Universiteit
Leiden
The Netherlands

Bachelor Computer Science & Economics

Developing a Prioritization Approach
for AI-related Data Governance

Xin Yi Lin (s3707628)

First supervisor: Drs. Niels van Weeren
Second supervisor: Prof.dr.ir. Joost Visser

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

15/08/2025

Abstract

Background: As data becomes increasingly valuable in the era of Big Data and artificial intelligence (AI), organizations are developing data governance (DG) frameworks to comply with regulations and maximize business value. However, these frameworks often remain more focused on compliance rather than aligning with business needs.

Aim: This study explores how AI-related DG practices can be assessed against regulatory requirements using such an "AI DG framework" (a capability maturity model), how the framework can be refined with best practices, how business value derived from DG can be measured, and how various DG practices can be prioritized in a way that balances regulatory risk and business value.

Method: Employing a combined approach of literature review, document analysis, questionnaires and interviews, we collected data from across five markets of a large HR firm. Data was analyzed by mapping information against predefined maturity level criteria, thematically coding interviews, and conducting a gap analysis between best practices in the framework, literature and the markets.

Results: The maturity assessments revealed similar patterns across markets, namely high compliance in security controls, but lower maturity in roles, data quality and data lineage. Thematic analysis identified data quality, data architecture and metadata management as the top drivers of business value. Integrating best practices from literature and the markets, recommendations were provided to improve the existing framework. Finally, a prioritization model was proposed that highlights areas with high expected business value and low maturity as focus areas for improvement.

Conclusion: This research provides a methodology for DG maturity assessment, insight into the business value of key DG areas and common data challenges, a refined AI-related DG framework, and a risk-benefit-based prioritization model. Future work should explore quantitative methods, place more focus on AI governance practices, and refine the prioritization model.

Contents

Chapter 1

Glossary and acronyms

1.1 Glossary

AI data governance (AI DG) - a term coined by the organization in this study to refer to data governance practices relevant for the development and deployment of AI systems. The company's "AI Data Governance Framework" (Appendix A), which is named after this term, mainly focuses on general data governance for data analytics and insights. However, as the framework also includes AI governance practices, it highlights the importance of data governance in the context of AI. This is especially relevant in light of recent regulation, such as the EU AI Act [29]. In this study, we maintain the same focus on data governance as a whole while also shedding light on the most important AI governance practices.

BlueDolphin - a software used by the organization in this study to document processes, data flows and system architecture.

Confluence - a collaboration and documentation tool used by the organization in this study to share knowledge across the company.

Global Aligned Information Model (GAIM) - a data classification scheme developed by the organization in this study to provide their markets a standardized way for categorizing data.

HR firm - the term we use to refer to the organization in this study.

OneTrust - a software platform used by the organization in this study to manage data assets.

1.2 Acronyms

AI - artificial intelligence.

CIA - confidentiality, integrity and availability (information security triad).

CDM - canonical data model.

DG - data governance.

IAM - identity and access management.

MDM - master data management.

PLS-SEM - partial least squares structural equation modeling.

PQ - priority quotient.

Chapter 2

Introduction

Rapid advancements in artificial intelligence (AI) powered by Big Data have been transforming businesses [26, 22] offering benefits such as increased efficiency and productivity, automation of processes and improved prediction accuracy [39, 4, 50, 85]. However, these opportunities are accompanied by major risks and challenges which arise when data is used for the implementation of AI systems [80]. In order to address these risks and challenges, effective data governance (DG) is essential, which Grover et al. define as "*the overall management of the availability, usability, integrity and security of data* [38]."

One key reason for DG is regulatory compliance. With the aim to mitigate risks related to security, privacy and ethical concerns, organizations developing and implementing AI systems must adhere to regulations and policies. One of these regulations is the new EU AI Act, which entered into force in August 2024 [29], making this a relevant topic for organizations employing AI.

The other role of DG is in maximizing the business value of data. The complexity of Big Data presents many challenges related to data quality, such as incompleteness, inconsistency and inaccuracy. In order to generate reliable and valuable business insights and gain competitive advantage, high-quality data is crucial. Therefore, Adewusi et al. emphasize the importance of clear DG frameworks in order to ensure data quality [3, 38].

2.1 Problem statement

While DG frameworks are widely adopted in organizations, they often inadequately address the business value of DG efforts, focusing more on achieving compliance with regulatory requirements [62, 87, 74, 9, 38, 58, 67]. This gap can also be observed in the case study of a global HR firm where this research is conducted. The organization's compliance-focused "AI DG Framework" (Appendix A) is a capability maturity model that addresses important DG controls. Nevertheless, it does not yet provide a clear guide to evaluate these practices based on their actual contribution to business objectives. As the company prepares to design a DG improvement plan, the need arises for a prioritization approach that helps identify which controls should be addressed first.

In light of the above, the research proposes a prioritization approach for the AI DG Framework to support in achieving business value in a compliant manner. Additionally, we aim to improve the current AI DG Framework with best practices. This will be achieved by (1) using the AI DG Framework to assess DG maturity in the organization, (2) analyzing how DG practices impact business value, and (3) developing a prioritization model that balances the maturity levels of each DG practice with its associated business value.

2.2 Research questions

The study aims to address the following main research question and sub-questions:

How can AI-related data governance controls in large organizations be prioritized to balance regulatory risk and business value?

RQ.1 How can the maturity of AI-related data governance practices be assessed against regulatory requirements?

RQ.2 How can the business value of AI-related data governance practices be measured?

RQ.3 How can the current AI DG Framework be refined with best practices identified from literature and the maturity assessment?

RQ.4 Which areas should be prioritized when implementing an AI-related data governance framework to create and maintain a risk-to-benefit balance?

2.3 Overview of the thesis

The structure of this study is as follows: this chapter presents the topic, problem statement and the research questions. In Chapter 3, we provide a background on the topic DG and the research context. Furthermore, DG best practices, methods to investigate business value of data-related practices, and prioritization approaches are reviewed. This is followed by Chapter 4, which details the methodology used throughout the study to answer the research questions and proposes a prioritization model. Results, such as maturity scores and thematic findings, are reported in Chapter 5, which also demonstrates the proposed model using the results. These results are interpreted and reflected on in Chapter 6. Finally, Chapter 7 summarizes the findings and provides future research recommendations.

Chapter 3

Background and related work

This chapter first provides a background on DG and explains its dual-purpose in mitigating data-related risks and generating business value. Then, we introduce the context of this study, and explore specific DG practices and methods for measuring the business value of DG. Lastly, we touch on existing prioritization approaches.

3.1 Background data governance

Data governance (DG) is defined as the set of processes and policies for ensuring the availability, usability, integrity, quality, and security of data assets [72, 1, 11, 37]. Aimed at managing data effectively, it focuses on the exercise of authority and control rather than the actual execution of data management [52]. Researchers such as Abraham et al. state: "*The purpose of data governance is to increase the value of data and minimize data-related cost and risk* [1, 37]", clearly distinguishing two purposes:

1. Ensuring regulatory compliance.
2. Enhancing business value.

The first purpose is demonstrated by Julakanti et al., who state that the establishment of DG frameworks is necessary within organizations due to regulations and standards such as the GDPR [71]. They confirm through case studies that well-structured governance frameworks lead to fewer data security issues and a higher capability of meeting regulatory requirements [48].

Ladley, whose work focuses more on the business value aspect, explains that DG programs are often created as a result of three concepts [52]. First, master data management (MDM), as DG helps in ensuring that an organization's most important data categories are managed well. Second, data quality: DG helps in maintaining and evaluating the quality of data, and integrating data quality standards into business operations. Third, business intelligence (BI), since DG helps in aligning data analysis with business operations and data quality.

The difference in emphasis between these two authors when describing the origin of DG initiatives in organizations further illustrates the two main purposes of DG.

Recent technological advances, particularly in artificial intelligence (AI), have further shaped the regulatory landscape for organizations that use data and AI. The European Union's AI Act, which entered into force in August 2024 [29], has driven the development of many AI assessment and AI governance (AIG) frameworks in literature. For example, Papagiannidis et al. interviewed representatives of three companies, thematically coded these interviews to find governance best practices, and proposed an AIG model including inhibitors, enablers and outcomes of AIG [63]. Similarly, Lucaj et al. conducted interviews with company representatives and researchers, through which they identified key themes related to AIG and proposed policy recommendations and practices [53]. A third example develops a semi-quantitative AI risk assessment framework by adapting a risk assessment framework provided by the Intergovernmental Panel on Climate Change with the proportionality test, a weight formula which allows for evaluating different types of risks and values in the context of the AI Act [23]. These frameworks and recommendations emphasize the growing need for governance structures that ensure ethical, transparent, and

accountable data use, in addition to more traditional DG aspects.

At the same time, the emergence of big data has offered organizations numerous opportunities for data-driven decision-making through BI, big data analysis (BDA), and data science. As data becomes more central to business processes, DG plays an increasingly critical role in enabling competitive advantage [3]. In a case study by Brous et al., DG was identified as one of the key success factors in realizing value from data science efforts [15].

Despite this, many existing DG frameworks fail to fully capture the business value of DG, focusing more on regulatory compliance. Several studies have highlighted the lack of understanding regarding the impact of DG, data assets, and BDA on firm performance, pointing to ineffective investments and missed opportunities for value creation [62, 87, 74, 9, 38, 58, 67]. This also reflects the challenge of defining and measuring the value of data, as it is intangible in nature [18]. As a result, assessing how DG efforts translate into increased data value is difficult. As Attard and Brennan argue: "*quantification of data value would optimize data governance* [9]." Currently, there is no standardized or objective way to quantify the value generated by different DG practices to evaluate its impact on business outcomes. Subsequently, this creates a challenge for DG improvement projects: it is unclear which aspects should be prioritized, which is critical given that organizational resources are limited. Ladley emphasizes: "*The DG framework will need some sort of guidance as to what is most important. The result will be fewer meetings if a prioritization scheme of relative importance of data areas is recognized* [52]."

Similar challenges are found by the organization in which this study is conducted. The following section introduces the organization and its approach to DG.

3.2 Research context

The study is conducted at a human resources consulting and staffing firm, hereinafter referred to as "the HR firm." The company's main business model matches "talents" (individuals looking for a job) with "clients" (companies looking to hire employees). This matching process is referred to as "assignment." Like many organizations operating in the era of digitization and AI, the HR firm increasingly leverages data and AI to deliver data-driven services, increase operational efficiency and improve decision-making.

In response to the new AI regulations, the HR firm developed an "AI Data Governance Framework" (AI DG Framework), combining requirements from the AI Act, GDPR and other relevant regulations, and covering principles from sources such as the DAMA DMBOK. This framework aims to outline the company's approach to DG in order to ensure compliant, responsible and ethical data usage. Moreover, to increase scalability in the context of digital transformation, the framework intends to support a harmonized DG structure within the organization.

Designed as a capability maturity model, the framework includes five DG categories: roles & responsibilities, data lineage, data quality & purpose, data architecture & platform, and security. Each DG category contains a number best practices, referred to as "controls." There are a total of 28 controls. For each control, there are five levels of maturity described, linked to five risk levels:

1. Ad hoc (high risk): processes are performed as needed and only at the project level. Data issues are fixed reactively rather than proactively through improved processes. Data is not considered a strategic resource.
2. Managed (medium risk): processes are planned and executed within policy guidelines, but the tools and skills for managing data are still inadequate. Data management is taken more seriously.
3. Defined (minimum requirement): standard processes help provide consistent data quality to meet business needs and regulatory compliance. Management oversight has been introduced along with monitoring and feedback loops.
4. Measured (benchmark): process metrics are judged against agreed upon variances. Data is treated as an asset and everyone is concerned with its accuracy and timeliness. Applications are written to capture data issues that are resolved quickly to avoid fines or reputational damage.
5. Optimized (exceeds benchmark): process performance is continuously improved through feedback from various sources. Data is regarded as a critical asset and vital element for our business to operate.

For reference, the AI DG Framework is included in a shortened form in Appendix A. The appendix presents the DG control objectives, summarized descriptions, and their associated maturity level criteria, but omits the specific controls provided in the full framework as this is similar to the control objectives.

This study is part of a broader initiative within the HR firm to start implementing the AI DG Framework in an initial selection of ten markets that the organization operates in. The first part of this initiative is a maturity assessment of these ten markets. The scope of this research is limited to the first five markets that will be assessed, namely the United Kingdom (UK), the Netherlands (NL), Belgium (BE), Germany (DE), and a cross-border service (CBS). This maturity assessment has three objectives:

1. Determine the DG maturity level of each market using the framework.
2. Identify local challenges.
3. Identify best practices that have been effective, which can be included in the framework.

The second part of the initiative concerns the development of an improvement plan; a recommendation of actions to help markets raise the most important DG controls up to a desired minimum level.

