern these entities with the same rules, policies and procedures in a structured manner. This study aims to find suitable data properties to label data entities to a corresponding data category. More importantly the goal is to find data properties that classify the data entities in categories with the same policy requirements. A governance structure, rules, policies and guidelines can be designed for each data category when the data entities are labeled correctly.

Data quality dimension literature might provide the desired data properties suitable for differentiating data categories for data governance. “A Data Quality Dimension is a term used to describe a data quality measure that relates to multiple data elements including attribute, record, table, system or more abstract groupings such as business unit, company or product range [Expertian, 2020].” The quality dimensions measures provide an objective measure to classify data into categories. Wang and Guarascio state 20 quality dimensions for data quality. These dimensions can be applied to help analyze data quality and formulate quality data policy [Wang and Guarascio, 1991]. The result of our research and analysis of data consumers yielded the following data quality dimensions. (1) Believability (2) Value Added (3) Relevancy (4) Accuracy (5) Interpretability (6) Ease of Understanding (7) Accessibility (8) Objectivity (9) Timeliness (10) Completeness (11) Traceability (12) Reputation (13) Representational Consis-

teny (14) Cost Effectiveness (15) Ease of Operation (16) Variety of Data & Data Sources (17) Concise (18) Access Security (19) Appropriate Amount of Data (20) Flexibility.

A grouping representation of the 20 most common dimensions as proposed by wang and Guaracio can be found in figure 3. The figure shows the overarching and relations of the 20 quality dimensions proposed by the study.

Value Added (2) Cost Effectiveness (14)			
Ease of Understanding(6)	Believability (1)	Relevancy (3)	Timeliness (9)
Interpretability (5) Representational Consistency (13) Conciseness (17)	Accuracy (4) Objectivity (8) Completeness (10) Traceability (11) Reputation (12) Variety of Data & Data Sources (16) Access Security (18)	Appropriate Amount of Data (19)	Accessibility (7) Ease of Operation (15) Flexibility (20)

Figure 3: Grouped data quality dimensions as proposed by Wang and Guaracio

Wang follows up his previous work In 1996 and proposes a Framework for Data Quality. This study develops a hierarchical framework that captures the aspects of data quality that are important to data consumers. 118 data quality attributes collected from data consumers are consolidated into twenty dimensions, which in turn are grouped into four categories. Using this framework, information systems professionals will be able to better understand and meet their data consumers' data quality needs [Wang and Strong, 1996].

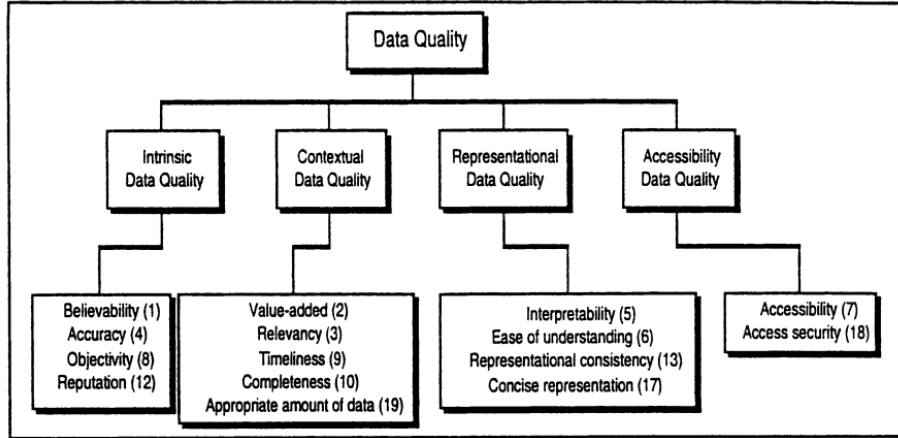


Figure 4: hierarchical Data quality dimension framework overview proposed by Wang 1996

Juddoo et al studied data governance in health industry and investigated the data quality dimensions within the context of big data. Their literature review concludes that accuracy, completeness, consistency, reliability and timeliness are the most important data quality dimensions for data governance in health care.

Shaik and Sasikumar propose a set of parameters for data classification. They indicate three data classification characteristics on which data must be classified and accordingly security considerations can be applied [Shaikh and Sasikumar, 2015]. The characteristics are Access control, content and storage and provide a set of parameters for each category. The parameters can be used for providing security levels based on the content and accessibility of data. The parameters for access control category defines the access restrictions applied on data. The parameters for access control are frequency of use, frequency of update, "visibility and accessibility" and retention. The content category poses properties of data with respect to its modification. The parameters for content are precision/accuracy, reliability/validity, degree of completeness, consistency and auditability. The storage category entails parameters for the technical storage properties of data.

3.2 Category classification

The literature mentioned in the previous section has provided many classification measures and dimensions to categorize data. The next step is to find a set of measures for creating data governance categories. The goal is to design a suitable data classification framework for data governance. Which means that data is classified into categories that share the same quality and governance requirements. Secondly, this will answer research question 5 of the study.

It is important to keep the data governance relation in mind while classifying data. An important side note is that the data classification is performed to map

and measure the as is properties of data in the organization. This process is used to determine the level of governance for the data entity. Data governance is a learning opportunity and process for an organization. The classification of a data entity might change overtime as the organization changes or matures in data governance. Therefore, the data properties are measured by how they are used at this moment by the organization.

There are two data quality measures that were consistently repeated by literature. These two measures are business value and access. The meaning of access is to have the right or opportunity to use or look at something [PMI, 2013]. PMBOK defines business value as the entire value of the business. This study will use the PMBOK definition for business value in regard to the data present within the application landscape of an organization [PMI, 2013]. Business value and access will be used as dimensions for the data classification framework. Each dimension will include multiple measures to design the framework. The classification measures for the access dimensions are accessibility, Access security, availability and ease of operations. The value dimension category will have Added value, Relevance, Ease of understanding and reliability as classification measures.

The 8 classification measures taken from literature are used for the draft design of the classification framework. The measures' validity and relevance for data governance are reviewed in the validation chapter of this study. The next sections will provide a clear definition and explanation how to properly classify data using the measure. A visual representation of the dimensions and data classification categories can be found in fig 5. The classification measures used in the framework are general properties of data and do not have an order of magnitude. Therefore, the terms high and low are used to describe and score the measurement-level of the data entity. All data measures contribute equally to their classification dimension.

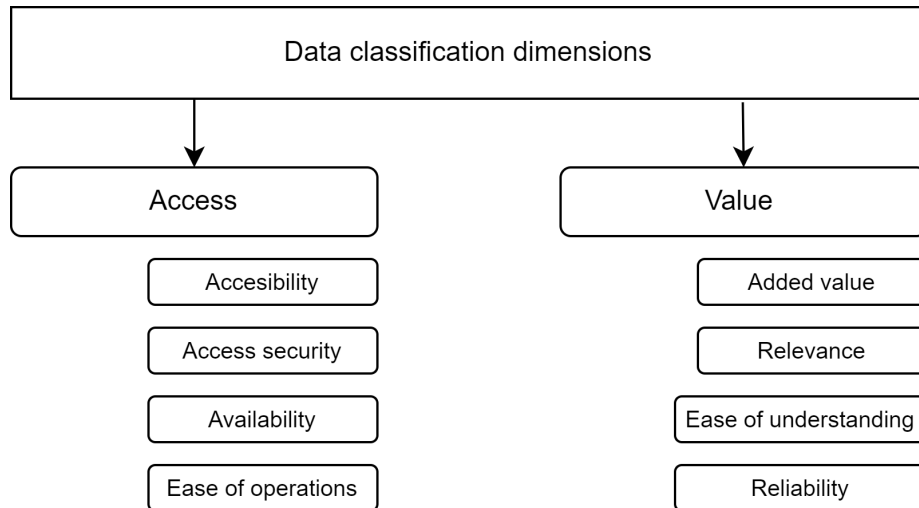


Figure 5: Dimensions and measures of the classification framework

3.2.1 Accessibility

Unfortunately the definition for accessibility is not given by the PMBOK guide. Therefore a global definition is used gathered from the Cambridge dictionary and data governance literature. The Cambridge dictionary describes accessibility as the quality or characteristic of something that makes it possible to approach, enter, or use it [dictionary,]. Panian describes availability in his paper about data governance as Ensuring that data are available to users and applications,when, where and how needed[Panian, 2010].The essence of the classification measure is to consider how many stakeholders a data entity is used. The accessibility of a data entity is high when it is used by multiple stakeholders in the organization’s operations. The organization is responsible for making the assessment whether the accessibility of a data entity is high or low. Secondly, The organization is also responsible for composing the threshold for classifying the accessibility high or low.

3.2.2 Access security

The sensitivity level of data varies from freely accessible to highly confidential. The access security measure audits the access security of a data entity. To rate this, the organization needs to look how strict the data entity is secured. when a data entity is freely stored for anyone the access security measure is obviously low. But what is the threshold for access security being high? Once again, an organization has control over the classification of their data entities. the organization should make a consideration about the definition of a data entity with high access security. Consideration of the threshold can be if a data is protected in a secured environment or protected with a password that only a

certain number of employees by the organization possess.

3.2.3 Availability

The availability measure of a data entity entails how available a data entity is. The availability of a data entity is high if it is highly available for a certain number of employees in the organization. The distinction between high and low availability is easy to make. For this measure it is suggested to classify the availability low if the data source is not freely available within the organization. On the other hand, the availability of a data source is high when it is available to use outside the system it originated from.

3.2.4 Ease of operation

The ease of operation measure was introduced by Wang and Guarascio. This quality measure is used to determine operational ease of the data entity. The organization weighs how easy it is to work with the data entity. Is the data entity easily joined, easily changed, easily updated, and how easily the data source may be used for multiple purposes?

3.2.5 Added Value

The most important measure of the Value category is the added value measure. The added value measure is the degree of value a data entity adds to the organization. What would happen if the organization does not store and process the data entity? This is the question the data governance teams must evaluate? If the operations of the organizations call mostly continue without the data entity then the added value of this entity is low. However, when the added value of the data entity is high then the organization cannot perform their daily operations without using the data entity.

3.2.6 Relevance

Another important data quality measure is relevance. Relevant data is more valuable than irrelevant data. Relevance is the degree of usefulness or appropriateness. The more relevant a data entity is the more value it possesses to the organization. Since the introduction of the GDPR Personal Identifiable Information (PII) sensitive data can only be processed if it serves a relevant purpose to the operations of an organization. In other words, PII sensitive data can only be processed by an organization if it is necessary to fulfill the operation. This means that since 2018 all PII sensitive data present in an organization is relevant.

3.2.7 Ease of understanding

Valuable data is easy to understand. The ease of understanding measure weighs the degree of understandability of the data entity. A data entity is not under-

standable if only a few stakeholders know how to interpret it. Data has value if it is easily understood by anyone.

3.2.8 Reliability

The previously mentioned quality measures relate to the external value of data. Reliability on the other hand focuses on the internal value of a data entity. Reliability is defined as the degree of trustworthiness. Data might add value and be relevant, but it cannot serve the organization unless it is reliable. A data entity is considered reliable if it performs consistently well and is trustworthy. Trustworthiness is a vague term; therefore, the organization has the task to compose a clear distinction for themselves what data entities are considered trustworthy and what not.

3.3 Classification framework

The result is a framework that can be represented as a 2x2 dimensional matrix with 4 data entity categories. The framework consists of two dimensions which were introduced in the previous section. The dimensions for the classification framework are Business Value and access. The dimensions were chosen due to their recurring citations in classification literature. The framework uses 4 classification measures per dimension in order to classify the dimension. The classification measures for the access dimension are: accessibility, access security, confidentiality, ease of operation. The classification measures for the business value dimension are: added value, relevance, ease of understanding and reliability. The framework will classify a dimension high if 2 or more of the data measures within that dimension are classified as high. In all other cases the dimension is classified as low. The resulting four categories feature distinct requirements for data governance. The data categories can be found in figure 6.

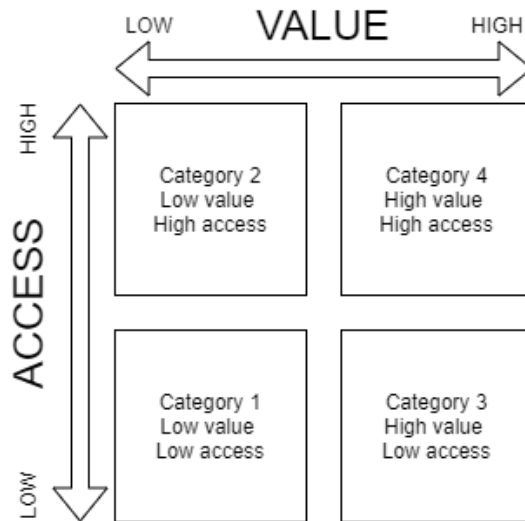


Figure 6: Categories and dimensions of the classification framework

The organization can use the classification framework to categorize their data into 4 distinct categories. The categories are explained below.

The first category entails the data entities that have both a low access and value classification. This category can be seen as support data. The data on its own is not valuable for the company and is not used by many stakeholders. This Category governs itself and without the interference of data governance elements. A potential data source in Transavia that can be classified in category 1 may be an excel sheet that is used in a single department. The excel sheet may be derived from other data and used for a single calculation. The entity value is low and is only used in a single department. Therefore it is classified in this category.

The Second category includes data entities that have a high access but a low value classification. These data entities are used by multiple stakeholders but not essential to the business of an organization. The data entity does not contain enough value to be essential for an organizational process. A potential data source in Transavia that can be classified in category 1 may be unchanging data such as ancillary prices. Ancillaries are an extra option for a booking for example an extra piece of luggage. The extra luggage ancillary is sold for a fixed price for a long period of time. The data is used by many stakeholders but the added value is low. The data is also not valuable for revenue optimization since they use fixed prices.

The Third category involves data entities that have a low access and a high value classification. The data entities in this category entail value to the business of the organization. Yet the data source is only used by one or few stakeholders. The data entity is relevant and valuable but only for a distinct part of the

organization. data about the length of the runway is an entity that is suitable to be classified in this category. Runway length data is used by pilots to optimize the used fuel by the airplane during takeoff. The longer the runway the slower the airplane can accelerate to reduce fuel consumption. This data source is only used in this use case and not in other parts of the organization. Fuel optimization is valuable and can save the organization lots of fuel expenses. Therefore runway length data is suitable in this category.