While the framework acknowledges data as a business asset, its controls and maturity levels are mainly derived from legal and risk mitigation requirements. This creates room for alignment between the focus of the framework and the HR firm's objectives. The company's broader goal is to enhance DG for effective utilization of data for business insights and realizing benefits. However, this business value perspective is not explicitly included in the current framework. This also creates the following challenge in the development of an improvement plan: "*How should improvement efforts (DG controls) be prioritized in the context of limited resources?*"

This research aims to address this gap by proposing a prioritization approach that considers two dimensions: results from the risk-based maturity assessment and the business value of DG activities. To investigate the latter, Section 3.4 touches on methods to measure the business value of DG in related works. In alignment with the third objective of the maturity assessment, we aim to refine the framework with best practices from the markets as well as from literature. Therefore, Section 3.3 reviews DG frameworks and practices. The last section, 3.5, explores existing prioritization models.

3.3 Data governance practices

To identify key categories of DG, we first conducted a broad exploration of DG and AIG frameworks in literature. The most common categories that have been found and will be discussed are: metadata management, roles and responsibilities, data quality, transparency and explainability, security and privacy, and training and awareness. Table 3.1 presents an overview of these categories. The next subsections investigate each category in more detail, focusing on identifying best practices and supporting these with evidence where available. Alhassan et al. note the inadequate exploration of specific DG activities in academic literature [7]. For this reason, grey literature was reviewed in addition to academic sources, as it provides more specific and practice-oriented insights.

#	**DG category**	**Sources**
1	Metadata management	[1, 6, 53, 41]
2	Roles and responsibilities	[1, 6, 48, 52]
3	Data quality	[1, 6, 45, 53, 41]
4	Transparency and explainability	[45, 23, 13]
5	Security and privacy	[1, 6, 80, 44]
6	Training and awareness	[12, 42, 52]

Table 3.1: Most common DG categories identified from literature

3.3.1 Metadata management

Metadata management refers to the systematic documentation and organization of information about data assets, including descriptions of their origin, structure, meaning, location and use [35, 81]. Essentially, metadata is "*data about data* [49]." Teymurove distinguishes three types of metadata, namely descriptions of technical attributes of data (technical), e.g. file formats, storage locations and database schemas; business contexts (business) described by data definitions, ownership and usage; and data lineage tracking (operational). Effective metadata management is critical for ensuring that data can be accurately located, interpreted, and reused [81].

The importance of metadata management has been highlighted both in academic research and industry case studies. For instance, several case studies have observed increases in data retrieval times and data handling errors [30, 31, 32, 33]. By improving their metadata management, the organizations managed to ensure consistency, streamline operations, increase efficiency and enhance data analytics. Moreover, a study on metadata management within DG frameworks highlights its role in transforming raw data into actionable insights by providing structure and control [81].

Several practices are commonly associated with metadata management:

- Metadata strategy: a structured plan that outlines how an organization will manage metadata across data systems. This includes defined objectives, roles, standards and technologies needed to support metadata-related activities. It ensures alignment with both regulations and business objectives [1, 81, 34, 21].
- Maintaining data catalogs: providing searchable indexes of available datasets along with descriptions and access information. While this is often a centralized platform [81], organizations such as JPMorgan Chase have successfully adopted a hybrid model through a data mesh architecture and a data mesh catalog [10]. The architecture is a decentralized model where data ownership is distributed to domain-specific teams, allowing those closest to the data to manage it. At the same time, an enterprise-wide mesh catalog centralized visibility into data assets. This approach offered several benefits: it enhances regulatory compliance through improved data auditing, supports better data quality control by detecting inconsistencies and outdated information, and creates flexibility by enabling faster data access.
- Data lineage tracking: documenting and visualizing how data moves through an organization's systems, from its source, through transformations, to its final use [81]. A related practice, data lifecycle management, also includes determining the data definition, retention and retirement [49, 1]. These practices combined enhance transparency and accountability, increase data quality by allowing easier error detection, and prevent costs of holding unnecessary data [84]. For example, Panasonic implemented this through a centralized DG platform which tracks data flow throughout its entire life cycle [76].
- Data taxonomy: a classification system containing metadata standards such as ownership, purpose and sensitivity, used to label and organize data consistently [49, 83, 1]. Guided by this taxonomy, tagging and classification are processes for applying these metadata standards. These practices foster collaboration and a common understanding among teams, and help in creating well-structured data catalogs.

3.3.2 Roles and responsibilities

Roles and responsibilities form the backbone of effective DG, as they ensure clear accountability and that data assets are managed, protected and utilized optimally [7, 1, 17, 52]. Having well-defined roles not only streamlines data-related processes but also fosters a culture of responsibility and trust across the organization. Furthermore, it is essential to have roles and responsibilities documented and embedded into the organizational structure to ensure that they support company goals [1].

Several key roles were highlighted in literature:

- DG council: a group of executives from departments involved in data processes. They set the overarching direction for DG initiatives and policies, and ensure alignment with organizational objectives [52, 17, 43, 61]. While [17] positions this group at the strategic level, the DAMA DMBOK considers it part of the tactical layer [43]. DAMA and [61] assign strategic responsibilities, such as

aligning DG initiatives with business objectives and providing sponsorship, to a Chief Data Officer or a similar C-level executive, for example, a Chief Technology Officer.

- Business owners of data assets (Data owners): typically senior business leaders who are accountable for ensuring data quality, compliance and appropriate usage of data (including access rights) within their department or domain [49, 61, 43].
- Data custodians: often part of the IT department, they are responsible for the technical environment, ensuring data is stored, archived and protected appropriately. As a bridge between strategic intent and operational execution, they are also tasked with resolving issues experienced by data users, which may be reported to the DG council [17, 43].
- Data stewards: responsible for executing tactical plans by data owners and custodians in specific domains, they bridge the gap between business requirements and technical requirements. Their responsibilities extend to organizing meetings with the data users and providing them educational support [17, 61, 43].
- Data users (operational): consists of all individuals who interact with data in their roles, such as data scientists and data analysts [6]. These individuals are responsible for adhering to data policies and reporting data issues [17, 43].

3.3.3 Data quality

Data quality refers to the degree to which data is accurate, complete, fresh (up-to-date), consistent, valid, unique (deduplicated) and reliable [1, 49, 59, 52]. High-quality data is essential for effective decision-making, efficiency and regulatory compliance, especially with the involvement of AI [86, 59]. Amongst others, Al-Badi et al. describe that measuring and monitoring data quality should be regarded as a prioritization within DG [6]. This is also reflected in a literature review of DG case studies, which suggested that the documented cases often concentrated on the single aspect data quality [84]. The process of conducting regular data quality checks is further emphasized throughout literature [1, 6, 49, 84]. Benefits of actively monitoring data quality were evident in a case study of Rijkswaterstaat; it contributed to finding weaknesses in DG, aligning processes and achieving cost efficiencies [15].

Best practices for ensuring data quality include the following:

- Establishing data quality standards and metrics: defining clear metrics, KPIs and rules/standards for data quality that align with organizational objectives [49**?** , 42]. This means that data quality dimensions may be defined differently and measured through different metrics depending on an organization's context [49, 55]. Nevertheless, Parmiggiani and Grisot explain through their case study that the focus should be on ensuring data is fit-for-purpose rather than acquiring as much data as possible, which can undermine overall data quality [49, 64]. This practice is typically the role of data stewards [68].
- Continuous (automated) profiling and monitoring: metrics such as the aforementioned data quality dimensions could be assessed using various automated methods, including statistical analysis, rule-based checks, default values, latency measurement, hashing, stability checks, ML techniques for error correction, imputation and rule compliance [59]. Tools such as the Great Expectations tool, which has built-in data validation, profiling and documentation features [79], have proven to be effective in detecting quality issues. For example, [60] describes its ability to assess data quality in a Brazilian governmental data warehouse where it identified problems in public procurement and expenditure data.
- Data cleansing: correcting or removing data that does not adhere to standards [42]. Automated data cleansing processes could be implemented using technologies such as ML algorithms, Asynchronous Search Algorithm and Java Parser [59]. There are several studies that have developed methods and solutions for this, including [51, 36, 78].
- Reporting: summaries describing the data quality against the predefined KPIs, which could also be in the form of dashboards. Along with enhancing quality monitoring and supporting early detection of issues, it is a way to communicate to stakeholders. This practice ensures alignment between business and IT [17].

3.3.4 Transparency and explainability

Transparency refers to the openness and clarity with which organizations handle data, ensuring that stakeholders understand how data is collected, processed and used. Explainability, particularly in the context of AI and ML, involves making the operations and decisions of complex AI models clear and understandable to humans [61, 20]. These mechanisms are crucial for building trust, ensuring compliance with regulations, and facilitating accountability in automated decision-making processes. Given that complex AI systems are often a black box, these aspects mainly recur in AIG frameworks [45, 63, 23, 13, 25], while they are less emphasized in more traditional DG frameworks that focus on data in general.

The following transparency and explainability mechanisms are often mentioned in AIG frameworks [45, 63, 13]:

- Implementing Explainable AI (XAI) techniques: adopting XAI methods helps making complex AI models more transparent, making their decision-making processes more interpretable. This is essential for stakeholders to trust and effectively oversee AI systems. For instance, techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) provide insights into model predictions [23, 20].
- Clear documentation: maintaining comprehensive documentation of data sources, processing methods, and decision-making criteria enhances transparency [63].
- Continuous monitoring and evaluation: regularly assessing AI systems and data processes for transparency and explainability helps identify risks and areas for improvement [47]. Using dashboards these results can be visualized to easier communicate information regarding AI systems to humans [63].

3.3.5 Security and privacy

Security and privacy measures encompass the policies and procedures implemented to maintain key attributes of data, such as accessibility, authenticity, availability, confidentiality, integrity, privacy, and reliability [1]. In other words, it ensures that data remains protected against unauthorized access, manipulation or loss while supporting its trustworthy use across the organization.

The following security and privacy practices are common within DG:

- Risk assessments: analyses conducted by security officers on risks in regards to critical business data in order to implement protection measures [49].
- Data storage based on business value: strategically allocating data to suitable storage platforms based on business relevance enables organizations to distribute it efficiently across multiple systems. This not only optimizes use of storage and reduces costs, but also aids in fulfilling archival compliance requirements [9, 49].
- Access controls: controls that define to which extent an individual can interact with the organization's systems or files. Using access controls, the authority to access data should be minimized to the extent that a user only sees what is necessary [45, 63]. In role-based access control, access rights are determined by the role a user has within the organization, while attribute-based access control takes a more dynamic approach, using user traits, environmental context and resource characteristics to govern access decisions [8]. Role-based DG also was embedded in Panasonic's data lineage platform, for which the company developed schemas for each business unit depending on their data needs [76].
- Encryption: transforming data into an unreadable format to ensure that sensitive information remains protected from unauthorized access or interception. Common methods are hashing, symmetric encryption and asymmetric encryption [44, 8].

3.3.6 Training and awareness

While training and awareness are often overlooked in DG frameworks, Ladley emphasizes their importance in supporting cultural change [52]. Training and awareness involve ensuring that stakeholders understand the value of DG, are informed about data policies and standards, and are equipped with the necessary skills to fulfill their roles in maintaining data. This practice fosters accountability, promotes

data stewardship across the organization, and helps embed DG into daily operations. It is considered as a success factor for DG programs. Bernadi et al. found that "*The most reported error prevention activities were the continuing education of professionals with regular training of data collectors during their studies* [12]." The author also recommends to address behavioral and structural issues, alongside improving technical capabilities.

The following practice is therefore defined:

- Fostering collaboration and training: communicating, educating and training stakeholders involved in data processes regarding the importance of data and skills necessary to handle data correctly [12, 42, 52]. Ladley distinguishes three training levels [52]:
  - Orientation: providing an overview of DG.
  - Education: building awareness and understanding so that employees know how to follow and apply policies and processes.
  - Training: delivering practical instructions to develop the skills needed to work effectively with new tools and processes.

3.4 Measuring business value

In this section, we explore methods for estimating value from data-related practices. Although research on the direct measured impact of DG is limited, insights can be drawn from related studies that examine the value of information (technology), as well as the relationships between BDA capabilities and business performance. This approach aligns with the theory of Resource-Based View (RBV), which is also a foundation for many of the papers examined below [65, 66, 69, 5, 46, 57, 73]. The RBV suggests that configurations of valuable resources and capabilities can be sources for sustainable competitive advantage [70]. Organizations with similar resources, such as data or information, may perform differently depending on how these resources are combined with capabilities, such as DG practices, BDA capabilities and IT capabilities [65].

3.4.1 Cost-benefit analysis

One way to evaluate the value of data-related activities is through cost-benefit analysis (CBA), which attempts to quantify value by directly comparing input costs with outcomes through metrics.

For instance, Wittard et al. attempt to quantify the value of improved DG using a framework that combines two existing frameworks [87]: the "Five Safes" framework [24] for data collection and the "Theory of Change" framework [82] for analysis.

- The study collects data through interviews with project managers. This process is guided by the Five Safes framework, a structure that enables DG to be organized as a set of connected dimensions, namely: safe projects (scope), safe people (roles), safe settings (access), safe data and safe outputs. For each of the five safes, challenges, best practices and implications were identified from the DG improvement project.
- Based on these findings, a mapping is conducted to develop a theory of change for the quantitative assessment. This mapping involves constructing a pathway with intermediate outputs between the initial challenge and the intended objective. The theory of change was then used as a CBA framework where input costs and benefits were estimated for the activities that lead to the intermediate outputs. Assuming that benefit from an event is at least equivalent to the costs of an individual's participation, benefits were measured across three categories: direct internal value (wages of the individuals attending a workshop), indirect internal value (time savings from improved processes measured in average hourly wage) and external tangible value (number of publications produced).

A comparable method is provided by Otto in his case study of the relationship between product data management and business goals [62]. This research also uses a framework combining two existing concepts: business engineering and business dependency network.

- Following business engineering, the scope for analyzing benefits was determined as the three layers of the organization where a transformation project has impacts, namely: strategy, organization and business processes, and information systems.
- The business dependency network was employed to analyze the paths between IT enablers and the company's business goals, which is mediated by enabling changes, business changes and business benefits. The benefits of the business changes were then quantified using cost analysis, comparing costs from before the changes to costs after improving the processes. This includes determining total sales, general and administrative expenses related to product data management for new product creation and existing product maintenance, and calculating the costs per new and existing product.

Another CBA approach applied in a related context is found in a study by Pathak et al., which focuses on using financial ratios [65]. The study analyzes the impact of IT practices on an organization by estimating their value and comparing it to industry benchmarks. First, the study presented a comprehensive literature review, which explored various input and output measures for evaluating business value of IT. Four categories of measures were identified:

- Financial measures, such as gross margin and cost ratios.
- Operational performance measures, such as efficiency and productivity.
- Process measures, such as improvement and satisfaction.
- Perceptional measures, such as information availability.

For their analysis, the researchers narrowed their focus to the measures expenses and revenues. To account for the differences between firms in the industry, they used ratios for these measures, namely: IT expenses to total revenue ratio (strategic) and IT expenses to total expenses ratio (operational).

These studies demonstrate how value of data-related initiatives can be translated into a cost-benefit estimate by defining metrics and determining intermediate outputs. While they provide a way to quantify DG value, their limitations are the need for assumptions, extrapolation. In addition, these methods rely on detailed project-level data and cost estimates, such as IT expenses or wages and time savings from improved processes, which are very time-consuming to calculate for each DG practice and not available to our research. Instead, our prioritization model conceptually draws inspiration from these approaches.

3.4.2 Causal inference through statistical modeling

A similarity in several CBA approaches is the use of path modeling to connect enablers, such as DG practices and IT changes, to business outcomes. These models emphasize how smaller practices or changes contribute indirectly to business value by enabling intermediate improvements. In the studies examined above, such models were referred to as "theories of change" and "means-end relations" [62, 87]. This same logic is the foundation for statistical modeling techniques applied in the works that will be reviewed below. While the earlier approaches use it for CBA, these works aim to empirically validate the assumed relationships in these paths using observed data.

Partial least squares structural equation modeling

One recurring method in literature for exploring relationships between data-related capabilities and firm performance is partial least squares structural equation modeling (PLS-SEM) [66, 69, 2, 19, 5, 46, 57, 73]. It is a statistical method commonly used in social sciences to study complex relationships between latent constructs, as it models the latent constructs through multiple observed indicators, also referred to as measurable items. Hence, it is particularly useful in data-related contexts where constructs are complex to measure or quantify [40].

Several studies use PLS-SEM to model the impact of BDA on business outcomes [66, 69, 19, 5, 46, 57]. For example, in a study of SMEs in Singapore, constructs like system quality, information quality, lack of infrastructure and data security concerns were used to assess their influence on perceived business value and performance [66]. All variables were measured via 5-point Likert-scale survey items, such as 'The data analytics software should provide a complete set of information' for the variable information quality, and 'Data analytics enhances employee productivity' for the variable DA business value.

Another example is a study by Raguseo and Vitari, who segmented business value related to BDA into transactional, strategic, transformational and informational value [69]. Rather than separately measur-

ing BDA and treating business value as a mediator between BDA and firm performance, their model positioned these business value components as the direct outcomes of BDA. For instance, informational value was measured using the items 'Enabling faster access to data', 'Improving management data' and 'Improving data accuracy', which are comparable to the variable information quality in the previous study ([66]) and certain topics of DG.

In two other models, information quality was broken down into similar indicators, namely completeness, accuracy, format and currency [5, 46].

Several papers also applied this technique in the context of DG [2, 67, 57, 73]. A study demonstrated that DG positively affects sustainable knowledge creation and firm performance using PLS-SEM [2]. The researchers chose PLS-SEM over covariance-based structural equation modeling (CB-SEM), a similar statistical technique, due to its flexibility with complex models and its fit for exploring relationships.

Similarly, Pestana proves with PLS-SEM that data-driven culture, data processes and data stewardship contribute to value creation in different steps of the business process. The writer also mentioned this technique's flexibility with non-normal distributions and suitability with complex models [73]. Additionally, its ability to deal with smaller sample sizes was mentioned as a benefit over CB-SEM.

While these papers differ slightly in the constructs and outcomes they measure and the items used to measure them, the consistent use of PLS-SEM highlights its usability in contexts with latent variables such as DG. The use of Likert-scale surveys is also comparable to our DG maturity assessment where we use five levels of maturity. The similarities with these studies and our research proves its potential use in assessing the business value of DG practices. However, the small sample size of five markets limits the feasibility of applying PLS-SEM in our study. Nevertheless, we draw inspiration from these papers and recommend this method for future studies with larger samples.

Econometrics studies

Another group of studies have also investigated similar relationships by applying econometrics methods, which, rather than relying on self-reported perceptions, use financial data with statistical models that control for confounding effects. Examples of the data-related capabilities examined in these studies include BDA [58], data-driven decision-making [16] and BI systems [28]. For instance, Müller et al. employed a Cobb-Douglas production function, a formula that estimates outputs (Sales) produced by specific inputs, which often used for measuring firm performance [58]. In the function used in the study, a binary dummy variable 'BDA' was added to the independent variables in order to assess effects of the presence of BDA assets in an organization. Using three regression methods (OLS, FE and 2SLS), the coefficients of this function were estimated, where the coefficient of the BDA variable represents percentage change in productivity attributable to the presence of BDA assets.

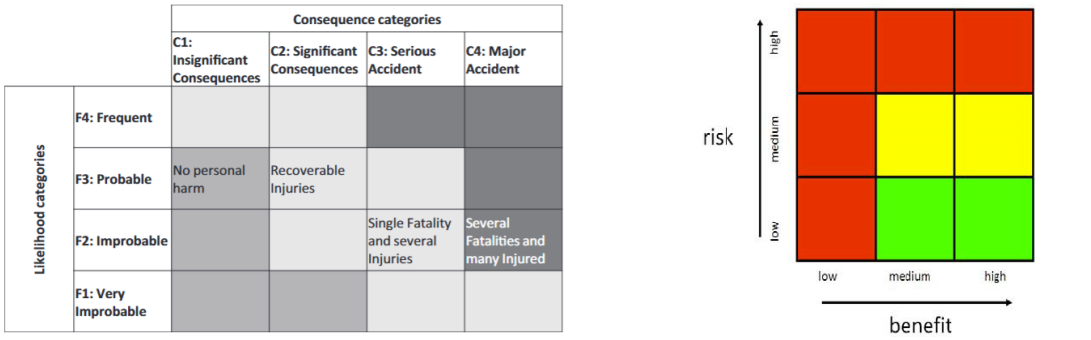
Overall, these econometric methods offer a valuable quantitative evidence of the relationship between data-related capabilities and business value. However, this approach requires detailed objective financial data and perfectly measured variables, which are not available in this case study. Therefore, this research draws on their insights conceptually rather than replicating the method.

3.5 Prioritization models

Organizations frequently face the challenge of allocating limited resources to activities that yield the highest value for the business while simultaneously considering risks [77]. To support effective decision-making, various prioritization models have been developed.

One example is the risk matrix, a tool for systematically assessing and comparing the potential impact (consequence) and likelihood (probability of occurrence) of risks or benefits within an organization [27]. It is commonly used in investment decision-making, and widely adopted into various other contexts where organizations create custom versions of the matrix. Impact and likelihood are often rated on a 1-5 scale or as low/medium/high. As a result, risks can be plotted on the matrix, which contains color-coded cells, allowing the user to assess the risk level and clearly see a prioritization between the risks. An example is provided in Figure 3.1a. The (top) right cells indicate higher risk events that should be prioritized. Despite the simplicity and flexibility of risk matrices, there are a few criticisms. The most important examples are the degree of subjectiveness in the assessment of the dimensions and the increased complexity of prioritization when risks land in the same cell.

Figure 3.1b illustrates a risk-benefit matrix, a variation of the risk matrix for conducting benefit analyses [14]. The researcher proposes an approach for weighing decisions regarding solar energy use in historical buildings. Risk represents the potential compromises that exist in using it, while benefit captures the positive value derived from it. The bottom right area here is seen as priority, as it provides quick wins, while the top and left cells (either low risk or low benefit) suggest low-priority.



(a) Example of a risk matrix [27]

(b) Example of a risk-benefit matrix [14]

Figure 3.1: Comparison of risk matrix approaches

A third prioritization model is the Eisenhower Decision Matrix developed by U.S. President Dwight D. Eisenhower [54]. It is a tool that helps individuals and organizations categorize tasks based on urgency and importance. As shown in Figure 3.2a, the matrix has four quadrants, recommending the following priority order: do (do now), decide (do later), delegate and delete. While the model provides a clear and actionable framework, it also has the limitations of the risk matrix. Additionally, it lacks consideration for resource constraints.

A similar model is the Accessibility Governance Matrix presented by [56], which combines the concepts of the Eisenhower Matrix with the Analytical Hierarchy Process (AHP) for a case study on geographical accessibility planning. There are two types of models proposed. The first model prioritizes based on a priority quotient (PQ):

$$\text{PQ} = E_f \times U_f \times I_f \times C_f$$

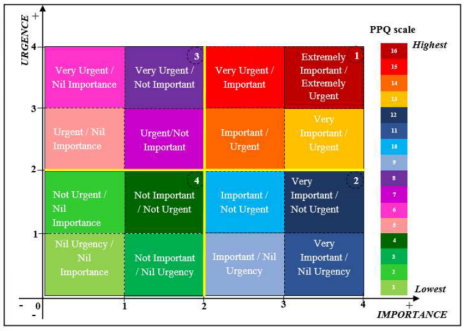
where: E_f = ease , U_f = urgency, I_f = importance, and C_f = consequence; all rated on a 1-4 scale.

The second model, a simplified version of the first model, is designed for non-technical managers to prioritize geographical districts. Prioritization is based on a partial priority quotient (PPQ):

$$\text{PPQ} = (U_f \times I_f) - AI$$

Rather than four factors, the simplified model only considers urgency and importance. Furthermore, an accessibility index (AI), is used as a correction factor to deprioritize already-accessible districts. The AI is calculated using AHP, a decision-making method for ordering activities into a hierarchy by comparing and assigning weights to all possible pairs of activities. The resulting PPQ can be plotted in the Accessibility Governance Matrix shown in Figure 3.2b.

(a) The Eisenhower Decision Matrix [54]



(b) The Accessibility Governance Matrix [56]

Figure 3.2: Comparison of Eisenhower Decision Matrix approaches

In the specific context of DG, Ladley has suggested an approach similar to these models for prioritizing subject areas within a DG program [52]. Figure 3.3 displays a graph with the two axes risk and business value. Business domains are plotted on this graph based on their estimated risk and business value, resulting in four quadrants. Items in the top right draw the most attention, as they represent high value and high risk. The least important items, low value and low risk are found in the bottom left, which are low-priority areas.

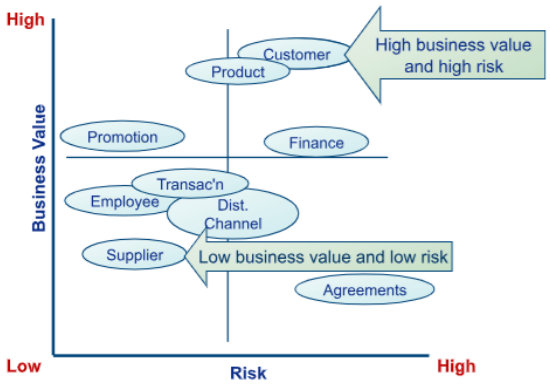


Figure 3.3: DG prioritization model suggested by Ladley [52]

These models illustrate how organizations can compare factors such as risk, benefit, urgency or importance to prioritize actions. Drawing inspiration from these examples, this study will adapt their foundations to develop a prioritization model for DG in Chapter 4, weighing the factors regulatory compliance risks and business value.