The Fourth data category has both high access and value classification. This data category features all valuable data that is used by multiple stakeholders in the organization. The data classified in this category can be considered the core data of the organization. The organization would not function the same without the use of data entities classified in this category. The most obvious data entity that can be classified in this category is booking data. Booking Data is used in multiple departments of transavia and regards personal customer information and the flight they booked. The data is very valuable for the organization and has a high quality requirement. The booking data entity is classified in category four due to its value and access requirement.

This Section reviewed literature about data classification. The literature provided this study with several data quality and classification measures. The measures were used to design a classification framework suitable for data governance. The Next section will introduce data governance and its context.

3.4 Summary

This chapter provided answers to research question 4 and 5 of this study. These research questions were answered by providing an extended literature review of data classification. Additionally a design for a classification framework for data was designed and proposed. The framework introduces 2 dimension Access and Value. These dimensions were derived from data classification Literature. Each dimension featured 4 data classification measures to support the classification of the dimension The classification measures for the access dimension are: accessibility, access security, confidentiality, ease of operation. The classification measures for the business value dimension are: added value, relevance, ease of understanding and reliability. The merged 2 dimensions create 4 distinct data classification categories with different data governance requirements. The next chapter will introduce data governance and state its core elements.

4 Data governance

This chapter provides an extended literature review of data governance. The literature review is intended to introduce the reader to data governance which will simplify the explanation and steps taken in the design process of the framework. This chapter is used to answer research question 1 and 2 of this study. The sections will review data governance literature and frameworks to find the core elements of data governance. Academic research on data governance is still in its infancy [Weber et al., 2009]. Therefore the goal for this chapter is to find the most relevant and comprehensive data governance framework. The framework with the most overlap to the core elements is considered to be the most relevant of a comprehensive framework for data governance. The most relevant framework for data governance is used to be combined with the classification framework proposed in the last chapter. The classification framework and the most relevant framework for data governance are merged into a draft framework for data governance.

4.1 What is data governance?

The following section will answer research question 1 of the study. “What is data governance and what is the benefit of data governance?” Data governance literature will provide the answer to this research question.

Data governance is “The exercise of authority, control, and shared decision making (planning, monitoring and enforcement) over the management of data assets.” Data governance includes High-level planning and control over data management and should be complied by every-one within an organization.[DAMA, 2009] Stedman and Vaughan stated potential Benefits and drivers of data governance [Stedman and Vaughan, 2020]. These are: increased data quality, lower data management costs, data process standardization, increased data security, increased access to needed data for analysts and business users and improved business decision-making.

Data governance can provide a handhold for implementing data rules and guidelines within an organization. These rules and guidelines can help standardize the data processing in the organization and reduce risk. Secondly, data governance can help an organization increase and monitor their data quality. Lastly, data governance may enable new revenue opportunities for an organization by enabling better decision making. These benefits are not easily visible unless they are monitored and measured over time.

Implementing data governance can be hard. Firstly, since there is no one, size fits all solutions for data governance. Data governance affects all employees of an organization, and it often involves organizational and sometimes cultural change. implementing organizational and sometimes cultural change is a challenge by itself. Moreover, can implementing data governance be difficult because the number of data sources and stakeholders may be overwhelming to govern within larger organizations. Lastly, since data governance compliance is hard to monitor. How do you know the employees that perform the data processes

comply with the newly designed data governance program?

4.2 Data governance vs IT governance

Data governance should not be confused with IT governance. [Nfuka and Rusu, 2010] state IT governance as “integral part of enterprise governance and have potential to provide mechanisms for leadership and organizational structures and processes that ensure the organization’s IT sustains and extends the organization’s strategies and objectives”. The Difference between IT governance and data governance is that IT governance focuses on governing IT equipment whereas data governance focuses on governing organizational data within its technological environment. The two concepts are often confused with each other. Firstly, because the names resemble each other and Secondly, because IT and data governance both support alignment between IT and business. IT governance is a much more mature field of study compared to data governance [Haes and Grembergen, 2008][Neff et al., 2013]. There are many IT governance frameworks which have been tested in practice compared to data governance frameworks. Compared with the extensive body of IT governance research, academic research on data governance is still in its infancy. [Weber et al., 2009]

4.3 Data governance frameworks

This section will summarize the existing data governance frameworks found in the literature search. The data governance frameworks are reviewed and compared with each other in order to answer research question 2: “What are the current solutions and frameworks for data governance?” The frameworks are compared to find differences and similarities between the frameworks. The goal for this section is to find the core elements of data governance. A list of core elements will be created after the literature review. This List of core data governance elements is fed back to the framework. The framework that matches the most with the core elements will be considered the most relevant data governance framework. The most relevant framework will be combined with the classification framework to design a new approach for data governance.

4.3.1 Abraham and schneider

Abraham and Schneider’s propose a conceptual framework for data governance. Their framework is very extensive and consists of 4 dimensions, antecedents and consequence. The dimensions are governance mechanisms, organizational scope, data scope and domain scope. Antecedents describe the external and internal factors that precede or predict the adoption of data governance practices. Consequences refer to the outcomes of data governance. [Tallon et al., 2013].

Governance mechanisms comprise formal structures connecting business, IT, and data management functions, formal processes and procedures for decision-making and monitoring, and practices supporting the active participation of and collaboration among stakeholders [Abraham et al., 2019].

The framework proposes 3 distinct governance mechanism categories:

- Structural mechanisms
- Procedural mechanisms
- Relational mechanisms

Structural governance mechanisms determine reporting structures, governance bodies, and accountabilities [Borgman et al., 2016] Roles and responsibilities are assigned within the organization to enable accountability. A list of roles and responsibilities and the allocation of decision-making authority as proposed by Abraham and Schneider can be found in table 4.

Role or decision-making authority	Responsibility
Executive sponsor	The executive sponsor is ideally a high-level executive i.e., C level and is responsible for providing strategic direction, business prioritization and funding for data management initiatives.
Data governance leader	The data governance leaders' responsibility is to manage the data governance program and provides guidance concerning design, delivery, maintenance of data and oversees compliance with data policies. The data governance leader also coordinates tasks for data stewards and provides reports on data governance performance.
Data owner	The data owner are business executives accountable for data assets within their business unit.
Data steward	Data stewards are subject matter experts or business leaders that have detailed knowledge about business and data requirements. Data stewards should be able to translate these requirements into technical specifications
Data governance council	The data governance council is a cross function governance body which is responsible for the strategic direction for the entire data governance program and aligns it with organizational goals
Data governance office	The data governance office is a staff organization which supports the governance and decision-making activities of data stewardship and the data governance council. The data governance office oversees the communication channels meeting scheduling and coordination of issue resolution and stakeholder education.
Data producer	The data producer is the creator, aggregator or maintainer of a data asset within the organization
Data consumer	The data consumer is the user of the data. The data consumer specifies requirements and reports data-related issues

Table 4: Role or decision-making authority as proposed by Abraham and Schneider.

Procedural mechanisms aim to ensure confidentiality integrity and Availability of data within an organization. And can be divided into 9 categories. These categories are:

- 1. Data strategy
- 2. Policies
- 3. Standards
- 4. Processes
- 5. Procedures
- 6. Contractual agreement
- 7. Performance measurement
- 8. Compliance monitoring

- 9. Issue management

A data strategy can be described as a high-level course of action driven by strategic business objectives. The data strategy consists of a vision statement, business case, guiding principles, long and short-term objectives and an implementation roadmap. Data policies provide rules and guidelines for creation, acquisition, storage, security, quality, and permissible use of data [DAMA, 2009]. Data policies are fundamental for ensuring a well-organized data governance program. Data standards ensure consistency and normalization of data within an organization. Data standards are defined internally by data stewards and data architects. Another fundamental element for ensuring a well-organized data governance program is the data processes. Data processes are standardized, documented and repeatable methods used to govern data [Al-Ruithe et al., 2019]. Actions may include:

- developing and maintaining rules for data handling.
- modeling and documenting the data lifecycle.
- Assessing the current state of the data governance program.
- Comprise processes for:
 - the alignment and validation of policies
 - decision-making
 - performance measurement
 - issue resolution

Procedures are “the documented methods, techniques, and steps followed to accomplish a specific activity or task” [DAMA, 2009]. Contractual agreements are needed for Data provisioning and data sharing between internal departments or external organizations. Performance measurements assess the effectiveness of data governance within the organization by measuring the level of goal attainment. Compliance monitoring aims to measure if the policies, standards and procedures are complied. As stated earlier this is hard to measure for an organization. A solution often used is auditing which results in unbiased assessments and recommendations for improvement of the program. Corrective and preventive actions can be taken according to the audits results. Lastly, issue management refers to the identification, management, and resolution of data-related issues [DAMA, 2009].

Relational mechanisms enable and give structure to collaboration between stakeholders and can be divided into 3 elements. The communication, training and coordination of decision-making. The communication is focused on continuously generating awareness for the data governance program among stakeholders. Generating and reemphasizing the importance of the data governance program will improve the compliance of the program and establish shared commitment with the stakeholders. Training programs ensure stakeholders have knowledge and qualifications to support the implementation of data governance [Tallon et al., 2013]. The coordination of decision making are practices and

methods for aligning business functions. There are 2 approaches to be considered horizontal and vertical. The vertical approach focuses can be classified as more hierarchical where the decision-making authority is on top of a pyramid like structure. The horizontal approach (cooperative) focuses more on collaboration between teams such as working groups and committees to collaboratively solve problems within the organization.

The organizational scope dimension of Abraham and Schneiders framework represents the expense and unit of analysis of the program. Abraham and Schneiders split the organizational scope into two. Firstly, the intra-organizational scope which can be described as the data governance within a single organization. Secondly, the inter-organizational scope is described as the data governance between organizations or within a firm ecosystem.

The data scope dimension of Abraham and Schneiders framework defines the data that will be subject to the data governance program. Traditional data builds the basis for an organization's operations [Lee et al., 2014]. It comprises master data, transactional data, and reference data. Master data describes the key business objects within an organization [Loshin, 2008]. Data governance in the context of big data, has partially different requirements on data governance than traditional data. A review for data governance for big data does not fall within the scope of this study and will be left out of the literature review. Data governance for traditional data often aims to ensure the consistent use of traditional data across the organization [Dreibelbis, 2008]. Data policies and processes for monitoring conformance to those policies are often used to achieve this [Loshin, 2008]

The domain scope dimension of Abraham and Schneiders framework defines the data decision domains within the organization. The focus of a data governance program may differ between organizations. The domain scope divides the domain scope into 6 categories and enables organizations to focus on specific categories within the data domain scope. The categories and their definition can be found in table 5

data quality	The ability of the data to satisfy its usage requirements in its context. Data governance programs with a focus on improving their data quality should develop a data quality strategy. Data quality should be monitored into spot issues prematurely and ensure healthy data within an organization
data security	The preservation of security requirements concerning the accessibility, authenticity, availability, confidentiality, integrity, privacy, and reliability of data (Carretero et al., 2017,) An organization can improve their data security by executing a risk assessment, set up data security roles and define policies standards and procedures for data security.
data architecture	The data architecture domain scope focusses on defining the enterprise data objects and developing an enterprise data model on a conceptual, logical, and physical level. Data governance programs which focus on data architecture a
data lifecycle	The data lifecycle is the approach of defining, collecting, creating, using, maintaining, archiving, and deleting data [Khatri and Brown, 2010]. Analysis of information flow, identification of business processes that use data and identifying potential overlaps in data storage are used measures for organizations that want to improve their data lifecycle.
meta data	Meta data is for classifying the sensitivity level, provenance and retention periods of the data. Defining and introducing common meta data standards and the specifying processes to build a meta data repository can help organization be more aware of their meta data and
data storage and infrastructure	Data storage and infrastructure focusses on where and how data is stored within an organization. Data governance programs that focus on data storage and infrastructure starts with assessing the current application and storage landscape. Improvements can be made by planning of software applications and storage capacity to support data quality, data security, and data lifecycle.

Table 5: domain scope principles Role or decision-making authority as proposed by Abraham and Schneider.

Antecedents describe the external and internal factors and drivers of an organization to implement data governance or describe the adoption of the data governance program. Antecedents impact the implementation and level of adoption. Antecedents can be internal or external. Internal antecedents may arise from strategic, organizational, system-related, and cultural factors. External antecedents may arise from legal and regulatory requirements imposed from an external party.

Consequences are the outcome of the data governance program. Abraham and Schneider propose 2 types of data governance consequences: intermediate performance effects and risk management. Intermediate performance effects are a broad term. Positive intermediate performance effects of a data governance program might be increased number of sales and customer spending, improved data quality and increased dynamic and operational capabilities for the organization. The second consequence is management of data-related risk. A data governance program reduces data management risks by creating risk-mitigating policies and introducing controls for monitoring compliance [Khatri and Brown, 2010]

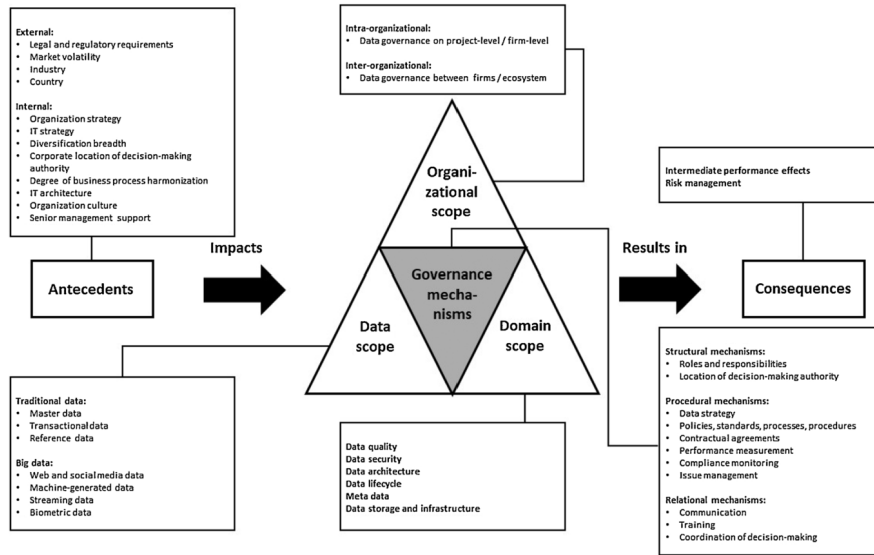


Figure 7: The data governance framework proposed by Abraham and Schneider [Abraham et al., 2019]

The framework for data governance proposed by Abraham excels in explaining the governance mechanisms. The governance mechanisms are categorized into structural, procedural and relational mechanisms. The mechanism categories describe the desired and compliant responsibilities and interactions for the data governance program. The mechanisms are the heart of a data governance program and if not designed correctly the program will die due to insufficient compliance.

4.3.2 Khatri and Brown

Khatri and Brown propose a data governance framework for developing a data governance strategy and an approach for managing data as an organizational asset. The framework consists of 5 decision domains: data principles, data quality, meta data, data access and data lifecycle. These 5 decision domains are interconnected, and a visual representation can be seen in figure 8

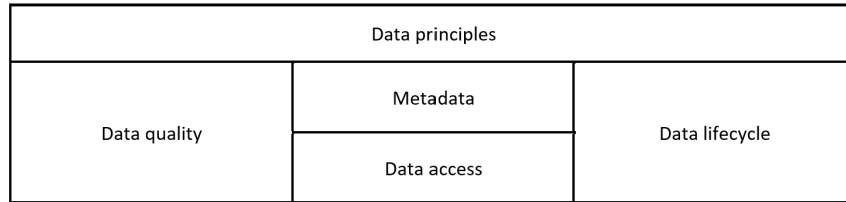


Figure 8: Decision domains as proposed by Khatri and brown

The data principal domain emphasize that data is an organizations asset. It defines the desirable behaviors for all intended users of the data by introducing specific policies standards and guidelines for managing it. The data principal domain establishes the direction for all other decision domains since this principle sets the boundary requirements for the intended uses of data [Khatri and Brown, 2010]. These boundary requirements set standards for data quality which in turn influence the meta data and data access. The data quality domain assesses the quality of data processed by an organization. The meta data assesses the type of data processed and stored by an organization. Lastly the data access domain asses the security and risk involved with the confidentiality integrity and availability of data. The data lifecycle plays a key role in initiating the conversion of data principles into IT infrastructure. The 5-decision domain proposed by Khatri and Brown are like the domain scope proposed by Abraham and Schneider. However, the framework proposed by Khatri and Brown is centered around developing a data governance strategy whereas Abraham and Schneider’s approach centers around the implementing a data governance program. Potential roles and locus of accountability are designated to the employees after the purpose and goal of the decision domain is clear. figure shows the decision domains, their goal, decision domain questions and the potential roles and locus of accountability.

Data Governance Domains	Domain Decisions	Potential Roles or Locus of Accountability
Data Principles • Clarifying the role of data as an asset	<ul style="list-style-type: none"> • What are the uses of data for the business? • What are the mechanisms for communicating business uses of data on an ongoing basis? • What are the desirable behaviors for employing data as assets? • How are opportunities for sharing and reuse of data identified? • How does the regulatory environment influence the business uses of data? 	<ul style="list-style-type: none"> • Data owner/trustee • Data custodian • Data steward • Data producer/supplier • Data consumer • Enterprise Data Committee/ Council
Data Quality • Establishing the requirements of intended use of data	<ul style="list-style-type: none"> • What are the standards for data quality with respect to accuracy, timeliness, completeness and credibility? • What is the program for establishing and communicating data quality? • How will data quality as well as the associated program be evaluated? 	<ul style="list-style-type: none"> • Data owner • Subject matter expert • Data quality manager • Data quality analyst
Metadata • Establishing the semantics or "content" of data so that it is interpretable by the users	<ul style="list-style-type: none"> • What is the program for documenting the semantics of data? • How will data be consistently defined and modeled so that it is interpretable? • What is the plan to keep different types of metadata up-to-date? 	<ul style="list-style-type: none"> • Enterprise data architect • Enterprise data modeler • Data modeling engineer • Data architect • Enterprise Architecture Committee
Data Access • Specifying access requirements of data	<ul style="list-style-type: none"> • What is the business value of data? • How will risk assessment be conducted on an ongoing basis? • How will assessment results be integrated with the overall compliance monitoring efforts? • What are data access standards and procedures? • What is the program for periodic monitoring and audit for compliance? • How is security awareness and education disseminated? • What is the program for backup and recovery? 	<ul style="list-style-type: none"> • Data owner • Data beneficiary • Chief information security officer • Data security officer • Technical security analyst • Enterprise Architecture Development Committee
Data Lifecycle • Determining the definition, production, retention and retirement of data	<ul style="list-style-type: none"> • How is data inventoried? • What is the program for data definition, production, retention, and retirement for different types of data? • How do the compliance issues related to legislation affect data retention and archiving? 	<ul style="list-style-type: none"> • Enterprise data architect • Information chain manager

Figure 9: Decision domains as proposed by Khatri and brown

4.3.3 Ladley

John Ladley focusses on the design, deployment and maintenance of a well-structured data governance program with his book. Ladley proposes an iterative methodology consisting of 8 processes. A visual representation of the processes can be seen in figure 10. A data governance program starts with the scope and initiation. The ability of an organization to govern and to be governed is assessed after the scope is understood. The assessments goal is to find out the capacity for the organization to change. Some organizations have the spirit to change but do not have the capability and capacity to do it. The pre-program assessment is fundamental for the acceptance and implementation of the pro-

gram. After the assessment phase comes the vision phase. The vision phase aims to move stakeholders and high-level management toward a greater level of understanding about the data governance program. The next step is align and business value. The data governance program starts to be specified in the functional design phase. The outcome of this process is designing the principles processes and policies of the data governance program. The next step is to turn the principles processes and policies into an organizational framework and identify the stewardship/ownership/custodian for the data. With the next step is to make a roadmap for the data governance program since the organizational framework is completed. This step entails the details about implementing the data governance program. Next the data governance program is enrolled and sustained according to the proposed roadmap. Enrolling and sustaining the data governance program is the last step of the cycle proposed by [Ladley, 2012].

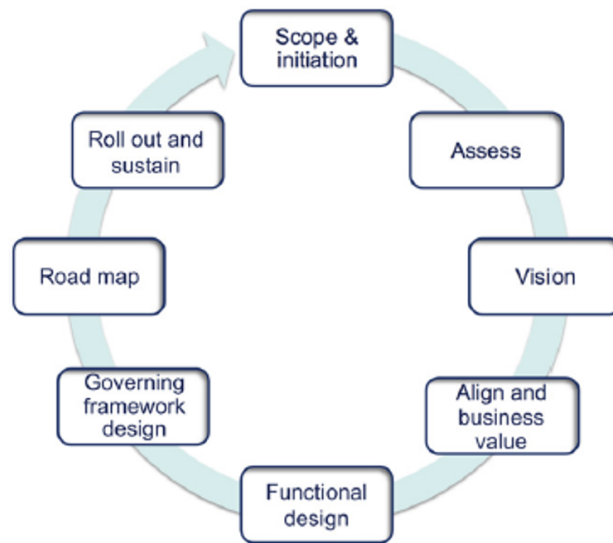


Figure 10: processes for designing and deploying a data governance program. [Ladley, 2012]

4.3.4 Panian

Panian states his personal experience for data governance. The study starts with stating the goals and drivers for data governance. The goals of data governance are to Ensure data meets the needs of the business, Protect and manage data as a valued enterprise asset and Lower the costs of managing data. the most common drivers for data governance are: firstly, Growing revenue by increasing cross-sell/up-sell rates and improving retention among existing customers with a deep understanding of their existing customers. Secondly Lowering costs by increasing operational efficiency with business process automation and elimi-

nating redundancy. Lastly, Ensuring compliance with external regulations and internal governance policies by streamlining the collection of reporting data and increasing auditability [Panian, 2010]. Panian states 6 key data attributes and four components for data governance. "By examining the challenges that arise repeatedly across organizations, enterprise data must have the following six attributes:

- Accessibility: Ensuring that all enterprise data can be accessed, regardless of their source or structure.
- Availability: Ensuring that data are available to users and applications, when, where and how needed.
- Quality: Ensuring the completeness, accuracy and integrity of data.
- Consistency: Ensuring the meaning of data is consistent and reconciled across all systems, processes, and organizational units.
- Auditability: Ensuring there are controls and an audit trail on the data.
- Security: Ensuring secure access to the data.

Data governance manages and develops these key data attributes, enhancing the overall value of the data as an asset to the organization." [Panian, 2010] An effective data governance framework involves four key components. Panian's key components for data governance are Standards, policies and processes, Organization and Technology. "A key function of data governance is to establish the standards for data in an enterprise" [fis, 2012]. These standards are data definitions, taxonomies, master data definitions and enterprise data models. Policies and processes lay the foundation for an effective data governance practice. They ensure right-full creation, development, control, management and audit of data. Companies need to define data and data-related business rules, control access to and delivery of data, establish ongoing monitoring and measurement mechanisms, and manage changes to data. Arguably the most important issue that companies must address when launching a data governance initiative is how to design the organizational structure.[Panian, 2010]. Organizational structure refers to defining the roles and responsibilities within the organization that are accountable for the data. Several roles at different level of the organization must be drawn up. These roles are designed for both business and IT personnel from executive councils to day-to-day implements. Panian's final component for data governance success is technology. many organizations start and launch their data governance program using manual tools. However, most quickly realize that this kind of manual approach is severely limited. It is difficult to ensure high data quality and availability, security is at risk given the ad hoc nature of the approach, and maintaining detailed documentation is an almost insurmountable task. [Panian, 2010] Technology can help automate and scale the development and enforcement of data governance standards, policies, and processes. A data integration platform will help to automate data-related processes. A visual representation of the key components and data attributes as proposed by Panian can be seen in figure 11.

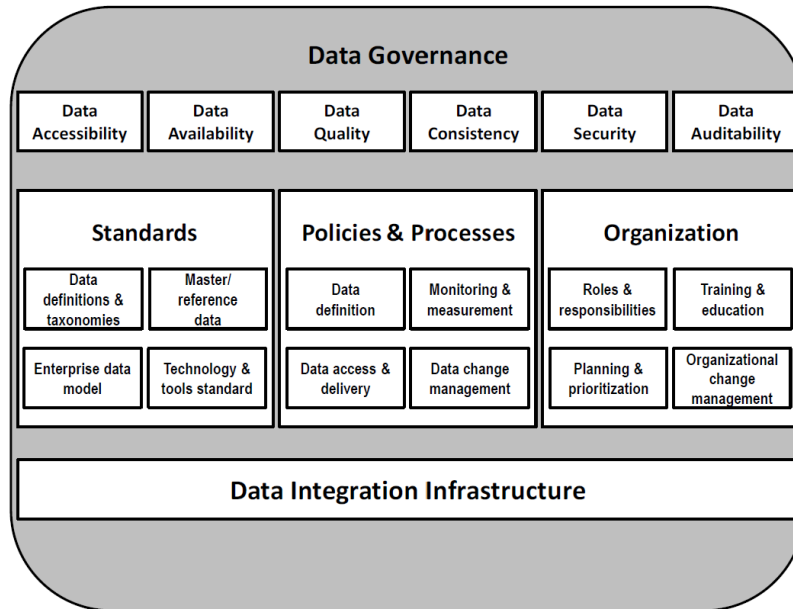


Figure 11: Panian’s key components and data attributes for data governance

Weber et al focus with their Paper "one size does not fit all A contingency approach for data governance" on the contingency factors of data governance. The paper also reviews the Components of a Data Governance Model in the context of data quality management (DQM). The data governance model consists of DQM decision areas and main activities, roles, and responsibilities, namely, the assignment of roles to decision areas and main activities. A company designs its individual data governance configuration by defining data quality roles, decision areas, and responsibilities, and by subsequently arranging the component into the model.[Weber et al., 2009]. Weber et al use the RACI chart from the Responsibility Assignment Matrix (RAM) introduced by.

4.4 core elements of data governance

Section 4.3 summarized the existing data governance frameworks and described their elements. This section will discuss the similarities and differences between frameworks. The goal for this section is to find the answer for the research question 3 of this study: "what are the core elements of data governance?". Core elements of data governance gathered by other literature reviews about data governance will be reviewed as well.

A list of common elements of data governance is concluded from the existing frameworks for data governance. The most used elements in these frameworks are:

- data standards
- data strategy
- contractual agreements
- compliance monitoring
- procedures, policies, rule and guidelines
- Organizational backbone for data governance
- decision making authority
- roles and responsibilities
- decision domains
- data definition

A common term used in multiple frameworks is defining data standards for the organization. Data standards are the guidelines by which data are described and recorded. they make it easier to create, share, and integrate data by ensuring that the data are represented and interpreted correctly.[USGS, 2021].

data strategy was mentioned a couple of frameworks. Gartner describe a data strategy as a highly dynamic process employed to support the acquisition, organization, analysis, and delivery of data in support of business objectives [Gartner, 2021]. Organizations that desire to implement data governance may have a data strategy in place or none. Data governance enables a great opportunity to align an existing data strategy to protect data for an organization or set up one when none exist.

Contractual agreements can be used to that establish a legally binding agreement between two data processing parties. The contract's terms and conditions will require the parties to either do or refrain from doing specific actions [Council,]. the most common agreement in data governance is the data sharing agreement. Data sharing agreements are formal contracts that detail what data are being shared and the appropriate use for the data [Usgs,]. Data sharing agreements provide insight and structure for who has access to what data source for an organization. data sharing agreements also sate the obligations rules and guidelines for the data provider and consumer.

If correctly implemented a structured way of working will be enabled data governance. Procedures, policies, rules and guidelines help an organization provide structure in to standardizing the governance over data. The range for this element in data governance is broad. Procedures, policies, rules and guidelines can be organization wide or entailed to a specific business unit or team. Secondly, they can be very detailed or very general. an example of an organizational wide general rule can be: Do not share any information queried from a source system via email. Procedures, policies, rules and guidelines may also be very specific and detailed to ensure a correct execution. A more specific procedure

can be the formal procedure for requesting the use and insight of a data source. the organization may design a policy with a standardized procedure for these requests by teams and business units.

Data governance will only work when the whole organization endeavors it to be successful. An organization backbone for data governance is crucial for its the success, support and compliance. an organizational backbone serves as a coordinating body that brings together a diversity of stakeholders and leads a synchronized effort to achieve the common goal (data governance) [Altarum,]. The backbone structure ensures the success of the data governance program by providing support for and measuring the compliance of data owners, stewards and other stakeholders.