Chapter 4

Method

This chapter outlines the different methods used to collect and analyze data. The last section of this chapter describes the approach we used to develop a prioritization model. The data collection phase and maturity level assessment were performed together with a data protection consultant and data protection officer from the HR firm.

4.1 Data collection

Data collection included a preparation phase, questionnaires and interviews; a process that was repeated for each of the selected markets. A high-level overview of the data sources per market is provided in Table 4.1.

Market	Informal Interviews	Questionnaire	Workshop	Deep dives
UK	0	No	1	2
NL	0	Yes	1	2
BE	0	Yes	1	2
CBS	2	No	1	0
DE	0	Yes	1	1

Table 4.1: Overview of data sources per market

4.1.1 Preparation

As preparation for the questionnaires, the 28 DG controls from the AI DG Framework were first consolidated into fourteen questions. This consolidation ensured that we captured a higher-level overview of similar topics, avoided asking complex details, and maintained a concise questionnaire.

Furthermore, an initial data analysis was conducted to review available information from the markets, such as policies, documentation on data management platforms and data quality dashboards.

During the preparation phase, we also conducted informal interviews with a variety of individuals, mainly key IT system owners. These interviews aimed to identify data sources for the initial data analysis, discuss the feasibility of extracting relevant information from data management platforms, and determine which individuals could further provide us necessary data. In some instances, these conversations also offered valuable market-specific insights regarding their DG organization. This was particularly the case for CBS, as illustrated in Table 4.1.

4.1.2 Questionnaires

The fourteen consolidated questions formed the baseline of the questionnaires. The initial data analysis further informed the design of the questionnaire for each market, resulting in two types of questions. The first type addressed the topics for which there was no information available; these were left open for respondents to answer. The second type focused on topics with a pre-filled answer, which asked respondents to validate this information and make updates if it was incomplete or outdated. In several cases, these validation questions were followed up by a more detailed follow-up question to explore the topic further. This approach helped reduce respondent burden and increase efficiency during the questionnaire phase while allowing us to gradually ask more detailed questions.

To strengthen the credibility and allow for verification of the answers, each question also required the respondent to provide evidence to their responses, e.g. a link to where documentation about this was stored.

4.1.3 Interviews

With the exception of two markets, we conducted three semi-structured interviews for each market (as shown in Table 4.1):

- Workshops with individuals in DG roles or roles involved with data, depending on the organization/maturity of a market's DG. Shifting the focus from the fourteen higher-overview questions to the specific DG controls described in the AI DG Framework, the workshops aimed to elaborate deeper on questionnaire responses and inquire about technical details. Additionally, for the purpose of proposing a prioritization approach based on business value, the workshop also included questions on perceived and measured business value of DG.
- Deep dives with client data owners or roles involved with client data processing. Since individuals in these roles are more business-oriented, the deep dives not only addressed unresolved topics from the AI DG Framework, but also focused on data challenges from a business process perspective.
- Deep dives with talent data owners or roles involved with talent data processing. Same purpose as the client data deep dives.

Each interview lasted approximately an hour. The number of participants varied widely per interview and market, ranging from one individual to a group of six, which reflects the structure of the DG organization within a market.

With permission of academic and company supervisors, who deemed this method appropriate and efficient, part of the interviews was transcribed using the AI assistant Gemini. The transcriptions were validated against the recordings, corrected where necessary, and summarized. Interviews that were not conducted in English were recorded and manually transcribed. This was the case for all three interviews with the markets the NL and CBS, and the workshop with BE.

4.2 Data analysis

To address the different research questions, data from informal interviews, questionnaires, workshops and deep dives was analyzed using three different approaches.

The maturity assessment (RQ1) was conducted using a structured template, the business value analysis (RQ2) followed a coding approach, and the gap analysis (RQ3) involved comparing the HR firm's AI DG Framework with findings from literature and the maturity assessment.

4.2.1 Maturity assessment

To determine the maturity levels of the DG controls in each market, we created a structured template based on the AI DG Framework. This template was implemented as a table in Google Sheets, with one sheet per market. Each DG control in the AI DG Framework was represented as one row in the table. The template included the following columns:

- **DG Control**: the specific DG control as described in the AI DG Framework.

- **Maturity Levels**: a reference column describing the criteria for each maturity level (1-5) for the DG control.
- **Client Data**: all information gathered that concern the handling of client data.
- **Talent Data**: all information gathered that concern the handling of talent data.
- **General**: all information gathered about the overall DG approach in a market.
- **Evidence**: the source or location of the information in previous columns, e.g. references to a policy document (Data Protection policy) or data management platform (OneTrust, Confluence).
- **Reason for level**: a brief reasoning for why a certain maturity level was assigned, based on all gathered information.
- **Current Level**: the maturity level assigned for this DG control in that market.

After the document reviews and interviews were completed, information was manually added to the corresponding columns per DG control. It was frequently the case that a single piece of information was relevant for multiple DG controls. For instance, the quote '*[Person 1] is accountable for arranging all data access. Together with [Person 2], she also arranges ad-hoc access for specific projects; she approves requests from roles that require certain data access for a certain period of time.*' is relevant for both 'Clear roles and responsibilities' and 'Role based access controls'. In such instances, it was entered into each applicable row to ensure the maturity of each control could be independently assessed based on all relevant evidence.

For each control, we compared the collected information against the maturity level definitions to determine the appropriate level. This structured method provided a good overview of the gathered information per DG control and facilitated quick assignment of maturity levels. The results of the maturity assessment are described in Section 5.1.

It is important to note that the interviews during the data collection phase were not solely focused on gathering the information needed to assign maturity levels for each DG control. Rather, our focus was often on understanding market-specific challenges and potential improvement areas, which aligns with the second objective of the maturity assessment as described in Section 3.2. This decision, approved by the data protection officer, was necessary due to time constraints during the interviews and the need for the outcome of this project to aid the development of an improvement plan for each market. Consequently, for some DG controls we did not collect sufficient information to accurately determine their maturity levels. In many cases, it was because these controls, e.g. (test-)data distribution, were less immediately relevant for the markets' current state of DG. Questions regarding more foundational controls, such as data quality and roles & responsibilities, were prioritized as they were more frequently mentioned as challenges that impacted the business processes.

In addition, we noticed that certain controls and their defined maturity levels were too advanced relative to the current state of DG practices in many markets, making them difficult to measure. For example, some level descriptions required a measured implementation percentage (25%, 50%, 75%, 100%) for controls where a foundational activity or measurement process was not yet in place. In both of the described instances, we either did not assign a maturity level to the DG control, or we assigned an average of two levels. Cases where no maturity level was assigned, are marked in the results tables using the abbreviation "N/A" (not assessed).

4.2.2 Business value analysis

To analyze the business value of DG practices for the purpose of developing a prioritization plan, we thematically coded all interview transcripts using ATLAS.ti, a qualitative data analysis tool. This method draws inspiration from the approaches described in [63] and [53], introduced in Section 3.1. Using the same method, we also attempted to identify best practices used in the markets.

The coding process was guided by the following questions:

GQ.1 Which DG practices or aspects do stakeholders explicitly associate with business outcomes, and have they observed positive/negative effects?

GQ.2 What are the most impactful or recurring data-related challenges stakeholders encounter in their business processes?

GQ.3 What DG practices have been effective or are seen as a strength in this market?

Applying the inductive approach described by Saunders et al. [75], the coding process consisted of three steps:

1. **Open coding**: each line in the transcripts was read to find quotes related to the three guiding questions. These quotes were labeled with descriptive codes that captured their core meaning. Codes were reapplied across similar text segments, allowing us to identify recurring answers and patterns between different interviews in one market and between different markets. To facilitate the second step of this process, a systematic coding prefix was applied on the codes based on the three guiding questions. ATLAS.ti automatically recognizes prefixes and groups codes accordingly. This approach allowed for more efficient grouping and retaining distinction between remaining challenges in a market and already implemented practices. The prefixes applied were the following:
   (a) 'PBV' for codes related to perceived business value from DG practices (GQ.1).
   (b) 'DC' for codes that described the most important self-reported challenges (GQ.2).
   (c) 'BP' for codes that we identified as best practices (GQ.3).
2. **Axial coding**: the initial codes were first refined and reorganized by consolidating overlapping codes, and deleting irrelevant codes. Then, we explored the relationships between codes by grouping them into broader categories.
3. **Selective coding**: with the aim of identifying a narrative that relates the categories together, we investigated the categories established in the previous step.

The results of this analysis are described in Section 5.2.

4.2.3 Gap analysis

To find opportunities for improvement in the AI DG Framework, a gap analysis was conducted. This process involved comparing the DG practices identified in the literature and in the five markets with the practices proposed in the AI DG Framework. The identification of market practices was based on the thematic analysis described in the previous subsection, particularly guided by the third guiding question (GQ.3). The best practices from each market were consolidated and subsequently compared to the list of practices found in the literature review (Section 3.3) and the controls outlined in the AI DG Framework. This comparison enabled us to assess differences, similarities and the completeness of the current framework. These findings are presented in Section 5.3.

4.3 Model development

The last objective of this study is to design a model that will aid the HR firm in determining a prioritization plan with the most impactful controls, balancing regulatory compliance risk and business value. Therefore, we propose a DG prioritization model that builds on the principles described in Section 3.5.

The design of the model mainly draws inspiration from two approaches. First, the prioritization graph suggested by Ladley [52] plots DG domains on a grid with business value and risk, identifying the areas that should be prioritized first (high value, high risk). This model provides a simple way to visualize prioritization based on the two dimensions that our study also focuses on. However, it does not clarify how business value and risk should be assessed or measured for each DG control. To quantify these dimensions, we define a priority quotient (PQ) inspired by the partial priority quotient used for the Accessibility Governance Matrix [56]. For each DG control, a PQ can be calculated which indicates its urgency in the prioritization plan.

The formula is defined as follows:

$$\text{PQ} = \text{Business value} - \text{Risk} = \frac{(PBV + PCI + C)}{3} - ML$$

where:

- PBV = perceived business value
- PCI = perceived challenge intensity

- C = complexity of implementation
- ML = maturity level

Our model uses the priority quotient above to balance the two primary factors business value and risk. The core idea is that DG controls with higher expected business value and higher regulatory risk should be prioritized first, as these controls offer the greatest benefits while also posing the greatest threat if left unaddressed. In other words, improving implementation in these areas can yield the most benefits while also mitigating the most compliance risks.

Risk is measured using the maturity level (ML) of a DG control, as this indicates the residual risk. Its position as a correction factor in the formula is inspired by the accessibility index used in the Accessibility Governance Matrix, which deprioritizes already-accessible districts [56]. In our model, a high maturity level means that the DG area is already well-developed (low risk), so there is less residual risk left to mitigate. For this reason, maturity level acts as a deprioritization factor in the formula. This results in a lower urgency in the prioritization for higher maturity levels. For instance, a control that offers high business value but has a high maturity level will have a lower priority compared to a control with the same business value but a lower maturity level.

Business value is defined by three variables: perceived business value (PBV; consequence/importance), perceived challenge intensity (PCI; urgency), and complexity of implementation (C; ease). These variables are inspired by the variables in the Accessibility Matrix [56]: ease, urgency, importance, urgency. All variables are rated on a 1-5 scale, following the scales in the AI DG Framework (Appendix A) and in the models discussed in Section 3.5. Unlike the priority quotient in the Accessibility Governance Matrix [56] in which factors are multiplied, these variables are summed. This method was chosen since each variable in our model contributes independently. To ensure that the maturity level can meaningfully reduce the priority quotient, we normalize the business value side by dividing the sum by the number of variables (3). As a result, both factors have a range of 1-5.

The following subsections explain each variable in more detail and provide guiding criteria for determining the scores on a 1-5 scale. In the last subsection, we provide an alternative version of the formula that incorporates an additional variable PLS-SEM, a statistical technique discussed in Section 3.4.

4.3.1 Model variable - PBV

Perceived business value reflects the degree to which a control is expected to contribute to business value, as perceived by stakeholders. It is informed by interview insights gathered from all markets. This aligns with guiding question GQ.1 from Section 4.2.2, which explores which DG practices have been most impactful or are expected to generate business value. By coding interview data, we can compare more easily how often each DG practice is mentioned in connection with valuable outcomes.