4.5 closing and continuation

The core elements for data governance are clear. The governance mechanisms proposed by Abraham and Schneider correspond most closely to the core elements of data governance. Therefore the data governance elements of proposed by the framework if Abraham and Schneider are for the design of the conceptual categorized data governance framework of this study. Moreover Abraham and Schneider’s framework quoted many other framework in their design process. This is mainly due to the framework being proposed recently in 2019. The framework builds on top an extends of previous published data governance literature.”The conceptual framework builds on the rich data we have collected during our literature search process” [Abraham et al., 2019]. For example the domain scope dimension proposed by Abraham and Schneider are mainly based on Khatri and Browns quality dimensions. The modernity and relevance of the framework makes it the best choice to build upon.

Previously, the complexity and differentiate of data was emphasized. Recent frameworks including the framework proposed by Abraham and Schneider assume that the governance for every data entity is the same. The conceptual framework proposed in this study will provide a different approach for data governance which extends current data governance literature. The addition of data classification in data governance will enrich the data governance landscape. The proposed framework will enable to govern data with governance elements suitable for this category in which they are classified. Which differs from governing every data entity the same. The next chapter will propose the draft framework for data governance. The classification categories are thoroughly explained and extended with suitable data governance elements.

4.6 Summary

This chapter reviewed existing literature about data governance in order to answer research question 1 and 2. A description of what data governance entails was given. Secondly, the benefits and challenges were highlighted and discussed. The infantile state of data governance frameworks was highlighted. Lastly, the existing frameworks and solutions were evaluated to find the core

elements for data governance. The most relevant framework was chosen in order to be used in the design for the draft framework. The framework that shares the most similarities with the core elements of data governance was considered most relevant. The framework proposed by Abraham and Schneider was chosen since it shares the most similarities and it extends other frameworks reviewed in this chapter.

The next chapter will propose the draft framework for data governance. The draft framework for data governance will combine the proposed classification framework and the core data governance elements found in literature. The core elements of data governance are assigned to the data categories as a requirement to properly govern data within that category.

5 Data governance framework

The Previous two chapters highlighted literature to find the core elements of data governance, classification measures and classification dimensions for data governance. This Chapter will design a draft data governance framework. This framework is the artefact for the design cycle approach mentioned in chapter 2. This chapter will design the artefact. The next chapter will validate the designed artefact for its ability to solve the problems it claims to solve.

The previously found most relevant framework for data governance and the classification measures are combined into a draft framework for data governance. The draft framework will answer research question 6 “What is a suitable draft framework for data governance?”.

The chapter will start with a summarizing the governance mechanisms of the most suitable data governance framework. The Governance mechanisms are mapped and merged to the categories of the classification framework. Each category features its own requirements for data governance. The end result is a draft framework for data governance. The framework will be able to categorize data into categories and provides a list of governance mechanisms for each category. An organization will be able to classify its data sources and get the most suitable governance mechanisms in return for each data source. The functionalities and elements are defined as well as explained. Likewise, the connection process of governance mechanisms and the classification framework is explained throughout the chapter. The chapter is completed when all aspects of the draft framework are proposed and explained.

5.1 Overarching elements and structure

Data exists because it is used or stored by an organization, this agreement always holds true. There are overarching elements for data governance that play a role with data in a general sense or ensure the implementation and success of the data governance program. These overarching elements are suitable for all 4 data categories of the classification framework. This section will introduce the overarching elements of Abraham and Schneiders framework and explain the relation with the classification framework. Other roles and responsibilities which focus on the policy and governance of an individual data entity will be explained later in this chapter. These Roles do not apply to the full data landscape of an organization. Therefore, they are explained in the Category section of the classification framework in which they are relevant. Overarching elements involve the organizational management of the data governance program. Abraham and Schneider propose roles and responsibilities for the data governance management and steering of the program.

The first Overarching role of data governance is the executive sponsor. The executive sponsor provides strategic direction, business prioritization, and funding for data management initiatives [Weber et al., 2009]. Business prioritization and funding are essential for the success of data governance.

The second overarching role is the data governance leader. The data governance leader is responsible for the day-to-day management of the data governance program.[Loshin, 2008]

The data governance council is a hierarchy-overarching, cross-functional governance body [Weber et al., 2009] It establishes the strategic direction for the entire data governance program and aligns it with organizational goals [Cheong and Chang, 2007]. Moreover the governance council guides the monitoring of the data governance program and introduction of improvements to the program.

A data governance office establishes communication channels and oversees issue resolution within the data governance program. In addition, the office is involved in educating stakeholders to ensure and strengthen the adoption and compliance of data governance.

Three other important overarching governance mechanisms are performance measures, compliance monitoring and issue management. Performance measurement aims at assessing the effectiveness of data governance by measuring the level of goal attainment [Al-Ruithe et al., 2019]. they entail business goals such as revenue growth, increased profitability, and cost savings [Tallon et al., 2013]. Compliance monitoring aims at tracking and enforcing conformance with regulatory requirements and organizational policies, standards, procedures, and SLAs [Al-Ruithe et al., 2019]. Compliance monitoring encompasses auditing, which aims at providing stakeholders with objective, unbiased assessments and recommendations for improvement [DAMA, 2009]. Issue management refers to the identification, management, and resolution of data-related issues [DAMA, 2009]. Issues need to be resolved regardless of the category the data entity is in. the organization should design procedures for issue management. Therefore issue management is added to the overarching structures of the data governance framework.

The overarching roles, performance measures and compliance monitoring are not centered around the governance of an individual data source. yet, they are rather concerned with the success of the data governance program. Moreover, These measures are essential for the success of data governance. A data governance program is destined to fail without overarching roles, measures and monitoring. There an organization should always implement these mechanisms to ensure the overall success of the data governance program.

5.2 Decision domains

As stated in the literature review, Abraham and Schneider introduce a set of domain scopes for the data governance program. These domain scopes are derived from the data decision domains proposed by Khatri and Brown in 2010. The Decision domains or so-called domain scopes entail focus areas for the data governance program. The decision domains proposed are data quality, data security, data architecture, data lifecycle, meta data and data storage and infrastructure. The decision domains are explained in this section and are reintroduced with the categories of the data classification framework. The data

classification categories share different focus areas that align with the decision domains proposed by Khatri and brown.

5.3 classification of data

As introduced in chapter 3 this study proposed a data classification framework. The classification framework is used to label data entities into 4 data categories. The framework uses 2 dimensions to label the data entity.

The Access dimension evaluates the access requirement of a data entity. The access dimension consists of 4 data measures which support the classification of the data entity.

The First data measure for this data dimension is accessibility. Availability entails ensuring that data are available to users and applications, when, where and how needed [Panian, 2010]. The accessibility of a data entity is high if it is used by multiple stakeholders within the organization. Highly accessible require governance mechanisms that focus on the data interaction.

Access security is the second measure of the access dimension. The access security dimension evaluates the security requirement of the data entity. Data entities with a high requirement for security must be governed strictly. interaction and security mechanisms can help govern data classified with a high access security measure.

The third data measure for this dimension is confidentiality. The confidentiality measure of a data entity resembles the access security measure previously mentioned. There is a difference between them. The confidentiality measure considers the privacy requirement of a data entity whereas the access security measure considers its safety and storage requirement. The confidentiality level of data describes the ease of retrieval. A data entity with low confidentiality can be easily accessed by many stakeholders whereas a highly confidential data entity can only be viewed by a few stakeholders. Many governance mechanisms can ensure the preservation of the confidentiality of the data entity.

The Last data measure for the data access dimension is ease of operation. The ease of operation measure considers the workability of the data entity. The higher the ease of operation the more stakeholders would like to benefit from using it.

Next to the access dimension the value dimension was introduced as well. The value dimension considers the value of an individual data entity. This dimension uses just as the access dimension 4 data measure to support the classification.

The first measure of the value dimension is the added value of the data entity. The added value measure considers how valuable the data entity is for the organization. The organization has to consider if they can perform their daily operation without relying on the data entity for input. The more value a data entity provides to the company the more it needs to be protected to ensure the value is retained. Valuable data require governance mechanisms to ensure and retain value.

The second data measure is relevance. This measure considers the usefulness of the data entity. A highly relevant data source contains value to the organization and has the requirement to be governed. Governance mechanisms that ensure quality en relevance can help govern these data entities.

The third data measure is ease of understanding. This data measure considers the interpretability of the data entity. A data entity is highly interpretable if it is easy to work and transparent.

The Final data measure for this dimension is reliability. Reliability considers the quality aspects of the data entity. The data quality majorly influences the reliability of the data entity. Governance mechanisms can help maintain and improve the data quality of the data entity to preserve its reliability.

The result of the classification is a measure for both dimensions. The values of the classification can either be high or low. Together this creates four data categories in which the data entities of an organization can be classified. The visual representation of the data categories can be found in figure 6 of chapter 3. The next 4 sections will describe a single category of the data classification framework with the addition of governance mechanisms. The governance mechanisms proposed by Abraham and Schneider are connected and included to one or multiple suitable data categories. The goal is to propose a set of governance mechanisms to ensure the governance requirements for each data category. This means that all governance mechanisms are evaluated for every category. When a governance mechanism tackles a requirement of that category than it is added. The result will be a set of governance mechanisms suitable for each category. The completed set of categories enriched with the governance mechanisms is the artefact of the design cycle in this study.

5.4 Classification of mechanisms

This section will evaluate and classify the governance mechanisms proposed by Abraham and Schneider. This classification will assign the governance mechanisms to categories of the classification framework. The classification will evaluate if a governance mechanism supports the governance requirements of the data entity. The effort to implement a governance mechanism is balanced with the properties of the category. In addition, the responsibility and accountability requirement of a category is taken into account before a mechanism is assigned.

5.4.1 Structural mechanisms

First the Roles and responsibilities as proposed by Abraham and Schneider are assigned to the categories. Earlier overarching roles were discussed. This section will classify the roles that impact a single data entity and not the programs as whole. There are 4 Roles that Abraham and Schneider propose for individual data entities. These roles are: data ownership, data stewardship, data producers, and data consumers. A detailed explanation of the roles can be found in Table 4 located in section 4.3.1.

Introduction of data ownership introduces responsibility and accountability for one stakeholder to the data source. The data owner is responsible and accountable for quality management, use compliance and security of the data entity. Introducing data ownership is only relevant if the value dimension of the entity is high. Creating responsibility and accountability is a timely and difficult process. The costs and effort cannot outweigh the benefits if the value of the data entity is low. To conclude, the data ownership mechanism is added to category 3 and 4 of the classification framework.

Data stewardship is a less drastic governance mechanism to implement. Data stewards are subject matter experts or business leaders that have detailed knowledge about business and data requirements. Data stewards should be able to translate these requirements into technical specifications [Abraham et al., 2019]. Data Stewards not only use the data entity they also maintain it. Collecting and stating the users of the data entity can be an impractical timely effort. Therefore assigning stewardship is only relevant if one or both dimensions of the classification is high. Insight into the use of the data entity can be obtained by assigning stewardship within the high access requirement categories. Stewardship can be used as a formal term for valuable data entities for adding and assigning rules, policies and procedures to data stewards. The data stewardship mechanism is added to category 2, 3 and 4 of the classification framework.

The data producer is the creator, aggregator or maintainer of a data asset within the organization [Abraham et al., 2019]. The data producer is often also the data owner if implemented. There are cases where The data producer is not the data owner. This Often happens when the customer is the provider of the data. an example can be: A customer is the data provider for their bank account number. Contrary, the owner of bank account data is the finance department.

Writing down the data producer is relevant for all categories of the classification framework. Stating the data producer provides traceability for the data entity regardless of the access or value requirement. The effort of stating the data producer does not outweigh the value of traceability of the data entity.

The data consumer is the user of the data. data consumers use data but do not have any responsibilities for the data entity. A stakeholder is considered a data steward as soon as responsibilities come to play for the data entity. Again this is not the case for the data consumer. The data consumer is relevant for all categories of the classification framework. Therefore the data consumer is used in all 4 categories of the classification framework.

5.4.2 procedural mechanisms

This section will assign the procedural mechanisms proposed by Abraham and Schneider to the categories of the classification framework. Procedural mechanisms aim to ensure confidentiality, integrity and Availability of data within an organization. And can be divided into 9 categories. These categories are: Data strategy, Policies, Standards, Processes, Procedures, Contractual agreement, Performance measurement Compliance monitoring, Issue management.

A data strategy consists of a vision statement, business case, guiding princi-

ples, long and short-term objectives and an implementation road-map. Valuable data entities always benefit from a well defined data strategy. Data strategies help increase and retain the value of a data entity therefore they are always beneficial for highly valuable data sources. Therefore the data strategy is added to category 3 and 4 of the classification framework. A data strategy might be valuable for category 2 as well. Yet the strategy for this category is not centered around value. Instead this data strategy focuses more on the guiding principles. Therefore a smaller data strategy document is added to category 2 of the classification framework. This smaller data strategy only states the vision and guiding principles of the data entity.

Data policies provide rules and guidelines for creation, acquisition, storage, security, quality, and permissible use of data [DAMA, 2009]. In other words these policies are used to retain the reliability and security of the data entity. Therefore this procedural governance mechanism is assigned to the high value categories of the framework. These categories are category 3 and 4.

Data standards ensure consistency and normalization of data within an organization. Data standardization is relevant for data entities that are used by multiple stakeholders. This mechanism provides structure and consistency for the stakeholders of the data entity. The structure and consistency improve the overall availability and accessibility of the data entity. therefore data standards are added to the data categories with a high access requirement. This study acknowledges that implementing data standards may be hard to implement. data classified in category 1 is allowed to be stored without a standard, since this category is not valuable and accessible enough to be a focus area of governance.

Data processes are standardized, documented and repeatable methods used to govern data [Al-Ruithe et al., 2019]. standardized processes make data processes uniform, predictable and repeatable. Standardized processes inform stakeholders how they are allowed to use and process the data entity. This can be easily monitored for compliance. The standardized processes are relevant for data entities with a high access classification score. Therefore the mechanism is added to category 2 and 4 of the classification framework. In category 3 data standards are not advised since the amount of work to implement it does not stand up to the benefit for the stakeholders.

Procedures are: the documented methods, techniques, and steps followed to accomplish a specific activity or task [DAMA, 2009]. Procedures vary widely across teams, domains and organizations. The variety of procedures make them useful to implement for all categories. Procedures can provide techniques and methods for data producers and consumers for category 1. Procedures can help category 2 with providing structure for the interaction with the data entity. For category 3 and 4 procedures can help retain value through uniformity, transparency and standardization. Therefore procedures are added to every category of the classification framework.

contractual agreements state a structured collaborative use of the data entity. There are 2 types of contractual agreements stated by Abraham and Schneider. Service level agreements (SLA) defines what data services will be provided by an internal team or a third-party provider, how the services will be provided,

and what happens if expectations are not met[Al-Ruithe et al., 2019]. Since the SLA governs the interaction of multiple stakeholders it is added to the categories with a high access requirement. Which are category 2 and 4 of the classification framework.

A data sharing agreement (DSA) determines the legal and data governance aspects before two or more parties start sharing data. Data sharing agreements focus on transparent communication between stakeholders. Dsa state obligations between stakeholders in order to retain data quality. Therefore DSA are added to categories 3 and 4 which have a high value requirement that needs to be retained.

Governance mechnism	Classification category	Governance mechnism	Classification category
Execusive sponsor	1,2,3,4	data producer	1,2,3,4
Data governance leader	1,2,3,4	Data consumer	1,2,3,4
Data governance council	1,2,3,4	Data strategy	2,3,4
Data governance office	1,2,3,4	Data policies	3,4
Performance measures	1,2,3,4	Data standards	2,4
Compliance monitoring	1,2,3,4	Data processes	2,4
Data ownership	3,4	Procedures	1,2,3,4
Data stewardship	2,3,4	contractual agreements	2,4

Table 6: Governance mechanisms combined with the classification categories.

5.5 Category 1 low access low value

The Low value low access category contains all the least essential data entities of the organization. Data in this category is supported used by only a few numbers of stakeholders. Therefore, governance is not a priority for these entities mainly because the low value classification does not outweigh the effort of introducing data governance mechanisms. Organizations should be careful with classifying data in this category. This data category should not be a wild card to evade data governance or keep the old ways of working. A relevant governance mechanism for this category is to assess the data producers and consumers of this data entity. Stating the d provides data producers and consumers provides basic insight into the creation and usage of the data entity.

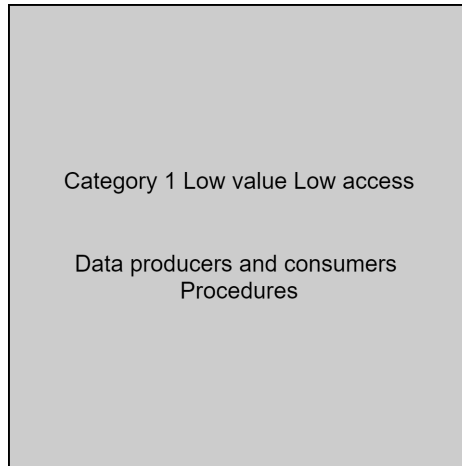


Figure 12: Category 1 of the draft framework

5.6 Category two high access low value

The high access low value category contains the data entities that are used by multiple stakeholders. The data entities classified in this have a requirement to be governed. Even though the value of data entities in this category is low. The governance mechanisms in this category are not centered around responsibilities and ownership. The business value is too low in this category for responsibilities and ownership. On the other hand, are the governance mechanisms aimed to improve and standardize the interaction and use of the data entity. The governance mechanisms most suitable for this data category are data strategy, data stewardship data policies, data standards data processes, procedures and contractual agreements. Next the governance mechanisms and their importance for this data category is explained. A data strategy is fundamental for data entities which are classified in the high access and low value category. The goal is to assess the current state of the data entity and to draft objectives and improvement for the future. A thoughtful consideration should be made about the Value classification as well. Why is the value of this data entity low, and is there a requirement to add value to the data entity? Stating business objectives for the data entity might help answer these questions. The data strategy should also include a vision statement, business case and guiding principles next to the business objectives. A vision statement helps an organization to think about the future of the data entity. The business case describes the justification and highlight the benefits of the data strategy. The guiding principles describe the set of moral values for expected behavior and decision making of the data strategy.

As stated before, this the goal for this category is not to assign data ownership to stakeholders. The value of the data entities is too low to introduce a strict policy at ownership and responsibility level. Stewardship on the other

hand may be of help govern data in this category. Data stewards are business leaders or designated subject matter experts, who have detailed knowledge about the business and data requirements and who can translate those requirements into technical specifications [DAMA, 2009]. The goal is to smoothen the interaction between stakeholders and use of the data entity. Enumerating stakeholders and assigning Stewardship to them enables insight into the stakeholders of the data source.

The next governance mechanism is data policies. “Data policies provide high-level guidelines and rules regarding the creation, acquisition, storage, security, quality, and permissible use of data” [DAMA, 2009]. Data policies are used to communicate key objectives, data accountabilities, roles, responsibilities, and data retention periods [DAMA, 2009]. The aim for the data policies in this category is not centered around data ownership but rather about the interaction between stakeholders. organized data policies can enable structured interaction for data entities in this category. The policies state the rules for proper usage of the data entity. the standardization of policies in this category increase the ease of operation of the data entity while reliability is retained. The goal for the data policies is to enable organization wide usage of the data entity without the entity losing quality or availability.

A Data standard is a technical governance mechanism to effectively standardize the data format for the entity. Data standards ensure that the data representation and the execution of data-related activities are consistent and normalized throughout the organization [DAMA, 2009]. Introducing data standards standardize the creation usage modification and storage to a prespecified format. The standardized format may increase the availability and quality of the data entity [DAMA, 2009]. Another governance mechanism to support the interaction of the data entity is to model and describe the data processes for the data entity. Clear data processes are considered a fundamental element of a successful data governance implementation [Alhassan et al., 2019]. Processes are standardized, documented, and repeatable methods used to govern data [Al-Ruithe et al., 2019]. Modeling the data processes can provide insight into how the data entity is used by stakeholders. Secondly, it standardizes the way of working with the data source. Modeling and standardizing the data processes is the most effective way to guide the stakeholders into a new way of working with the data entity.

Procedures provide documented methods for task accomplishment; they can vary widely. Procedures may describe how to establish the standards and stewardship for the data entity. The procedure for a shared data entity often requires contractual agreements between stakeholders to ensure clarity and proper use. There are 2 contractual agreement that often recur within the data governance domain. These are service level agreements and data sharing agreements. Service level agreements (SLA) defines what data services will be provided by an internal team or a third-party provider, how the services will be provided, and what happens if expectations are not met [Al-Ruithe et al., 2019]. A data sharing agreement (DSA) determines the legal and data governance aspects before two or more parties start sharing data. Sla’s are most suitable for data entities

in this category. a Sla can assign duties and services to stakeholders without the need of ownership of the data entity. Again, the value of the data entity is too low to assign ownership to one of the stakeholders. Instead the contractual agreements, standardized processes and procedures are used to govern and smoothen the interaction between stakeholders for creation usage modification and storage of the data entity.



Figure 13: Category 2 of the draft framework

5.7 Category 3 low access high value

The Low access high value category represents the valuable data entities in the organization that are used by one or few stakeholders. The goal for this category is protection and value retention of the data entities. This category governs and protect business unit or domain specific data entities. The access requirement for the entity is low. Therefore, the governance within this category is not focused on the interaction with the entity. However, the focus for this category lies in ownership, quality conservation, and process standardization. The governance mechanisms suitable for this category are a data strategy, data ownership, process standardization, and procedures. The data strategy for this category focusses on the long- and short-term objectives of the data entity. the data strategy entails how value and quality is retained for the entity and an implementation roadmap for the objectives. A thoughtful consideration should be made for data entities classified in this category's. Why is the access requirement classification low and is there a requirement to add access to the data entity? does the data entity belong in this category or are there other factor that play a role. An organization should assess whether the data entity really shares a low access requirement. Data entities should not be classified in this category purely out of lack of availability. There will be source system which do not have the capability to make data available for other stakeholders. However,

this is not a valid position to classify the data entity in this category. a data entity should be classified in category 4 of the framework If there is a requirement for access. The organization must work out a data availability strategy when this happens.

Data ownership is an important role within this category. Assigning ownership to entities classified in this category is easier compared to other categories. Data entities in this category get more responsibility from stakeholders. Stakeholders naturally have a responsibility for a data entity in this category. The stakeholders manage and use the data entity and have an incentive for maintaining it. The higher the quality of the data entity the more valuable the entity is for the stakeholders. This gives the stakeholders the incentive to retain and improve the data quality of the data entity. Assign ownership does not introduce major changes to the stakeholder therefore it is easy to assign. The ownership states the responsibilities and accountabilities for the data entity. The most drastic addition for the new data owner is that they will be accountable for the data. The data governance team audits the compliance and progress of the data strategy. The data owner is accountable for the progress and compliance.

Data process standardization helps standardize the data processing of the data source. This category is suitable for data process standardization because of the low access requirement. Therefore, the data entity is not used by many stakeholders and most likely in only a few processes. The low number of process stakeholders in this category makes standardization more achievable. In other categories this is more difficult to achieve due to the processes having many different stakeholders. fortunately, this is not the case in this category. The goal for the process standardization is to make the processes that use the data entity repeatable and predictable. The process standardization will improve the quality and reliability of the data entity. Due to the process being described in advance how to be performed.

Procedures recur in almost all categories because they vary widely. The procedures provide methods and technical steps for the data entity. The procedures entail the interaction between stakeholders as well. However, there are little stakeholders in this category. Therefore, the procedures have a different goal in this category. The procedures focus on quality and reliability retention for the data entity. The procedures are not the same as the standardized processes. Procedures are much more detailed compared to processes. Procedures are often part of larger process. The procedures describe how steps for proper use, modification and storage of the data entity.

Additional procedures may be to data enablement procedures. Data enablement procedures entail the necessary steps taken for a data consumer to gain access to the data quality. This procedure might be beneficial to design for the organization to give a few other stakeholders access to the data source. The procedure for data enablement must consider whether the application is legitimate. The data entity might be classified incorrectly when the application is filed by many other stakeholders. The data entity should be classified in the core data category when this is the case.

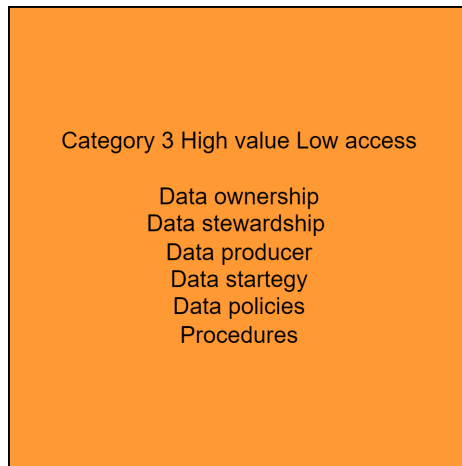


Figure 14: Category 3 of the draft framework

5.8 Category 4 Core data high access high value.

The High value high access category represents the most valuable data within the organization. This category entails all Personal Identifiable Information (PII) sensitive data as well as process essential data that is used by many stakeholders. This category has the highest requirement for data governance. This category features the combined requirements of category 2 and 3. Therefore, the governance mechanisms of categories 2 and 3 are combined to ensure the access and value of the data entity. All decision domains are important for this data entity and many governance mechanisms help to satisfy them. all governance mechanisms proposed by Abraham and Schneider apply to this data category. First an introduction and definition of the roles and responsibilities for this data category are given. Secondly the mechanisms to govern the data category are described. The section is concluded with a description of the performance and monitoring measures.

The Core data category contain the essential and most valuable data entities an organization possesses. The value of these data entities is high are used by multiple stakeholders. To ensure the quality and availability of the data entities a couple of roles and responsibilities are introduces. The Most important role within this category is the data owner. The data owner is responsible and accountable for a data entity in this category. The responsibilities of the data entity are: directing the data stewards, ensuring the quality and security of the data source and executing the data strategy goals. The data stewards use and help maintain the data entity but are not responsible and accountable for it. The data producer is the creator of the data source. Often the data producer is also the data owner, but this doesn't have to be true. For example, a customer of an organization is the data producer of their personal data. yet the data owner of this data is the product owner of the customer management system.

The data strategy is an essential governance mechanism for the core data category. The data strategy aligns the IT and business goals. The data strategy is centered around the vision, long and short-term business objectives and the implementation roadmap for the data entity. The business case and guiding principles are used to support and standardize the interaction between data owner and stewards.

The data policies (the high-level guidelines for creation storage quality and use) are managed by the data owner. The data owner supervises the compliance of the policies by the data stewards. The data policies are designed to guide interaction the responsibilities of the stakeholders. The data policies set and design the playing rules for the data entity.

Data standards ensure a uniform storage specification and representation for the data source. Data standards ensure that the data representation and the execution of data-related activities are consistent and normalized throughout the organization [DAMA, 2009]. Introduction of a data standard is essential for this category. data standardization improves the data availability, quality and usability. The data owners and stakeholders have approve an agreement states what data standard will be used for the entity. 2 governance mechanisms to facilitate this agreement are meta data definition and contractual agreements. Meta data definition captures and describes the meta data of the data entity. The definition improves the understandability and interpretability of the data entity for stakeholders. Secondly, the meta data definition improves the discoverability of the data entity.

contractual agreements ensure that the stakeholders are aware of their rights and obligations. The data sharing and service level agreement state what is expected for both the data provider and consumer. The data owner which is often also the provider oversees the compliance of the agreements. Communication with stakeholders is essential for the data owner to ensure the quality and value of the data entity are preserved . Service level agreements (SLA) defines what data services will be provided by an internal team or a third-party provider, how the services will be provided, and what happens if expectations are not met[Al-Ruithe et al., 2019]. A data sharing agreement (DSA) determines the legal and data governance aspects before two or more parties start sharing data. both agreements are required for data entities in this category. Both agreements are required between stakeholders in this category. Firstly, because it provides the data owner and data provider insight into the data receivers. Secondly because it enables a transparent communication about the obligations and rights for all parties involved.

Another requirement for data classified in this category is process standardization. The processes in which data from this category is used should be modeled and standardized within the organization. An organization must prevent data in this category from being used in an unstructured manner by data receivers. This category can harm the organization if stakeholders do not act cautiously. Therefore, the use is only allowed by a third party when a SLA and DSA is signed between the data owner and the receiver. The legitimate use processes have to be included in this DSA and SLA of the stakeholders to

provide transparency to the data owner, receiver and data governance team.

To provide structure and clarification data owners are highly advised to design procedures for the data entity. as mentioned over and over Procedures provide documented methods for task accomplishment; they can vary widely. Procedures standardize the interactions with the data owner. The data enablement interaction is the most important procedure in this category. the procedure describes the request of use for the data entity. The procedure describes every step which must be taken by the stakeholders.