We define the following scores:

1. Very low: the DG control is not mentioned by any market as beneficial or is only mentioned as non-beneficial. The category that the DG control belongs to is barely mentioned (contains the least number of codes).
2. Low: the category that the DG control belongs to is mentioned (contains codes about other controls), but the control itself is not explicitly discussed.
3. Moderate: the DG control is mentioned a few times, or the category it belongs to is among the most mentioned categories.
4. High: the DG control is mentioned multiple times or by multiple markets with strong statements about benefits.
5. Very high: the DG control is repeatedly emphasized by stakeholders across multiple markets as beneficial with strong statements.

4.3.2 Model variable - PCI

Perceived challenge intensity shows the extend to which a control is associated with current data-related challenges, as perceived by stakeholders in a specific market. It is informed by market-specific interview insights and coding of challenges and needs. This aligns with guiding question GQ.2 in Section 4.2.2,

which explores what practical DG challenges a market faces locally. By coding interview data per market, we can connect DG controls to experienced pain points, inefficiencies, or needs.

We define the following scores:

1. Very low: the DG control is not mentioned as a challenge in the market, or is only mentioned as a solved issue or minor inconvenience; its category is barely mentioned.
2. Low: the category that the DG control belongs to is mentioned several times as a challenge area, but the specific control is not explicitly discussed as problematic.
3. Moderate: the DG control is mentioned as a challenge or cost driver, or its category is one of the more frequently discussed challenge areas for the market.
4. High: the DG control is discussed as a source of multiple issues in the market with clear examples of challenges and costs.
5. Very high: the DG control is repeatedly described in strong terms as a large obstacle or source of costs, or is indicated as a major need.

4.3.3 Model variable - C

Complexity represents the expected cost and time required to improve a DG control. This variable is assessed based on how resource-intensive the implementation is likely to be. It considers aspects such as the technical difficulty, cost, and duration. The scores for this variable have an inverse relationship to the degree of complexity to account for the idea that higher complexity should reduce the priority score.

We define the following scores:

1. Very high: the DG control is highly complex to implement, as the current DG organization is not prepared to support the changes.
2. High: the DG control requires significant coordination, technical development, or process changes that take considerable time and investment.
3. Moderate: the DG control requires noticeable implementation effort, requiring multiple teams or systems to align, but is manageable with available resources.
4. Low: the DG control requires some planning and coordination but can be implemented with moderate effort and cost.
5. Very low: the DG control is quick and easy to implement with minimal resources and process changes.

4.3.4 Model variable - PS

In Section 3.4, PLS-SEM (PS) was a recurring statistical technique employed in studies investigating relationships between data-related initiatives and business outcomes. These studies had a number of similarities with this study.

Many of the reviewed papers used Likert-scale questionnaires to measure latent constructs such as DG impact, system quality, and perceived business value, which aligns with this study's aim of analyzing how DG practices contribute to business outcomes. Moreover, constructs like information quality were used to represent DG or data quality, further demonstrating the relevance of this technique for evaluating DG practices. Similar to these studies, our approach uses the maturity levels of specific DG controls as indicators of how well different areas of DG are implemented within an organization. Several studies also modeled business value or firm performance as dependent variables, which fits the structure of this study's model. These conceptual and methodological similarities support the choice of PLS-SEM as an appropriate tool for exploring the impact of DG maturity on business outcomes across markets. The outcome of the PLS-SEM analysis would provide statistical weights for each DG practice, representing its impact. However, our limited sample size prevents us from employing this technique in our study, as it is difficult to conduct a statistical analysis with a sample of five instances. Nevertheless, we recommend that future studies or organizations with a larger sample size test this approach, as it can provide a data-driven way of measuring the business value of a DG practice.

For the business value side, we recommend choosing measures such as those identified in the study by Pathak et al., described in Section 3.4.1 [65]. Specific measures depend on the organization's business model and the availability of measured financial and operational data. For example, in the context of the HR firm, operational performance measures could have been "fulfillment rate" (effectiveness), "time to fill" (efficiency), and "net promoter score" (client satisfaction).

The adjusted formula is as follows:

$$\text{PQ} = \text{Business value} - \text{Risk} = \frac{(PBV + PCI + PS + C)}{4} - ML$$

This proposed prioritization model is demonstrated in Chapter 5.

Chapter 5

Results

This chapter presents the results from the maturity assessment, thematic analysis, and gap analysis. The prioritization model proposed in Chapter 4, which makes use of these results, is demonstrated in the last section of this chapter.

5.1 Maturity assessment results

This section describes the results of the maturity assessment. We first discuss each market separately, followed by general observations across markets. The results are presented in tables where the first column contains the controls as described in the AI DG Framework (Appendix A), the second column provides the maturity levels, and the third column briefly describes the reasoning for assigning the level. Controls where no level could be assigned, are marked with "N/A", as explained in Section 4.2.1.

5.1.1 Maturity assessment results - UK

The results of the maturity assessment for the UK are presented in Table 5.1.

DG Control	**Level**	**Reasoning**
Defining and documenting roles and responsibilities	4	Roles and responsibilities are formally defined and documented. Documentation was not completely up-to-date.
Integrating roles and responsibilities into organizational structure	4	There is documentation of the organizational structure, but this was not completely up-to-date.
Reporting lines for DG	2	Reporting lines are understood but not specifically documented.
Defining master data	3	CDMs are mapped to GAIM, creating a local model. Data management approaches vary by business line.
Documenting data preparation activities	N/A	Not specifically discussed/known.
Developing data cleaning guidelines	2.5	Data cleansing is included in the sales framework. There are still conversations happening on establishing data cleansing processes.
Developing data transformation/enrichment standards	N/A	Not specifically discussed/known.
Classifying data according to GAIM (classification scheme)	N/A	Not specifically discussed/known.
Developing data minimization, aggregation and anonymization standards	N/A	Not specifically discussed/known.
Developing data sharing standards	3.5	There are controls in place to prevent exports. They also use role-based access controls in multiple systems.
Developing third party data standards	N/A	Not explicitly discussed/known.
Documenting purpose of personal data processing	3.5	Processing activities are documented. OneTrust is however not completely up-to-date.

Centrally documenting purpose of tools/applications	3.5	Assets are documented. OneTrust is however not completely up-to-date.
Categorizing data according to taxonomy	N/A	There is a local model which is mapped to GAIM. It has not specifically been discussed how much of the data is classified.
Formulating completeness criteria	N/A	Fields that are required are mandatory fields within UFO. Fields that are not necessary and that are also not mandatory are not often used. For how much of the data this applies was not specifically discussed.
Defining accuracy metrics	N/A	Fields that are required are mandatory fields within UFO. Fields that are not necessary and that are also not mandatory are not often used. For how much of the data this applies was not specifically discussed.
Measuring freshness	N/A	Fields that are required are mandatory fields within UFO. Fields that are not necessary and that are also not mandatory are not often used. For how much of the data this applies was not specifically discussed.
Measuring consistency	N/A	It was indicated that the data is not completely consistent; there exist multiple versions of the same data. For talent data, the systems require the data to go through Salesforce UFO, which enforces consistency. It was not specifically discussed how much of the data is consistent.
Measuring uniqueness	N/A	It was indicated that there exist multiple versions of the same data. There is also use of shadow data. However, there has been a data cleansing, related to the implementation of the sales framework. They also indicated that a lot of work has already been done to deduplicate data. It was not specifically discussed how much of the data is deduplicated.
Knowing (test-)data distribution	N/A	Not specifically discussed/known.
Measuring and reporting data quality	3.5	There is now an initial dashboard from this framework that provides a baseline to improve on and visibility of sales segmentation categories & allocation. So data quality is being tracked and reported on. They are working on corrective actions regarding behavior.
Developing data strategy and architectural guidelines	N/A	Data strategy and architectural guidelines were not discussed in enough detail to assign a level.
Inventory data assets in a data catalog	3	Data assets were cataloged in OneTrust. This list was not completely up-to-date.
Rating data assets on CIA	3	The CIA ratings are described in OneTrust. However, not all assets are up to date in OneTrust. Some data assets do not have a CIA rating in OneTrust.
Restricting access based on roles	4	There are specific permission sets to allow certain roles read or editing rights. Some parts of the system are more closed off. Access via sharing rules ins managed through Salesforce public groups. Access controls and permissions for primary data source are under constant review. Compliance and sensitive data is processed strictly within the system.
Periodically reviewing access rights	3.5	Access, which is managed by edit permissions and visibility, and divided based on roles, is constantly under review every 12-18 months.
Documenting data retention policy	4	Personal data is automatically removed 3 years after a candidate has not worked with the company. There are also other retention rules.
Storing data based on localization regulations	N/A	Not specifically discussed/known.

Table 5.1: Maturity assessment for the UK

The maturity assessment for the UK shows that roles and responsibilities are generally well-defined and integrated into the organizational structure, although documentation was outdated.

Controls related to data standards and quality, like data cleaning, transformation, classification, and quality measurement, are more mixed. Some areas like data cleansing and data quality dashboards are

in development, but these are not yet established as standard processes. Other parts such as having clear transformation standards or consistent classification were not discussed in enough detail to assess. This indicates that while operational processes exist, more formalization and documentation are needed to reach higher maturity.

Access controls and retention are relatively strong. Role-based access and permission reviews are done regularly, which results in an above-average maturity level for data protection. However, not all data assets have up-to-date documentation or CIA ratings. This suggests that active management of documentation can still be strengthened.

A general remark for the UK is that tools like OneTrust are being used well, but they are not always fully maintained. This shows that processes for keeping them current could be improved.

Overall, the UK's DG organization has clear roles and a good security focus, but more work is needed on formalizing data standards, keeping documentation up-to-date, and ensuring consistency across business lines.

5.1.2 Maturity assessment results - NL

Table 5.2 shows the maturity assessment results for the NL.

DG Control	**Level**	**Reasoning**
Defining and documenting roles and responsibilities	2.5	Some roles are formally assigned, some informally. No specific data owners but there is EDU in place formally.
Integrating roles and responsibilities into organizational structure	2.5	Responsibilities are appointed informally. There is a division between responsibility and accountability. Responsibilities of EDUs are defined, which also makes them feel responsible for the data itself. Roles are not clearly integrated into organization.
Reporting lines for DG	N/A	Data users can communicate data issues about data to EDU, however, clear reporting lines were not discussed.
Defining master data	4	They use a local model that resembles GAIM, which is mapped to GAIM where possible.
Documenting data preparation activities	N/A	Not specifically discussed/known.
Developing data cleaning guidelines	N/A	Cleaning is part of ETL done via DBT, which indicates that it is standardized. How broadly it is used is unknown to us.
Developing data transformation/enrichment standards	N/A	Transformation is part of ETL done via DBT, which indicates that it is standardized. How broadly it is used is unknown to us.
Classifying data according to GAIM (classification scheme)		They have their own classification system, and controls for when confidential information is shared. However, how much of the data is specifically classified is unknown to us.
Developing data minimization, aggregation and anonymization standards	N/A	There are minimization, aggregation and anonymization standards. However, it is unknown to us how broadly this is applied. Whether non personal data is used in non-production environments was also not discussed, but where possible, personal data is masked.
Developing data sharing standards	4	There are data sharing and access controls based on the data classification. How often this is reviewed and/or updated was not discussed in the workshops. When looking at reports, only the information about data that your role is allowed to access, is shown.
Developing third party data standards	N/A	Not specifically discussed/known.
Documenting purpose of personal data processing	2.5	Information in OneTrust is not completely up-to-date, some activities might be missing. This indicates that it is not regularly maintained.
Centrally documenting purpose of tools/applications	N/A	Information in OneTrust is not completely up-to-date, some assets are missing. This indicates that it is not regularly maintained. Non-personal data was not specifically discussed.
Categorizing data according to taxonomy	4	They use a local model that resembles GAIM, which is mapped to GAIM where possible.