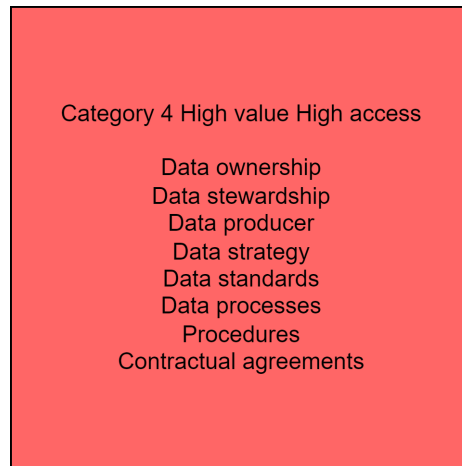


Figure 15: Category 4 of the draft framework

5.9 Summary

This chapter proposed the draft framework for data governance. The draft framework makes use of the classification framework designed in chapter 3. The classification framework was extended with the governance mechanisms from the most suitable data governance framework. The draft framework proposes a new approach for data governance. The framework uses the classification framework to categorize data into 4 governance categories. These categories differ in their governance requirements. Therefore different governance mechanisms are proposed to the categories to optimally support the governance requirements. An overview of the full framework can be found in Figure 16. The next chapter will validate the framework for its useability, validity and robustness.

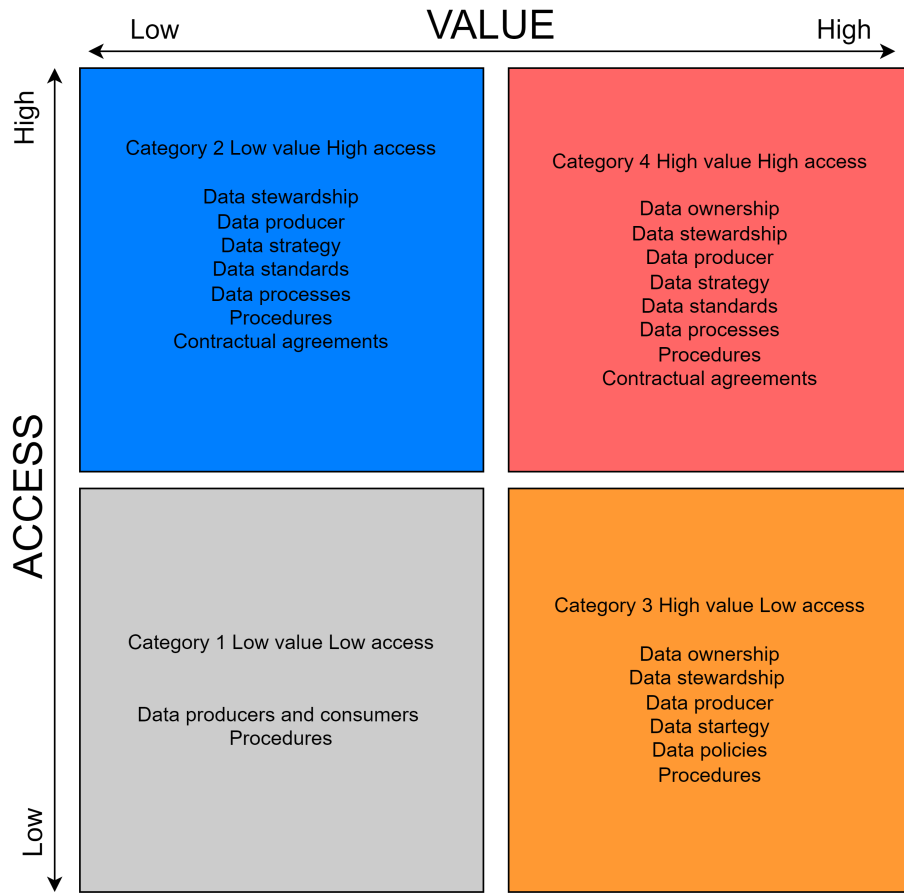


Figure 16: The draft framework for data governance.

6 Validation

This chapter will provide the validation of this study. Firstly the research's validity is evaluated. The measures to tackle the validity threats are evaluated and an overall evaluation of the research is given. Secondly the design artefact is validated. The design cycle of Wierenga was used to design the artefact. The validation will answer whether or not the designed artefact answers the main research question of this study. 2 validation strategies are used to validate the designed artefact. Firstly expert interviews are used to validate the elements of the designed artefact. The simplest way to validate an artifact is by expert opinion. The design of an artifact is submitted to a panel of experts, who imagine how such an artifact will interact with problem contexts imagined by them and then predict what effects they think this would have.

Secondly, Wierenga's Technical action research inference design and validation (TAR) is used to validate if the artefact is capable of solving the problem it was designed for. Technical action research (TAR) is the use of an artifact prototype in a real-world problem to help a client and to learn from this. Usually this is one of the last stages in scaling up a technology from the laboratory to the real world [Wieringa, 2014]. Use cases and workshops within the research object are used for the validation of the technical action research. The Use cases and workshops will simulate the artifact prototype in a real-world problem to validate the artefact.

6.1 Tackling of validity threats

In the beginning of the study multiple validity threats were stated. The internal validity threats are proposed for the feedback provided by stakeholders in the use cases and workshops and by experts in the interviews. Experts and stakeholders from the research object may provide incorrect information for the validation of the artefact. Therefore multiple expert interviews and workshops are conducted to validate this study. The validation and conclusion of the study will be based on a large sample of feedback rather than single instances.

The Second validity threat is the generalizability of the study. Data governance is still a new field of study. Also, the problem investigation phase emphasized the fact that there is no one size fits all approach for data governance. The workshops within the research object might be too specific to generalize the findings of the study. Therefore the validation is not only based on TAR within the research object. The supportive power of expert interviews will enable this study to be validated and generalize the suitability of the designed artefact.

6.2 Research validity

A lot of effort was put into performing a valid study. The literature review stated the infantile state of data data governance literature. This provided pros and cons for the study, the recentness of articles and papers found in the literature was an advantage of the infantile state of data governance literature.

Literature found was often not older than 4 years old and often directly relevant to the study. The lack of data governance frameworks was a disadvantage of the infantile state of data governance literature. Ultimately this had an effect in finding the core elements of data governance. therefore a search for the most relevant framework was opted for. This provided much more structure and a complete set of data governance elements. Choosing the most relevant framework brings a threat to the study. The validity of the choosing process of the most relevant framework. This study aimed and was designed as transparent as possible. Therefore a list of data governance elements was created from the found data governance frameworks. The framework which mostly overlapped the list of data governance elements was appointed the most relevant data governance framework. The study is designed to be repeated in the future. A follow up to this study can be to redo it when data governance literature is more mature. This may provide new insights. A new data governance framework might be designed which better merges with data classification.

6.3 Expert interviews

A total of 3 expert interviews were conducted to validate the designed artefact of the study. All experts work in the field of IT consultancy and have experience with the domain of data governance. The interviews are semi-structured. There are a couple of standard questions which are asked in every interview to the experts. These questions can be found in Appendix A. The standard questions are the foundation of the validation of the artefact. The semistructured interviews enable the interviewer to ask more in-depth questions when necessary. The following couple of sections will review the feedback provided by the experts in the interviews.

6.3.1 Interview one Albert Sprokholt

Albert sprokholt is a data management consultant of Anderson Macgyver. He helps organizations solve strategic problems about information management and ICT. He has also experience in designing suitable business processes for his clients.

The interviewee first provided feedback to the designed dimensions of the classification model. “In my opinion you make a clear distinction between the technical aspects of data and the definition of data with your dimensions. Access specifies the technical access aspects data Whereas value evaluates the definition meaning of the data to the company. The interviewee provided a remark in the data quality measures for the value dimension.” reliability of a system is concerned with the technical aspect of data and therefore it may belong in the access dimension.” This was a little miscommunication. The definition I used in my study for reliability was the degree of trustworthiness of the data entity. Not the technical reliability of the system the data lives in. We both concluded that credibility could be a synonym for the reliability measure of the value dimension.

Next The interview steered towards the interface between the classification model and data governance. This part of the interview is intended to validate whether or not the categories within the classification framework are suitable data governance categories with distinct requirements. The interviewee said that all categories with a high classification have distinct requirements for data governance. His concern was with the low access and low value category. This category should not be neglected in its governance. Please be careful not to neglect this category and make sure that the minimal quality standards are ensured with as little effort as possible. “The high access high value category works the other way round. The organization should in this category focus to ensure all quality standards regardless of the effort.” “ Another concern The interviewee had was how privacy sensitive data is classified in the framework”. My response was that privacy sensitive data always is classified with a high value requirement. The introduction of the GDPR obligated organizations to better govern privacy sensitive data. This resulted in organizations having to classify privacy sensitive data as valuable due to this external obligation. The privacy of these data entities can be protected with process standardization, procedures and accountability. The interviewee agreed but stated that my explanation was not visible at first glance. He advised me to review this part of the model and strengthen it.

Hereafter the Objectivity of the classification framework was validated . The goal for this question is to evaluate the objectivity of the classification model. Is the model capable of objectively categorizing data in the four data categories. The interviewee responded: “ your model is objective in the context in which you work. However outside of your work context your model is rather subjective. Your model is not objective enough to overlap all branches but in specific branch context it can be objectively used.” I informed The interviewee about the situation within Transavia and their concern with the model being possibly too subjective. The interviewee responded that this is not necessarily a problem of the model. The Subjectivity can be tackled through transparent communication by the stakeholders of the classification. The classification is definitely going to evoke different opinions by the stakeholders. The most important thing is that stakeholders do not lose sight of the goal. The goal is to classify all data present within the organization; this can only be done when they agree with their opinions.

The Next topic of the validation interview was governance mechanisms. First the complete set of governance mechanisms was validated. The interviewee reviewed the set of governance mechanisms and asked 2 questions: “ how does your model deal with issue management?” and “how does your model deal with technical system changes?” . My response was “ That issue management is included into the data governance framework but as an overarching structure. Issues need to be resolved regardless of the category the data entity is in. The organization should design detailed procedures for issue resolution within each category. I may have described this too little, and I will highlight this in the manuscript. procedures for Technical system change can be designed into any category of the framework but this is not highlighted in my manuscript.”

The interviewee ended with an addition for future research of the data governance model. “ A great addition to add believability and value to your model might be to add critical success factors to the categories. The classification is more valuable if you know which critical success factors need to be achieved in each category. Your governance mechanisms in these categories then have a foundation to achieve and fulfill the critical success factors.

6.3.2 Interview two Peter Kromhout

Peter Kromhout is a data architect and consultant for Xebia. Xebia is a digital transformation consultancy. Peter has experience in digital transformation and has helped a large Dutch bank with data enablement, lineage and governance. He can be considered a domain expert for data management and governance. The interview was firstly focused around validating the dimensions of the classification framework. The interviewee was firstly interviewed about the suitability of the designed dimensions for data classification. “Access is a suitable dimension for classification, access requirements can be stated objectively and therefore a classification can be done with this dimension” The interviewee was concerned with the value dimension. “ I do not think value is a well-suited dimension for data classification. ”Value can differ between stakeholders and can be very subjective. I do not think an objective classification can be accomplished with this dimension”

Hereafter the interviewee was interviewed about the classification suitability for data governance. The interviewee reacted: “ data governance relates to a broad scale of subjects and classification is one of them. Organizations use data governance to gain insight and control about the use of data. Feeding this back to your model, I think access security is the most valuable measure regarding data governance. access security is most suitable since this measure can provide insight and enable control over data. The other measures for the access category can be used to gain insight and context of a data source but are not as concerned with controlling the data source as access security. The answer to satisfy the accessibility, availability and ease of operation measure is not data governance but rather data distribution and technology. The value dimension and its measures are in my opinion too subjective for data governance. I think objective decision making cannot be achieved with this dimension!“

The interviewee emphasized the close relationship between data governance, data distribution, data enablement, data security and data management. Data governance should be used for decision making concerning the protection and usage of data. Data distribution, enablement, security and management facilitate the implementation of the made decisions. The category validation resulted in an interesting discussion. “The categories are too complex in my opinion given your classification framework. Many organizations are not mature in their data management and strategy” The interviewee emphasized that for those organizations a simpler classification framework with fewer categories is more suitable. “ You may want 2 categories initially for your classification. An example can be a sensitive data and insensitive data category. The goal is to implement

data governance on the sensitive data category first since this category has a higher requirement for governance. insensitive data can be subject to the data governance program as the organization matures and gets accustomed to data governance. Another initial classification which leads to 2 categories can be based upon reliability of the source. In this case the classification is a consideration if the data entity is retrieved from a golden source system. Golden source systems or source of truth systems support, operate or perform business processes and ensure that the organization functions. The data which is created and stored in these systems is highly valuable for the organization. Implementing data governance for golden source systems or sensitive data is a great way to start with data governance. This could be the start of a governance growth model.”

The next subject of the interview was the validation of the governance mechanisms. This part of the interview was focused on validating the existing set of mechanisms for data governance and to figure out whether mechanisms have not been identified. The interviewee believed that the set of mechanisms was complete. However he was concerned with the data standard mechanism. “ I think data standards don’t really fit with data governance. Data standards are too broad of a concept to be applicable for data governance. In my opinion, a large part of this falls outside the boundaries of data governance. data standards are applicable for data governance if an organization chose to implement an organization wide data model. The guidelines and policies for data standards for data standards are subject to data governance. all other facets of data standards fall out of the boundaries of data governance. “ The interviewee concluded that a data governance program can be achieved with the set of governance mechanisms proposed in this study. He emphasized that a governance program is destined to fail when it is not combined with a well-suited data architecture and landscape. “ The data governance program will become a paper tiger when it is not combined with a suitable architecture and data landscape. In this case, the governance is worded very nicely yet highly impractical. the combination of data governance, architecture and landscape will resonate and create combined value of which the organization can benefit. ” The draft framework for data governance was discussed and validated next. Most of the framework was validated by the interviewee. There was only 1 point of criticism. The interviewee emphasized that data strategies for individual entities are rarely implemented. A data strategy is in his opinion an overarching structure for the governance program. “ The data strategy should formulate how to properly interact with data in an organization’s landscape. This is not done on entity or attribute level. ”

The interview concluded with the improvements for the draft framework. We agreed that the first improvement for the framework should be to objectify the classification of data entities. The classification framework was designed to be easily implemented for employees from the business and IT department. The difference in values and perspective between the two was not considered. We acknowledged this resulted in a subjective classification framework. The second improvement that resulted from the interview was to reduce the number

of categories to two for less mature organizations. In this implementation the categories entail whether or the data entity is a focus area for data governance or not. an organization can mature in this model by bringing more data entities in the focus zone of data governance.