Formulating completeness criteria	N/A	There are some data quality controls in place, such as mandatory fields. However, it is unknown to us for how much of the data this is applied.
Defining accuracy metrics	N/A	For urgent processes such as financial processes, the data is accurate. However, for other processes, there have been cases where data was not completely accurate. For how much of the data the accuracy has been measured, is unknown to us.
Measuring freshness	N/A	Talent data is up-to-date if the talent feels the urgency to keep it up to date. Otherwise it will become outdated quickly. How much of the data is up to date is unknown to us.
Measuring consistency	N/A	Between different systems, the information is inconsistent, which leads to extra work for consultants to find the accurate data. This will improve when transitions to [new platform] are complete. How much of the data is consistent is unknown to us.
Measuring uniqueness	N/A	There are data duplications & shadow data. How much of the data is unique is unknown to us.
Knowing (test-)data distribution	N/A	Not specifically discussed/known.
Measuring and reporting data quality	1	Data quality is not being reported on.
Developing data strategy and architectural guidelines	3	There is a document with the architectural principles of the data platform. Some details have changed, but the main points are still valid. How broadly it is applied is unknown to us. They have also indicated that systems are not well integrated, which leads to non optimal use of data and inconsistency. This will be improved however when transition to [new platform] is completed.
Inventory data assets in a data catalog	5	They have their own data catalog, which is up-to-date because it is a mandatory part of data ingestion.
Rating data assets on CIA	4	There is a CIA requirement from the IT. However, within OneTrust not all assets have a CIA rating; not all assets are included or up-to-date. During the deepdive it was mentioned that there were other CIA ratings used in the NL (outside of OneTrust). This was not specifically discussed.
Restricting access based on roles	4.5	Access to data is based on roles. This is also enforced when someone's role changes. When data is downloaded, it is unknown what happens afterwards.
Periodically reviewing access rights	2.5	It is documented which roles have access, with room for improvement, especially for historical data access.
Documenting data retention policy	2.5	Data retention agreements are defined at the source system. The source sends delete/anonymize orders to the Data Lake, which then processes them down the chain accordingly in the data warehouse and Tableau. How often this is reviewed was not discussed.
Storing data based on localization regulations	5	All data in stored and processed within the EU. Partially within AWS services, Snowflake services and Tableau services.

Table 5.2: Maturity assessment for the NL

The maturity assessment for the NL shows mixed results. Some elements have a high maturity level: there is a well-developed local data model that aligns with the global framework (GAIM) where possible, and all data is stored and processed within the EU, which supports compliance with localization and privacy requirements. The data catalog is a strength too, which is up-to-date because it is a mandatory part of data ingestion.

Furthermore, access controls are mature; there are role-based access controls and enforced adjustments when roles change. Minimization, aggregation and anonymization standards are also in place. While there are requirements from the IT-department on CIA ratings, not all assets and CIA ratings are consistently updated in tools like OneTrust. This shows that tool maintenance and completeness can be improved.

On the other hand, some controls around roles and responsibilities and reporting lines are less mature. There is no clear appointment of data owners, and some responsibilities are assigned informally. This indicates that accountability and clear governance structures still need to be formalized.

For data quality, there is room for improvement. There is no reporting on data quality, and issues like inconsistencies and duplication remain, partly due to system integration challenges. These should, however, improve after the transition to [new platform].

Finally, the documentation of purposes is in OneTrust, but this information is not fully up-to-date, showing again that documentation maintenance is a recurring area for improvement.

Overall, the NL shows good technical controls and compliance, especially for storage, access, and classification. Nevertheless, they could benefit from establishing clearer ownership structures, better integration between systems, clear data quality standards, and processes for regularly updating documentation.

5.1.3 Maturity assessment results - BE

The results of the maturity assessment for BE can be found in Table 5.3.

DG Control	**Level**	**Reasoning**
Defining and documenting roles and responsibilities	3	Some roles and responsibilities are assigned. However, they are still in the process to document and appoint roles more formally.
Integrating roles and responsibilities into organizational structure	3	Responsibilities are appointed, but due to complexity not completely integrated into the organizational structure. They are still in the process of arranging this.
Reporting lines for DG	1.5	Clear reporting lines have not yet been established.
Defining master data	4.5	They have a local version of GAIM (BAIM), which maps to local definitions and fills in gaps missing from GAIM.
Documenting data preparation activities	N/A	There are targeted standards regarding data preparation. However, it is not clear to what extent the actvities are known.
Developing data cleaning guidelines	N/A	There are targeted standards regarding data cleaning. However, it is not clear to what extent the actvities are known.
Developing data transformation/enrichment standards	N/A	There are targeted standards regarding data transformation. However, it is not clear to what extent the actvities are known.
Classifying data according to GAIM (classification scheme)	4	They have data classified in GAIM that is linked to BAIM.
Developing data minimization, aggregation and anonymization standards	1	Anonymization efforts have constraints because of the Belgium government.
Developing data sharing standards	2	Some access is manually managed through authorization model. No clear sharing and visibility standards exist.
Developing third party data standards	N/A	Not specifically discussed/known.
Documenting purpose of personal data processing	3.5	Original purposes were described. The responsibility of updating OneTrust does not lie by the same person who enters new data.
Centrally documenting purpose of tools/applications	3	Most assets are included in the assets lists in OneTrust, but this is not completely up to date.
Categorizing data according to taxonomy	4.5	They have data classified in GAIM that is linked to BAIM
Formulating completeness criteria	2.5	Relevant data is available. For client, application level data is missing. Data migration has some criteria set, but locally this is not being monitored.
Defining accuracy metrics	N/A	For client, additional search for information and combining different sources is required. For talent, it is not sure if the data is accurate, as this relies on the talent. There are some data controls to help ensure accuracy.
Measuring freshness	N/A	Active talent data is generally up-to-date. There is a process in place to keep it up-to-date. The Dun & Bradstreet database is not always up-to-date.

Measuring consistency	N/A	For client, there are multiple sources with conflicting information. For talent, there seems to be one version of the truth. Controls to manage unique talent profiles are also in place.
Measuring uniqueness	N/A	For client, there are multiple sources with conflicting information. For talent, there seems to be one version of the truth. Controls to manage unique talent profiles are also in place. Shadow data does exist.
Knowing (test-)data distribution	N/A	The government tests with production data.
Measuring and reporting data quality	1	There is no documented process for when issues are identified
Developing data strategy and architectural guidelines	3	There is a data strategy and architectural guidelines. This is being implemented step-by-step.
Inventory data assets in a data catalog	3.5	The list of assets is mostly up-to-date. However, periodic reviews are not done for all assets.
Rating data assets on CIA	4	Assets in OneTrust have a CIA rating.
Restricting access based on roles	3.5	Security focuses on role-based access to systems. In addition, the data owner is involved in granting access to talent data.
Periodically reviewing access rights	N/A	Not specifically discussed/known.
Documenting data retention policy	3.5	Retention managed within systems in known. But it is not clear what happens after an export. In some systems, controls are in place to prevent this.
Storing data based on localization regulations	N/A	Not specifically discussed/known.

Table 5.3: Maturity assessment for BE

The maturity assessment for BE shows that several foundational areas are actively being developed, but there are differences in how well controls are embedded. One of the strengths is the local data model (BAIM), which extends the GAIM to fill local gaps. This shows alignment with the framework while adapting it to fit local needs. Data is well-classified and categorized, and assets are rated on CIA in OneTrust.

There are also role-based access controls in place, and the involvement of data owners in granting access further strengthens this practice. However, clear reporting lines and formal ownership structures are still lacking, with many responsibilities informally appointed and documentation incomplete. This means accountability is not yet fully embedded in the organizational structure.

Documentation of the purposes for personal data processing and applications is mostly covered in OneTrust. Nonetheless, this is not consistently updated, showing that regular maintenance is not a part of standard processes. While some guidelines exist around data strategy and architecture, these are still being implemented.

Notably, BE faces specific challenges around data minimization and anonymization, partly due to local government constraints. This shows that local regulatory context can directly impact maturity levels, and that this is an area in the framework that is currently difficult to improve for BE.

Data quality is an area for improvement: there are some controls, but no structured process for measuring and reporting issues, and inconsistencies exist especially for client data. For talent data, there is more control and clearer unique records, which shows that these data sets are at different maturity levels.

Overall, BE shows clear effort to improve its DG organization, with its maturity currently being around the minimum requirement of level 3. However, their weaknesses remain in reporting lines, data sharing standards, and data quality monitoring.

5.1.4 Maturity assessment results - CBS

The results of the maturity assessment for CBS are shown in Table 5.4.

DG Control	Level	Reasoning
Defining and documenting roles and responsibilities	2	Roles and responsibilities exist, but are not formalized, standardized or documented.
Integrating roles and responsibilities into organizational structure	4	There are is a DG team that covers five domains: data solutions, data literacy, access control, compliance, data quality
Reporting lines for DG	4.5	There are two main teams. They report to the CIO, but not to each other.
Defining master data	3	They are working on an extensive data catalog. MDM is on the roadmap, where it has been determined that Salesforce should be the master.
Documenting data preparation activities	3.5	The data quality control team aims to follow the complete journey of data to make sure that it is end-to-end. These are documented in google sheets, confluence and jira, but not in one central place. The goal is also to have this in one data catalog. The percentage of known data preparation activities is not explicitly discussed/known.
Developing data cleaning guidelines	N/A	Not explicitly discussed/known.
Developing data transformation/enrichment standards	N/A	Not explicitly discussed/known.
Classifying data according to GAIM (classification scheme)	4.5	The data is client-controlled. All data is pseudonymized/anonymized.
Developing data minimization, aggregation and anonymization standards	4.5	The data is client-controlled, so there is no personal data. All data is pseudonymized/anonymized.
Developing data sharing standards	N/A	Not explicitly discussed/known.
Developing third party data standards	N/A	Not explicitly discussed/known.
Documenting purpose of personal data processing	N/A	Not explicitly discussed/known.
Centrally documenting purpose of tools/applications	N/A	Not explicitly discussed/known.
Categorizing data according to taxonomy	N/A	They are working on an extensive data catalog. They are working on a business glossary that can combine the glossary of the client, CBS and GAIM.
Formulating completeness criteria	N/A	They have a designated quality control team that documents data quality in cyber processes.
Defining accuracy metrics	N/A	They have a designated quality control team that documents data quality in cyber processes.
Measuring freshness	N/A	They have a designated quality control team that documents data quality in cyber processes.
Measuring consistency	N/A	They have a designated quality control team that documents data quality in cyber processes.
Measuring uniqueness	N/A	They have a designated quality control team that documents data quality in cyber processes.
Knowing (test-)data distribution	2	They are working on this. Some things have already been done in SODA.
Measuring and reporting data quality	3.5	The two main teams report to the CIO, but not to each other. Data quality is being improved by data quality control. The complete journey is documented end-to-end.
Developing data strategy and architectural guidelines	4	Due to strict client requirements, they can adjust which tools they use and where data is stored. In addition, their data catalog enables harmonization across systems (including client).
Inventory data assets in a data catalog	3.5	They are working on an extensive data catalog which will include use cases, roles and responsibilities, data stewardship and ownership, data curators and a glossary.
Rating data assets on CIA	N/A	Not explicitly discussed/known.
Restricting access based on roles	4.5	Access control is managed and documented in CBS. Clients have requirements for access as well.

Periodically reviewing access rights	3.5	There is a specific DG team in charge of access control. Data access control will determine based on the audience who will get access.
Documenting data retention policy	4	Clients have requirements for data retention. This is contractually documented.
Storing data based on localization regulations	5	There are contractual agreements regarding data processing. Every customer has their own requirements, which CBS adapts to. Talent radar is a data warehouse. Where they can store data in whatever country the client wants.

Table 5.4: Maturity assessment for CBS

The maturity assessment for CBS shows a relatively high level of maturity, driven strongly by client requirements. CBS has a well-defined DG team that cover a broad range of DG dimensions; data solutions, literacy, quality, access control, and compliance. This setup ensures that responsibilities are integrated into the organizational structure, although formal documentation of individual roles is still under development.

Moreover, security is another strong point. All data is pseudonymized or anonymized, and data minimization is contractually required by clients, resulting in high maturity for privacy-related controls. Access management is also mature, with a dedicated DG team controlling access rights in line with client agreements. Data storage is an additional strong area, as it is both compliant with localization requirements and adjusts to each client's needs.

Although the extensive data catalog and business glossary are still in progress, they are designed to combine CBS, client, and GAIM glossaries into a standardized source. This shows clear harmonization efforts.

While it is unknown to us whether specific metrics for completeness, accuracy, and consistency exist, data quality is actively monitored by a designated team. End-to-end data journey documentation is also available and being improved upon. An improvement point is centralizing this, as current documentation is scattered over three platforms.

Overall, CBS shows a strong client-driven governance structure with a designated DG team and mature controls for access, privacy, and compliance. Areas for improvement are formal documentation of roles, consolidating data lineage documentation in a central platform, and finalizing the data catalog.

5.1.5 Maturity assessment results - DE

Table 5.5 includes the maturity assessment results for DE.

DG Control	**Level**	**Reasoning**
Defining and documenting roles and responsibilities	4	Roles and responsibilities are documented and defined for different processes.
Integrating roles and responsibilities into organizational structure	3.5	There is a master data management team, which is documented in the policy. Responsibilities lie with the managers of this team.
Reporting lines for DG	3.5	Reporting lines are documented in the policy.
Defining master data	3	There is a master data management team that reviews and validates every new account or change to master data. They ensure correct invoicing and payments, support credit checks and help in reducing duplicates. This is mainly the case for client data.
Documenting data preparation activities	4	The processing activities and data flow are known through OneTrust, BlueDolphin and interface descriptions. Blue Dolphin has 95% of DE's processes. Documenting data flow is still an ongoing process.
Developing data cleaning guidelines	N/A	Not explicitly discussed/known.
Developing data transformation/enrichment standards	N/A	Not explicitly discussed/known.

Classifying data according to GAIM (classification scheme)	3	Data in google drive is classified to a certain degree. There is an automated classification process in place. Some data is not yet classified.
Developing data minimization, aggregation and anonymization standards	3.5	In the Salesforce-based front-end and SAP, there's a full data protection and anonymizing concept. When a GDPR request is received from a talent or contact person, data can manually be anonymized by clicking a button. There are also automated routines for the entire database to anonymize data that is no longer needed or allowed to be held due to regulations. Note: non-production environments were not discussed.
Developing data sharing standards	3	There is a role-based access management system with different permission sets based on cost center, business unit, role and persona. Different systems have different access rights.
Developing third party data standards	N/A	Not explicitly discussed/known.
Documenting purpose of personal data processing	4	Documentation of the processing activities is as complete as reported by the departments. Checks are on an ad-hoc basis.
Centrally documenting purpose of tools/applications	4	Assets are documented in OneTrust but this list is not completely up-to-date.
Categorizing data according to taxonomy	2.5	The taxonomy is being implemented in the front-end system, which harmonizes job titles.
Formulating completeness criteria	3	The system itself has a wide range of completeness and consistency checks. There is also a global definition of "base data" for client data.
Defining accuracy metrics	N/A	MDM ensures more accurate data and less duplicates by validating data. MDM is mostly focused on client data.
Measuring freshness	N/A	Data in systems is mostly up-to-date. However, shadow data is used, which is not up-to-date.
Measuring consistency	N/A	The system has checks for consistency.
Measuring uniqueness	N/A	Duplicates exist for both client and talent data, but for talent data this is less the case. There are efforts to harmonize the duplicates.
Knowing (test-)data distribution	3	There are in general 2 data sets used for testing: (1) generic and specific data created to serve a test with particular data constellation to get a wide test-coverage mainly used in development systems, (2) data derived from a copy of the production data with (a) the same security level as the production data to test and compare to production results or (b) anonymized, scrambled or pseudonymized for a wider range of testers.
Measuring and reporting data quality	3	There are usage reports to show data quality status, without a link to business outcomes.
Developing data strategy and architectural guidelines	3.5	The data strategy is generally driven by the standard systems that are used. Data models in these systems are aligned, ensuring integration between them. Data strategy is driven by both technical requirements and business requirements. Changes to data or systems are managed through a standard change process applicable across all system modifications, so these systems are not adapted to whatever the organization wants.
Inventory data assets in a data catalog	3	List of assets in OneTrust is incomplete and not up-to-date.
Rating data assets on CIA	3	CIA rating in OneTrust is considered unreliable, because the information is based on the context of specific projects.
Restricting access based on roles	4	There are permission sets in place and an IAM system.
Periodically reviewing access rights	3	There is an IAM system in place.
Documenting data retention policy	3.5	Data retention measures are in place and reviewed regularly. Some parts are automated.

Storing data based on localization regulations	3.5	To the question whether this team knows where all the data is stored and processed, [data protection officer] answered that this information is documented in OneTrust assessments, and that [person 1] and [person 2] are very diligent in this.

Table 5.5: Maturity assessment for DE

The maturity assessment for DE shows a moderately high level of maturity, with a clear foundation of documented roles, responsibilities, and reporting lines in policies. DE has a master data management (MDM) team, mainly focused on client data, who ensure accuracy, reduce duplicates, and support invoicing and credit checks.

Process documentation is one of DE's other strengths: tools like OneTrust and BlueDolphin cover or are planned to cover nearly all processing activities and data flows, although some updates are still ongoing. Data classification and taxonomy implementation are partially in place, with automated classification for some data and a taxonomy system being rolled out to harmonize job titles. Both of these can be extended to more data sets.

There are strong controls for data protection and anonymization: Salesforce and SAP have both manual and automated routines for GDPR compliance, access rights are well-managed through an IAM system, and data retention is documented and partly automated.

Despite the high maturity levels on the mentioned practices, there remain some gaps. There are data quality measures like completeness and consistency built into systems and a MDM team. However, clear KPIs are missing for talent data. Shadow data and duplicates are also an issue, particularly for client data. Lastly, the data catalog in OneTrust is incomplete and CIA ratings for assets are unreliable due to project context-specific measurements.

Overall, DE has strong foundations with policies, effective MDM for client data, and good privacy and access controls. They are improving in areas where attention is still needed, such as the taxonomy and data uniqueness. Nevertheless, there remain quite some gaps between client data and talent data in different areas. Controls in place for either client or talent can be extended to the other data set. Other improvement areas are regular maintenance of OneTrust documentation and linking data quality reports to business outcomes.

5.1.6 General observations

Across the five maturity assessments, several patterns stand out. Many markets show relatively strong maturity in security-related controls, such as access controls, data retention, and anonymization.

In contrast, defining and documenting clear roles and responsibilities remains a common challenge. While responsibilities are often carried out in practice, formal appointments of data owners, clear reporting lines, and consistent role documentation are frequently lacking or only partially implemented. Similarly, the maintenance of tools like OneTrust and data catalogs is not yet a fully established routine in many markets, as information is often incomplete or outdated.

Another recurring area for improvement is data quality. Although most markets have some controls in place, such as MDM teams, ad hoc cleansing, or dashboards, there is often no structured approach for consistently measuring and reporting on data quality issues. System integration challenges, and as a result duplicated data, also remain frequent obstacles.

Among the markets, CBS and DE generally demonstrate higher maturity levels. CBS benefits from strong client requirements that enforce robust controls for access, privacy, and compliance, with a dedicated DG team covering the most important DG topics. DE shows a solid foundation with policies, a MDM focused on client data, strong privacy measures, and good process documentation. The NL, BE, and the UK score lower in most areas. Nevetheless, there remains room for improvement in all markets.

The last observation is that maturity levels regarding data quality criteria or measurements could generally not be determined in markets. As explained in Section 4.2, this was largely due to the AI DG Framework's criteria, which required measured percentages regarding a data set. This suggests that these criteria should be adjusted to fit what is possible in the markets now.

5.2 Business value analysis results

Thematic analysis of the interviews with stakeholders resulted in three types of codes: benefits from DG practices as indicated by stakeholders (perceived business value), market-specific data challenges, and market-specific best practices. The following subsections discuss the results regarding perceived business value and data challenges. Best practices are presented in the next section.

5.2.1 Perceived business value

As can be observed in Table 5.6, we identified 16 distinct examples where interviewees explicitly indicated how other processes, business outcomes, or projects are influenced by the presence or lack of certain DG practices. The identified codes are grouped by the DG category where the underlying DG practice belongs to, based on the six categories found in Chapter 3. In other words, the codes are classified under the area that was mentioned as being effective or necessary in enabling a certain outcome. A new category "Data architecture" was created to address practices regarding the organization and structure of data systems, as these did not fit in the existing categories. Below, we discuss the findings for each category.

Category	**Code**	**Markets**
Data quality (8)	Complete data is needed for effective client communication	NL
	Data cleaning helps data entry practices	UK
	Data of poor quality is more costly to govern	UK
	High quality data is needed for daily operations	DE
	High quality data is needed for ML, DS and AI projects	UK
	High quality data leads to better matches	NL
	Incorrect/incomplete data entry lead to downstream problems	NL, UK
	Master data management (data validation, enrichment, duplicate management) leads to higher data quality	DE
Data architecture (3)	Data fragmentation makes it hard to make data-driven decisions	UK
	Standardization of reporting reduces costs and increases customer satisfaction	BE
	Standardization, aggregation and migration of data improves data quality	NL
Metadata management (2)	Clear information and reporting needs enable effective migration	NL
	Clear documentation creates client trust	CBS
Roles & responsibilities (1)	Clear ownership enables effective migration	NL
Security & privacy (1)	Data protection reduces risks and costs	DE
Training & awareness (1)	Incorrect/incomplete data entry lead to downstream problems	NL, UK

Table 5.6: Business value from DG practices as indicated by the markets

Data quality was the largest category, with eight codes highlighting its importance. Markets indicated that complete and accurate data is essential for daily operations, effective client communication, and delivering better matches. Furthermore, high-quality data is a prerequisite for DS, ML and AI initiatives. On the other hand, poor data quality was described as increasing governance costs and causing downstream problems. DE reported that MDM activities like validation, enrichment, and duplicate handling, which are data quality activities, directly improved data quality. In the UK, cleaning the data led to better data entry practices.

The findings suggest that good data architecture is an important enabler for good decision-making, customer satisfaction, and data quality. Fragmented data was said to make data-driven decision-making difficult, while standardizing reporting was seen to reduce costs and improve customer satisfaction. Migrating and aggregating data were linked to higher data quality.

The other categories were less discussed. Nevertheless, they revealed insightful benefits of DG, such as that having clear documentation and reporting structures helps coordinate migrations effectively and

demonstrates transparency to clients.

Furthermore, clear roles and responsibilities were mentioned once as another important condition for smooth migrations, which shows how ownership structures support technical improvements.

Moreover, security and privacy practices were linked to reducing risks and unnecessary costs as they ensure that data is properly protected.

Lastly, training and awareness appeared in the context of data entry: when people understand correct practices, they help prevent mistakes that would otherwise create issues further down the line, which would have required more costs and resources.

5.2.2 Data challenges

Table 5.7 illustrates 23 distinct data challenges and needs, as indicated by interviewees from all markets. Similar to the previous subsection, these codes are grouped based on the six categories data quality, data architecture, roles and responsibilities, training and awareness, metadata management, and security and privacy. The findings suggest various needs and challenges, which are explored below.

Category	**Code**	**Markets**
Data quality (8)	Automated controls to improve and enrich data	BE, DE, UK
	Controls for monitoring data quality issues	BE, DE, NL
	Data freshness depends on talents	BE, NL
	Data quality KPIs to measure base level of quality	DE
	Duplicate data from migrations and system limitations	DE
	Inconsistent data across systems	BE, NL, UK
	Lack of data quality controls on data entry	BE, NL, UK
	Use of shadow data	BE, CBS, NL, UK
Data architecture (4)	Aligning data capture points	UK
	Data fragmentation over multiple systems	BE, NL, UK
	Standardization of reporting	NL
	Use of shadow applications	UK
Roles & responsibilities (4)	Data management team needed	BE
	Defining and appointing data owners	BE, DE, NL
	Formalizing and standardizing roles and responsibilities	CBS
	Lack of resources to fulfill responsibility	NL
Training & awareness (4)	Data entry focuses on immediate operational needs	NL
	Incentivizing good data entry to improve data quality	NL, UK
	Raising awareness around proper data entry	DE, NL, UK
	Raising awareness around the purpose and value of data	DE, NL, UK
Metadata management (2)	Insight into data lineage	BE
	Localized version of data model	CBS, DE
Security & privacy (1)	Historically granted access is unknown	NL

Table 5.7: Data challenges and needs as indicated by the markets

The codes show that data quality challenges were by far the most frequently mentioned, appearing in every market. Commonly shared issues include inconsistent data across systems, lack of controls at the point of data entry, and the use of shadow data. These problems were raised in multiple countries (BE, NL, UK, CBS), which suggests that they are not market-specific but rather structural. Similarly, several countries pointed out the need for automated controls for improving or enriching data and the need for monitoring data quality.

Data architecture problems were also seen across multiple markets. Fragmentation of data over multiple systems was a recurring point in BE, the NL, and the UK. Shadow applications and misaligned capture points were highlighted mainly in the UK, while standardization of reporting came up in the NL.

Roles and responsibilities challenges were described in nearly all regions too. The need for clear data owners was noted in BE, DE, and the NL, while a lack of resources to carry out responsibilities was mentioned specifically in the NL. CBS also raised the need to formalize and standardize roles more broadly.

Training and awareness issues appeared in four countries, with the NL standing out for multiple concerns around poor data entry practices and the immediate focus on operational needs. The UK, DE, and the NL all pointed to the need to raise awareness of the purpose and value of data.

Metadata management came up less often. CBS and DE both mentioned the need for a localized data model, while insight into data lineage in BE.

Finally, security and privacy challenges were only mentioned in the NL, where historically granted access rights are not always known.

Overall, the findings show that some challenges, especially around data quality and architecture, are experienced across multiple countries. Other challenges, like resource constraints, data lineage, or access issues, are more specific to certain local contexts.

5.3 Gap analysis results

This section presents the results from the gap analysis. We first provide an overview of the best practices identified from the five markets in scope. These best practices are then consolidated with the practices from literature (Section 3.3), allowing us to make a comparison with the AI DG Framework.

5.3.1 Market best practices

Table 5.8 presents the best practices identified through thematic analysis of the interviews. These practices are grouped by DG category, following the categories used in the previous subsection.