6.4 artefact validation within the research object

A workshop within the research object is used to validate the data governance framework. The workshop will test the artifact prototype in a real-world situation. The workshop was held with the Architecture team of Transavia. The Architecture team of Transavia consists of all enterprise architects within the organization. Each domain in Transavia has its own architect and can be considered a domain expert.

The workshop consisted of introducing and explaining the data governance and classification framework. Firstly, the dimensions and data quality measures were explained. Secondly the 4 categories within the framework were explained. The participants of the workshop were able to ask questions after the explanation of the framework.

After the explanation, the participants were able to work with the framework. This started the validation section for the framework. as stated before the goal is to validate if the framework is capable of solving the problem it was designed for. The participants were given the task to classify a given data entity with the classification framework. Hereafter, The reasoning of the participants were asked for the reasoning behind the classification to a specific category. Lastly, the participant was asked what the most suitable data governance mechanisms were for the data entity in their opinion. Next a discussion about the difference in classification between participants was started. The discussion ended with a verdict where all participants agreed in the governance mechanisms and classified category.

6.4.1 Workshop results

In the first workshop the team of participants were able to classify 3 data entities with the framework. The data entities used in the workshop were: Booking, Commercial available seats and Business to business contracts.

The first data entity given to the participants to classify was booking. A booking is described internally as a: “Reservation of one or more seats and SSR’s and ancillary services on one or more segments. Bookings may exist without any booked segments. The booking contains the itinerary. Itinerary refers to the entire reservation (one or more trips).” The booking data entity lives in 2 different systems. The classification of this entity was concluded fastly. all participants were in agreement that booking is classified in the high access high value category of the framework. Additionally all participants were also in agreement of the governance mechanism requirements of the booking data entity. The participants agreed that the complete list of mechanisms are required to ensure this entity is governed. The full list of mechanisms are; Ownership,

stewardship, data strategy, standardized processes and policies, data standard, procedures, and contractual agreements.

The second data entity for the participants to be classified was Commercially available seats. Commercial available seats are described within Transavia as: “ The number of seats on a Segment that are available for sale. The commercial seat capacity does not indicate how many of the available seats are already sold. The number of “seats available” in a CRS (Computer Reservation System) does take this into account.” Commercial available seats are used and calculated by the finance department and used for internal reporting.

This data entity resulted in divided participant opinions. All participants agreed that this data entity should be classified low for the access requirement since this data entity is only used in finance reports. The bottleneck arose with classification of the value dimension. The participant couldn't get to an agreement for the value classification of this entity. The participant argued whether or not the entity is used to sell more seats in the future. Therefore they asked for more background and context information about commercially available seats in order to classify its value dimension. After a voting round high value was chosen after a split decision. All participants agreed that a data strategy could benefit this data entity. The participants agreed that a data strategy could enable, increase, structure and retain the value of this data entity. This strengthened the classification in the low access high value category.

The last data entity to be classified by the participants was the business to business contracts data entity. A business to business contract is an Agreement between transavia.com and a business partner. These contracts live in one system.

The B2B-contract specifies · the period to which the contract applies · the Contract route to which the contract applies · the traffic days to which the contract applies · the number of seats that is allocated to the business partner · the price of these seats · the form of contract · additional terms, discounts and/or supplements

The classification of the business to business contract went exactly the other way around as the classification of Commercial available seats. Here, all participants agreed that this entity is very valuable. The participants had trouble with classifying the access dimension for this data entity. Both sides provided legitimate arguments for their classification. A legitimate argument for a low access classification was: “ this data entity has a low access requirement because it is highly confidential, we do not need interaction standardization for this entity since the entity is already highly protected”. A legitimate argument for a high access classification was: “The contract contains information used in multiple operational processes from pricing to traffic days therefore the access dimension should be high for this entity. ”.

6.4.2 Results and feedback

The Expert interviews and workshop provided interesting results for the validation of the draft framework. The weakest point of the draft framework is

its capability to objectively classify data to their suitable category. Both interviews concluded that the classification framework is subjective in a way. The subjectivity of the framework is due to differences in values and beliefs between stakeholders. The problem investigation of this study stated that data governance often entails organizational and cultural change within organizations. Employees are used to their way of working and do not want to change it. This can create a subjective classification between stakeholders. The framework was not designed with this human / sociological problem in mind. The subjectivity of the framework came to light in the workshop as well. Opinions for the classification of the data entries often differ. In most cases one of the 2 dimensions caused a split decision, and often resulted in a discussion between sides. The allocation of the governance mechanisms to the classification categories were validated. Two remarks were given for the allocation of the governance mechanisms to the classification categories. First, One interview suggested the addition of issue management to the categories. Issue management was placed as an overarching structure for this study and is resolved by policy design and procedures within each category. The following iteration of the framework can provide more guidelines and explain more in-depth how to design issue management within each category. Secondly, one interview suggested removing the data strategy from individual categories and to be added to the overarching structures. A general data strategy can indeed be added to the overarching structures of the framework. Yet, an exclusion to design an individual strategy for a data entity should stay within the categories. A data strategy can differ between data entities classified in the same category. Some data entities might be a focus point for quality, reliability or usage improvement. An individual strategy for these entities can help implement these improvement. The framework might be too complex to implement in immature organizations. Immature organizations are not interested in multiple categories. Data governance implementation for immature organizations is highly understudied in the infantile domain of data governance. An idea for future research would be to optimize the designed draft framework to be implemented within organizations that have no affinity in data governance.

6.5 summary

This chapter provided the validation of the designed artifact. The artifact was validated through semi-structured interviews and a workshop to test the artifact in a real-world situation. The validation resulted in feedback and improvement for the designed draft framework for data governance. The main remark was the subjectivity of the classification framework. The subjectivity can lead to discussion and disagreements between stakeholders of the data entity. The discussion and disagreements often are rooted due to the difference in perspective and objectives of stakeholders. The subjectivity can also be traced back to sociological and human problems that arise with data governance.

7 Discussion and conclusion

7.1 Discussion

This study designed a new approach to data governance to extend the existing base of data governance frameworks. The designed conceptual framework was validated through 2 expert interviews and a workshop. The expert interviews were used to validate the individual parts of the designed draft framework. The workshop was used to test the draft framework in a real world situation within the research object. The validation provided feedback, improvements and future work for the framework. Multiple discussion points can be concluded from the validation. The following sections will explain these discussion points.

7.1.1 Subjectivity in classification of data

A discussion point for the classification part of the framework is that it can result in subjective classification of data. The classification may result in a discussion between stakeholders. Compromises must be made between stakeholders in order to perform an agreeable classification. The discussion and subjectivity of the classification framework does not have to be a weak point of the framework. Perspectives and opinions will always be different between stakeholders. Which is why data governance is as much a human and sociological problem as it is a data management problem. The subjectivity of the classification opens the possibility for discussion. Stakeholders discuss their perspective and explain their justification for the classification. This can lead to more insight, and may align values between stakeholders. Section 1.2 stated that one of the goals of Transavia was to improve business and IT alignment through data governance. The interaction between business and IT wasn't measured during the execution of this study. Yet there is a noticeable change in the interaction between business and IT. The alignment between business and IT has improved through performing this study within Transavia. There is more talking and coordination between IT and business.

Multiple exploratory and preliminary interviews were conducted for the problem investigation and identification. The study and framework provided business and IT an incentive to communicate with each other, align values and create agreements. The problem identification resulted in a shared desired goal for data governance. Which is to protect sensitive data within the organization, reduce risk and gain competitive advantages from data. Even Though the subjective nature of the designed draft framework, Transavia has clearly benefited from the process. The improved communication, understanding of shared goals and agreements improved the overall alignment of business and IT. The improved overall alignment and communication will enable the business and IT units of Transavia to collaboratively align goals and objectives outside the boundaries of data governance in the future.

7.1.2 Complexity of the framework

A second remark for the framework was the complexity of the framework. The literature review emphasized the complexity of data governance. There are a lot of challenges when it comes to implementing data governance within an organization. Secondly the literature review emphasized there is no one size fits all solution for data governance [Weber et al., 2009]. Organizations differ too much in their domains, field of working, organizational and hierarchical structure and IT management maturity. This makes an unified solution impossible. The designed conceptual framework featured 4 data categories. Having 4 distinct categories for data governance may be too complex for a subset of the organizations. The new approach for data governance proposed in the designed conceptual draft framework can overwhelm organizations that lack IT management maturity. A suggestion from the expert was made to reduce the amount of categories in the framework to 2. Reducing the amount of categories can increase the understandability, adaptability and ease of implementation for these organizations. However this will come at the cost of flexibility within the framework. In this implementation it is more difficult to differentiate governance mechanisms between categories.

The lack of categories may be easier to implement for immature organizations. However 2 problems may occur with these little categories. The first problem is that the categories are too generalized. This will result in a subset of data entities not being governed optimally. The governance mechanisms designed in the 2 categories will be too generalized. Which will result in entities that are not governed optimally. A second problem with this implementation arises when an organization tries to optimally govern each data entity with governance requirements. The only solution to do this is to allow unstructured allocation of governance mechanisms to individual data entities. This solution is destined to fail. The unstructured allocation of governance mechanisms will create turmoil within the organization. The confusing situation arises without standards. The opposite of standardization is achieved when data entities have unstructured governance mechanisms.

7.1.3 Overarching structures

The third remark of the framework arose with the overarching structures of the study. The interviews concluded that the relation between overarching structures and the classification framework was not sufficiently highlighted. The overarching elements for data governance play a role with data in a general sense and are applicable to every category of the draft framework. Additionally, they may ensure the implementation and success of the data governance program. In both interviews the lack of explanation resulted in a loss in understanding of the framework. As a result, interviewees often recommended adding overarching structures to individual categories of the framework, not knowing these mechanisms were included in the overarching structures. This remark has to be taken seriously. There might be a concern with the understandabil-

ity of the framework. Stakeholders have to fully understand the framework to ensure adoption and suitable implementation of the data governance program. An extended explanation has to be given to explain the relation between the overarching structures and the mechanisms added to individual categories of the framework. Improved understandability can be achieved through training sessions with stakeholders and or improving the existing explanation given in this report. Further research can be done to improve the understandability of the framework and help stakeholders to work with it. Increased understandability will help demonstrate the value of the framework and improve the adoption of stakeholders.

7.1.4 Decentral vs central organizations

The hierarchical structure of an organization can affect the implementation outcome of data governance. One hierarchical organizational structure might be more suitable for a given data governance approach. This study was performed in collaboration with Transavia. Transavia has a flat hierarchical structure. Transavia's organization is build as an holacracy. "A holacracy is a system of corporate governance whereby members of a team or business form distinct, autonomous, yet symbiotic, teams to accomplish tasks and company goals. The concept of a corporate hierarchy is discarded in favor of a flat organizational structure where all workers have an equal voice while simultaneously answering to the direction of shared authority. [Hargrave, 2021]". The flat organizational structure may have impacted the validation of the research. The validation results of the designed conceptual framework may have differed if the study was performed with a more hierarchical research subject. This comes back to the quote of Weber " there is no single solution for data governance". It may be interesting in the future to study what hierarchical structure is most suitable for this conceptual framework.

7.1.5 addition of critical successfactors

An addition to the framework was given by the interviews to increase the business value of the framework. Critical success factors can be added to the categories of the framework. Critical success factors are requirements for business success. The value of the framework outside the domain of data governance can increase when critical success factors are added to the categories of the framework. The critical success factors help each category to state what has to be achieved for business success. Additionally the relationship between the governance mechanisms and critical success factors can be highlighted. The governance mechanisms can be used to explain how they aim to satisfy the critical success factors of the category. This will increase the overall value of the framework for the organization as well as strengthen its justification.

7.1.6 interpretation of discussion

The designed conceptual framework provides a new approach for data governance. There are remarks for the framework, it is not perfect. However perfection of the framework is never achieved. Data governance will always cope with a difference of opinion and subjectivity. Though this framework can be used to open the discussion for data governance and help support agreement making between stakeholders. The framework is suitable for expansion and future research. The framework can align and support debate to provide an organization with a data governance program. The capabilities, expansion and improvements of the framework can be studied in the future for further improvement.

7.2 Conclusion

The chapters of this study provided the answers to the research questions stated in chapter 2. A copy of the research questions can be found in table 7

Research question (RSQ)	Approach	Chapter/Section
How can we design a data governance framework that uses data classification?	Literature review, expert interviews, answers of sub questions	3,4 and 5
1. What is data governance and what is the benefit of data governance?	Literature review	4.1
2. What are the current solutions and frameworks for data governance?	Literature review	4.3
3. What are the core elements of data governance?	Literature review, Expert interview	4.4
4. What data properties are most suitable for differentiating data categories?	Literature review	3.1 and 3.2
5. What is a suitable data classification framework?	Literature review	3.3
6. What is a suitable draft framework?	Literature review, expert interviews, answer of sub questions	5
7. Can the draft framework be validated as a data governance framework based on data classification?	Expert interviews workshop	6

Table 7: Research question and approach table

In the beginning of this study the value of data was highlighted. To increase the value of data as a business asset, organizations need to establish standards, policies, and processes for the usage, development, and management of data, to create the right organizational structure, and to develop the supporting technology infrastructure [Panian, 2010]. Data governance can provide the solution to fulfill these requirements. Yet, the problem investigation cited that data gov-

ernance is an understudied field. Data governance is often hard to implement even with the high benefits for the organization. Data governance may result in increased data quality, lower data management costs, and increased data security. These benefits are not easily seen unless they are measured over time. Implementing data governance is hard since demonstrating the business value can be hard. Additionally, data governance often results in organizational and cultural change which may create resistance to data governance by employees. This section is the answer to the first research question.

The aim for this study was to provide an extensive literature review about data governance and data classification. This was done to answer research question 2 and 3 and 4. The literature review provided an overview of the current state of data governance and classification literature. The existing frameworks of data governance were reviewed. This overview answered research question 2. The most suitable existing data governance framework was allocated and used to find the core elements of data governance. This answered research question 3. Multiple lists of data quality measures were reviewed to find the measures most suitable for data governance. This was done to answer research question 4. research question 5 and 6 are used to design the conceptual framework for data governance. First a classification framework was designed to answer research question 5. The conceptual framework consists of two individual dimensions which result in four categories for data governance. Each individual dimension is used for data quality measures for its classification. An extended explanation of the dimensions and classification framework can be found in section 3.2 and 3.3. Hereafter governance mechanisms derived from the most suitable data governance framework were allocated to the categories. This resulted in a new conceptual framework for data governance. This draft framework answered research question six.

The conceptual framework was validated through two interviews and a workshop within the research subject. The validation provided valuable feedback on which conclusions can be drawn. The validation of the framework is the answer of research question 7. The main takeaways of the validation were concerns about objectivity and complexity of the framework. The resulting conceptual framework may be too complex to implement in immature organizations. The conceptual framework is destined to fail in these organizations due to a lack of understanding and affinity with data governance. The literature review stated how hard implementing data governance can be. More research can be done to simplify the designed framework and improve its understanding and adoption. Further research has been done before the framework can be released and implemented. an assessment of the complexity of the framework can be made in order to properly validate whether the framework is too complex.

The second concern derived from the interviews had to do with the objectivity of the framework. The classification of data within the framework is often very subjective and dependent on the executive stakeholder. This problem does not only occur within the domain of data governance. Perspectives and opinions between stakeholders always differ. Subjectivity will always exist and this will not be resolved by a more objective framework. The framework is intended to

support the resolution for data governance decisions. The stakeholders should come to an agreement on how to classify the data in the framework. The objectivity of the framework should be put aside and more focus should be put into how stakeholders can come to agreement. The framework can be extended with an agreements / settlement strategy for the classification of data. This strategy can be used to help resolve the disagreements between stakeholders and smoothen the classification process.

The conceptual draft framework for data governance was designed to study a new approach to implement data governance. The main research question for the study was: What would be a data governance framework that is based upon data classification? The goal for the designed framework was to introduce data classification to data governance and provide a new approach. This new approach uses data classification to divide and label categories to categories. Each individual category has its own data governance requirements and mechanisms to solve these. The new approach of the designed conceptual framework is the first of its kind in which a distinction in governance is made based on the classification.

The study has provided the answer to the main research question. a conceptual data governance framework based upon data classification was designed and validated. There is room for improvement and discussion. Transavia experienced increased business IT alignment even though the framework was only tested in the validation phase. The execution of this study provided an incentive for the business and IT and domains to start a conversation with each other about data. Stakeholders shared their perspective and goals for the data governance program with each other. This resulted in increased understanding between the stakeholders as well as finding out common goals. Communication is key in aligning opinions, goals and perspectives. The stakeholders and domain leads continue to communicate in order to align their goals and values in the future. The organization has improved their Business IT alignment through the study.

Data governance is a tool and approach to support the management of data assets. Yet, there are more approaches to help organizations use their data. Data governance is not a holy grail for resolving all data related problems within an organization. The main takeaway of this study is that data governance can help an organization govern their data. Yet, data governance should always be supported with aligned data management, data enablement, data lineage, architecture and modeling. The collective methods provide the fundament for proper use of data and this can only be created collaboratively.

8 Future research and acknowledgement

8.1 future research

Data governance is a relative new subject of study. A lot of research has to be conducted to make the subject of data governance as big as IT security or IT governance. The infantile state of data governance literature and framework makes it an easy subject for future research. This study provided a new approach for data governance in which data classification was used to categorize the data of an organization. In the coming years researchers have to design new approaches for data governance to improve the current frameworks and implementations of data governance. This is a process of trial and error. As mentioned in the literature review: there is no one size solution for data governance that offers an unified solution. The new approaches for data governance might enable data governance to be suitable for a small new set of organizations. a point in time may come in which any organization can implement data governance with a suitable framework. However, in this case a lot of framework still needs to be devised.

Future research can also be conducted to improve the collaboration between data management, data lineage, data modeling and data governance. effective collaborative use of these principles might be valuable. As of now almost none literature exists between the collaborative use and relation between data governance and data management, lineage and modeling.

Future research can also be conducted to extend the conceptual framework designed in this study. Firstly, the complexity of the framework can be studied. An assessment whether the conceptual framework is too complex may be very valuable. Especially when the research also reviews how the framework can be simplified. Overall this will make the framework more understandable and more easily implemented.

Secondly more research can be conducted on how to effectively implement the conceptual framework. A comprehensive case study for implementing the designed conceptual data governance framework may help other organizations implement data governance. The case study can also highlight challenges, bottlenecks and experiences that arose with the implementation. Lastly, research can be conducted to optimize the amount of data governance categories. The validation of this study highlighted that less mature organizations might benefit from less categories. Less categories may increase understanding and feasibility of implementing the data governance framework for these organizations. More research has to be conducted to approve or reject this hypothesis. Future research can also design a maturity model for organizations to grow and learn to implement data governance gradually.

8.2 acknowledgement

This study was performed from march 2021 till January 2022. The pandemic has caused this study to be performed fully digitally. Many participants from

Transavia work fully from home since the start of the pandemic. All preliminary and validating interviews were conducted in a digital environment. The pandemic has caused the execution of this study to be very formal. As a result, possible insights may be lost due to the interaction with the PC instead of an actual person in front of you. I have seen concentration loss from stakeholders in the interviews and workshop. An example is participants scrolling when another participant is talking. The digital approach for this study has caused minimal interaction with the stakeholders. Every meeting had to be planned digitally within timeslots. There was little room for discussion and creation of rapport with the interviewees. Studies have shown that interviews are more successful when there is an apparent and mutual liking between the interviewer and the interviewee. The digital approach made achieving rapport more difficult. The validation workshop also suffered due to a lack of rapport between me and the participants. The workshop could have been more interactive and effective when it was given at a physical location. I am sure that the workshop would have lingered more with the participant when it was given at a physical location. The same holds true for the interviews. Talking in person to each other leaves a better impression and is more memorable compared to video calling. I think more understanding and value could have been created if this study was performed without the preventive Covid-19 measures. I regret that the study couldn't have been performed less formally. Sometimes the most interesting conversations are conducted at the coffee machine. I regret that these conversations could not be included in this study.

Chapter 2 stated that the evaluation phase was scheduled to take place in October of 2021. However, the study was delayed because of my participation in the olympic games. This resulted in an overall delay of about 2 months for the design and validation phase. The delay didn't cause bias in the results. The delay was scheduled between the preliminary interviews and the validation. The initial planning also featured a month to design the framework before validating and testing it. Unfortunately this process now took 3 months to be executed.

I would thank all participants of the study that provided input for this study. This study would not have been possible without the collaboration and transparency of Transavia. I have experienced the cooperation as very pleasant. I will continue to help Transavia govern their data in the future as an enterprise solution architect.

I would like to thank the experts from Anderson Macgyver and Xebia. I am very humble with the feedback I received for the validation of the conceptual draft framework. I know your time is very precious. Therefore, I am very grateful you all took extensive time for me and the validation.

I want to close with thanking Nico Brand, my thesis supervisor of the university of Utrecht. Nico, thank you for the pleasant collaboration and all the help you gave me. I am grateful I got to work together with you and that you had the patience to work with me. Your expertise and network helped improve this study and I owe you a lot. I hope we can help each other in the future and work together to improve data governance and data management in the domain of enterprise architecture.

References

- [GDP,] The EU General Data Protection Regulation (GDPR).
- [fis, 2012] (2012). *Governing Your Data*, chapter 5, pages 61–76. John Wiley & Sons, Ltd.
- [Abraham et al., 2019] Abraham, R., Schneider, J., and vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda.
- [Al-Ruithe et al., 2019] Al-Ruithe, M., Benkhelifa, E., and Hameed, K. (2019). A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing*, 23(5-6).
- [Alhassan et al., 2019] Alhassan, I., Sammon, D., and Daly, M. (2019). Critical Success Factors for Data Governance: A Theory Building Approach. *Information Systems Management*, 36(2):98–110.
- [Altarum,] Altarum. What are Backbone Organizations?
- [Begg and Caira, 2012] Begg, C. and Caira, T. (2012). Exploring the SME quandary: data governance in practise in the small to medium-sized enterprise sector. *The Electronic Journal of Information Systems Evaluation*, 15(1):3–13.
- [Borgman et al., 2016] Borgman, H., Heier, H., Bahli, B., and Boekamp, T. (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*, volume 2016-March, pages 4901–4909.
- [Cheong and Chang, 2007] Cheong, L. K. and Chang, V. (2007). The need for data governance: a case study. *ACIS 2007 Proceedings*, page 100.
- [Council,] Council, C. Contractual agreements.
- [DAMA, 2009] DAMA (2009). *The DAMA Guide to the Data Management Body of Knowledge - DAMA-DMBOK*. Technics Publications, LLC, Denville, NJ, USA.
- [dictionary,] dictionary, C. accessibility.
- [Doyle, 2020] Doyle, A. (2020). What Is a Semi-Structured Interview?
- [Dreibelbis, 2008] Dreibelbis, A. (2008). *Enterprise master data management: an SOA approach to managing core information*. Pearson Education India.
- [Expertian, 2020] Expertian (2020). What is a Data Quality Dimension?

- [FU et al., 2011] FU, X., Wojak, A. ., Neagu, D., Ridley, M., and Travis, K. (2011). Data governance in predictive toxicology: A review. *Journal of Cheminformatics*, pages 3–24.
- [Gartner, 2021] Gartner (2021). Data Strategy Definition.
- [Groß and Schill, 2012] Groß, S. and Schill, A. (2012). Towards User Centric Data Governance and Control in the Cloud. In Camenisch, J. and Kesdogan, D., editors, *Open Problems in Network Security*, pages 132–144, Berlin, Heidelberg. Springer Berlin Heidelberg.
- [Haes and Grembergen, 2008] Haes, S. D. and Grembergen, W. V. (2008). Practices in IT governance and business/IT alignment. *Information Systems Control Journal*, pages 1–6.
- [Hargrave, 2021] Hargrave, M. (2021). Holacracy.
- [Khatri and Brown, 2010] Khatri, V. and Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, 53(1):148–152.
- [Ladley, 2012] Ladley, J. (2012). *Data Governance How to Design, Deploy, and Sustain an Effective Data Governance Program*. Elsevier.
- [Lee et al., 2014] Lee, Y., Madnick, S., Wang, R., Wang, F., and Zhang, H. (2014). A Cubic Framework for the Chief Data Officer (CDO): Succeeding in a World of Big Data. 13.
- [Loshin, 2008] Loshin, D. (2008). *Master Data Management*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [Neff et al., 2013] Neff, A., Schosser, M., Zelt, S., Uebernickel, F., and Brenner, W. (2013). Explicating performance impacts of IT governance and data governance in multi-business organisations. In *ACIS 2013: Information systems: Transforming the Future: Proceedings of the 24th Australasian Conference on Information Systems*, pages 1–11, Melbourne, Australia. RMIT University.
- [Nfuka and Rusu, 2010] Nfuka, E. and Rusu, L. (2010). Critical Success Factors for Effective IT Governance in the Public Sector Organizations in a Developing Country: The Case of Tanzania. *18th European Conference on Information Systems, ECIS 2010*.
- [Niemi, 2013] Niemi, E. (2013). Working Paper: Designing a Data Governance Framework.
- [Okoli, 2015] Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, 37(1):879–910.
- [Oudeman,] Oudeman, J. What does Data Governance mean and what does it do?

- [Panian, 2010] Panian, Z. (2010). Some practical experiences in data governance. *World Academy of Science, Engineering and Technology*, 38:150–157.
- [Perrons and Jensen, 2015] Perrons, R. K. and Jensen, J. W. (2015). Data as an asset: What the oil and gas sector can learn from other industries about "Big Data". *Energy Policy*, 81:117–121.
- [PMI, 2013] PMI, editor (2013). *A Guide to the Project Management Body of Knowledge ({PMBOK} Guide)*. Project Management Institute, Newtown Square, PA, 5 edition.
- [Shaikh and Sasikumar, 2015] Shaikh, R. and Sasikumar, M. (2015). Data classification for achieving security in cloud computing. *Procedia Computer Science*, 45(C):493–498.
- [Stedman and Vaughan, 2020] Stedman, C. and Vaughan, J. (2020). What is data governance and why does it matter?
- [Strategie 2020-2024, 2020] Strategie 2020-2024, I. (2020). Transavia Strategie 2020-2024.
- [Tallon et al., 2013] Tallon, P. P., Short, J. E., and Harkins, M. (2013). The Evolution of Information Governance at Intel. *MIS Q. Executive*, 12.
- [Usgs,] Usgs. Data Sharing Agreements.
- [USGS, 2021] USGS (2021). Data Standards.
- [Verhoeven, 2019] Verhoeven, N. (2019). *Doing Research, The Hows and Whys of Applied Research*. 5th edition.
- [Wang and Guarascio, 1991] Wang, R. Y. and Guarascio, L. M. (1991). Dimensions of data quality: Toward quality data by design. *IFSRC Discussion Paper*, (August):1–21.
- [Wang and Strong, 1996] Wang, R. Y. and Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4):5–33.
- [Weber et al., 2009] 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).
- [Wieringa, 2014] Wieringa, R. J. (2014). What Is Design Science? In *Design Science Methodology for Information Systems and Software Engineering*.
- [Wohlin, 2014] Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *ACM International Conference Proceeding Series*.

A Interview Questions

Access dimension

- Is access a suitable dimension for data classification?
- Is access a dimension a suitable dimension for data governance?
- Are the access dimension measures complete?
- Is there an access dimension measure missing?
- Are all data access measures of added value?

Value dimension

- Is value a suitable dimension for data classification?
- Is value a dimension a suitable dimension for data governance?
- Are the value dimension measures complete?
- Are there any value dimension measures missing?
- Are all data value measures of added value?

Governance mechanisms

- Does the study provide all governance mechanisms relevant within the data governance domain ?
- Are there governance mechanisms overseen by this study?
- Does the study propose too many governance mechanisms?
- Can a governance program be achieved with the mechanisms proposed in this study?

Classification framework

- Would the classification framework work?
- Does the classification framework provide suitable categories for data governance?
- Is the classification framework suitable for classifying data entities within the categories?
- Are there data entities that can not be classified within the classification framework?

Governance mechanism classification

- 4x (1 for each category) Does every category represent the necessary governance mechanism to fulfill its governance requirements?

- Is there a category that misses a mechanism?
- Is there a category with too many mechanisms?
- Do categories provide a suitable data governance framework?