Category	**Code**	**Markets**
Metadata management (8)	Automated classification of documents	DE
	Data catalog (use cases, roles and responsibilities, glossary)	CBS
	Documenting data architecture	UK
	Documenting data lineage	DE
	Localized version of data model	BE, UK
	Localized version of taxonomy	DE, UK
	Process catalog	BE, DE
	Using data build tool to define source, transformations and destination	NL
Security & privacy (8)	Ad hoc access for specific requests	BE, DE, NL
	Automated anonymization based on regulation and retention	DE
	Blocking data exports from system	BE, UK
	Built-in anonymization options	DE
	Data retention policy	DE
	Data storage customization	CBS
	Role-based access rights	BE, DE, NL, UK
	Using only pseudonymized data	CBS
Data quality (6)	Automated data quality checks	BE, CBS, DE, NL
	Automated duplicate handling	BE
	Data cleaning	UK
	Mandating necessary fields for data entry	DE, UK
	Tools to enrich data to increase its usability	UK
	Tools to merge multiple versions of data	UK
Training & awareness (5)	Flashcards to help people understand DG principles	CBS

	Formalizing feedback regarding data problems and mistakes	UK
	Onboarding and (mandatory) training	CBS, DE, NL, UK
	Reminders to keep data up-to-date	NL
	Understanding undesired data entry behavior	UK
Roles & responsibilities (5)	DG team covering different domains: data solutions, data literacy, access control, compliance, data quality	CBS
	Defining and documenting roles and responsibilities	CBS, DE, UK
	Executive data users representing users from departments	NL
	Master data management team (data validation, enrichment, duplicate management)	DE
	Steering committee for information security	CBS
Data architecture (4)	Migrating data from multiple/local systems into one platform	NL, UK
	Requiring critical processing to go through one system	UK
	Standardization of reporting	BE, CBS
	Standardized change process for changes to data	DE

Table 5.8: Best practices identified from the markets through thematic analysis

In total, 36 distinct practices were identified across six DG categories. Metadata management and security and privacy were areas where the most best practices were found, with eight practices each, contributed by all markets. This demonstrates that the markets are implementing good ideas for organizing, documenting, protecting, and controlling access. It also suggests that overall, these areas were the most developed across markets. For example, three countries have adapted the organization's data model and taxonomy, and adapted them to fit local needs. CBS is in the process of building an extensive data catalog containing use cases, roles and responsibilities, data stewardship and ownership, data curators, and a glossary. While other markets, such as BE, still face challenges in gaining insight into the data lineage and process flows, DE appears to have these areas well-managed, as indicated by their best practices on data lineage and process catalog.

In security and privacy, markets described a variety of measures. These range from technical controls, like blocking exports or using built-in automated anonymization, to policy-driven approaches such as data retention guidelines. DE seems to be especially well-organized in this area. Nevertheless, access rights, both ad hoc and role-based, are established in all markets.

For data quality, we identified six examples touching on automation, enrichment, and cleaning activities. Automated data quality checks were implemented in most markets (UK, BE, CBS, NL). For example, BE mentioned the following: "*There are various data controls, including validation of entered numbers, start and end dates based on certificates, and automated document recognition.*" It is notable that majority of the remaining best practices were identified in the UK. This suggests that this market is proactive in improving their data quality and that they are quite mature in this area.

Although this topic is not part of the AI DG Framework yet, efforts to strengthen training and awareness around data were observed in five practices. Most markets reported to have an onboarding and mandatory training procedure. As the UK states: "Annual, mandatory compliance training covering data quality and GDPR is implemented for all employees, with completion reported." Some markets, like CBS and the UK, also shared ideas such as using flashcards and feedback loops to build data literacy and establish better data management practices.

In the area of roles and responsibilities, five practices showed how markets structure governance teams and clarify accountability. Where CBS has established comprehensive DG teams and steering committees, and the NL has designated executive data users to represent departmental data needs, DE stands out as the only market with a dedicated master data management team focused on ensuring data quality.

Lastly, in data architecture there were four practices highlighting standardization of reporting, system migration, and consistent change processes. CBS, for instance, has a standard dashboard used for reporting that is customizable to some degree: "*There is a standard dashboard portfolio. For each customer you can then decide which components are included. For the customer it seems specialized for them, but actually it is off-the-shelf.*"

Overall, these insights show that while the specific practices vary to fit local contexts, all markets have found similar ways to address common DG needs. Some best practices, such as role-based access rights and automated data quality checks, were reported by multiple markets, while many other practices were more market-specific, emerging from only one country each.

5.3.2 Framework gaps

An overview of the gap analysis between practices from literature, markets, and the AI DG Framework is presented in Table 5.9. Below, we describe differences, similarities and gaps for each DG category according to the categories used in Subsection 5.3.1.

Literature	**Markets**	**Framework**	**Alignment**
Metadata strategy	Documenting data architecture	Developing data strategy and architectural guidelines; Inventory data assets in a data catalog; Defining and documenting roles and responsibilities	Partial
Maintaining data catalogs	Data catalog; Process catalog	Inventory data assets in a data catalog	Full
Data lineage tracking	Documenting data lineage; Using data build tool to define source, transformations and destination	Documenting data preparation activities	Partial
Data taxonomy	Localized version of taxonomy	Categorizing data according to taxonomy	Full
Establishing data quality standards and metrics	Automated data quality checks; Mandating necessary fields for data entry	Formulating completeness criteria; Defining accuracy metrics; Measuring freshness; Measuring consistency; Measuring uniqueness	Partial
Continuous profiling and monitoring	Automated duplicate handling; Tools to merge multiple versions of data	Measuring and reporting data quality	Partial
Data cleansing	Data cleaning	Developing data cleaning guidelines	Partial
Reporting	Formalizing feedback regarding data problems and mistakes	Measuring and reporting data quality	Partial
Implementing Explainable AI techniques	-	-	Gap
Clear documentation	Documenting data architecture	Documenting purpose of personal data processing; Centrally documenting purpose of tools/applications: Developing data cleaning guidelines; Developing data transformation/enrichment standards; Knowing (test-)data distribution	Partial
Continuous monitoring and evaluation	-	-	Gap
Risk assessments	-	Rating data assets on CIA	Partial
Data storage based on business value	Data storage customization	Storing data based on localization regulations	Partial
Access controls	Role-based access rights; Ad hoc access for specific requests	Restricting access based on roles; Periodically reviewing access rights	Full
Encryption	Using only pseudonymized data; Automated anonymization; Built-in anonymization options	Developing data minimization, aggregation and anonymization standards	Full
Training and awareness	Onboarding and (mandatory) training; Flashcards; Reminders; Understanding undesired data entry behavior	-	Gap
DG council	DG teams; Steering committee for information security	Defining and documenting roles and responsibilities	Partial

Business owners of data assets	Executive data users representing users from departments	Defining and documenting roles and responsibilities	Partial
Data custodians	Master data management team	Defining and documenting roles and responsibilities	Partial
Data stewards	DG teams	Defining and documenting roles and responsibilities	Partial
Data users	Executive data users	Defining and documenting roles and responsibilities; Reporting lines for DG	Partial
-	Migrating data from multiple/local systems into one platform	Developing data strategy and architectural guidelines	Partial
-	Requiring critical processing to go through one system	Developing data strategy and architectural guidelines	Partial
-	Standardization of reporting	Developing data strategy and architectural guidelines	Partial
-	Standardized change process for changes to data	Developing data strategy and architectural guidelines	Partial

Table 5.9: Gap analysis between literature, market, and framework best practices

Metadata management

In the literature, a metadata strategy is described as a structured plan covering objectives, roles, standards, and technologies for managing metadata across systems. In practice, the markets generally do not have such a comprehensive plan but instead show related, separate practices such as documenting data architecture. Similarly, the AI DG Framework does not address a single control for an overarching metadata strategy but covers the relevant elements across multiple controls. This means that while parts of a metadata strategy are addressed, they are not necessarily required to be under one plan. The framework does address "data strategy and architectural guidelines", but this control is largely focused on data architecture. Therefore, this can be seen as both an alignment and a gap. An improvement could be to extend the existing control or add a new control to require a single metadata strategy describing roles, objectives, standards, and tools. This will help ensure alignment between these elements.

The literature also emphasizes searchable data catalogs with datasets along with descriptions to improve visibility and control. This is an area where the literature, markets, and the framework are well-aligned on. Several markets are actively working on data and process catalogs. For example, CBS has an extensive data catalog describing use cases, roles and responsibilities and a glossary. The framework includes the practice of inventorying data assets in a data catalog along with metadata, descriptions, and other relevant information.

Data lineage tracking follows data through all steps of transformations in systems. It is highlighted in the literature as a key to transparency and data quality. In the markets, data lineage is documented through BlueDolphin and a data build tool. The framework partially covers this through the control of documenting data preparation activities. While data cleaning and transformation are also addressed in separate controls, these controls focus more on the development and standardization of cleaning and transformation guidelines. For this reason, it is suggested to extend the preparation control in the framework to include end-to-end data flow. BlueDolphin and a data build tool can be added as examples to help other markets realize this.

Finally, the literature calls for taxonomies to standardize classification. This area shows full alignment, as the framework includes categorizing data according to a centrally (HR firm) defined taxonomy. The markets UK and DE further strengthen this practice by adapting the taxonomy to fit local needs. This suggests that the framework could include a recommendation to apply a localized version of the taxonomy to help the other markets adapt the taxonomy more effectively.

Data quality

Both the literature and the AI DG Framework address the importance of establishing clear data quality standards and metrics. The literature strongly emphasizes defining data quality dimensions such as completeness, accuracy, consistency, freshness, and uniqueness, with relevant KPIs that are fit-for-purpose

rather than focusing merely on collecting large volumes of data. The framework reflects this by including controls for defining and measuring these dimensions. Reporting measured data quality is also part of the literature, as well as the framework, which shows that they are well-aligned on these two practices.

Across the markets, similar practices are visible, but they also reveal some practical aspects that are not fully covered in the framework. For example, several markets reported the use of automated checks and duplicate handling tools. While these align with the framework's guidance on defining data quality metrics, the framework does not directly require these practices. Another example is the feedback loops used in the UK, which go a step further than standard data quality reporting. Instead of summarizing data quality metrics, they trace mistakes back to their source and feed this information to the person responsible. This allows for accountability and targeted awareness surrounding data problems. Combining both practices would strengthen the control by linking measurement with concrete actions for improvement.

Data cleaning is also addressed in a different way in the three sources. In the literature section and the markets, the focus is on the practice itself and implementing automated cleansing processes, although it is still an ad hoc practice in the markets rather than a regular process. However, the framework only mentions the development of cleaning guidelines, without providing specific recommendations on the contents. For this reason, this control in the framework could be adjusted to combine these approaches, addressing guidelines as well as processes involving regular data cleaning.

In conclusion, the gap here is not so much a lack of coverage in the framework, but rather that some of these practical approaches are not explicitly defined as controls. Therefore, there is room to adjust or expand the existing controls to reflect these additional measures, so that practices are not only aligned with theoretical requirements (e.g. standards and guidelines), but also supported by practical methods that have already proved effective.

Transparency and explainability

Transparency and explainability are important principles in the context of AI (data) governance. While traditional DG frameworks tend to focus on the other DG categories, AIG also highlights the complexity and opacity of AI systems. Methods such as Explainable AI (XAI) techniques, clear documentation of decision logic, and continuous evaluation are commonly used to address this challenge.

In the AI DG framework, however, these aspects are not yet embedded as controls. The focus remains on ensuring basic transparency through documentation of processes, data preparation activities, and reporting. Some related controls, such as knowing the data distribution to ensure fairness in the use for AI and documenting the purpose of data processing and tools, overlap partially with what the literature suggests. Nonetheless, explicit explainability techniques for AI models are not prescribed.

At market level, no specific best practices were identified that directly address transparency and explainability of data, algorithms or AI-driven decision processes. There are two main reasons for this. Firstly, as the topic is not part of the current AI DG Framework, it was not discussed. Secondly, current priorities in markets are focused on establishing or strengthening the more foundational practices of DG, such as data protection and data quality, rather than on AIG practices.

Given the HR firm's current stage of DG maturity and focus, we do not suggest to add highly specialized AI explainability requirements as separate controls. Instead, it may be more practical to first ensure that the aforementioned basic transparency measures are consistently applied.

Security and privacy

The security and privacy aspects of DG ensure that data remains reliable and protected. In the literature, this includes measures like risk assessments, (role-based) access controls, strategic data storage practices, and technical safeguards such as encryption.

In the framework, risk assessments are partially addressed: data assets must be rated on confidentiality, integrity, and availability (CIA). However, this is a rating on a 1-3 scale and limited to asset-level classification. It does not fully reflect the broader risk management and protection focus described in the literature. In the markets, CIA ratings are applied to assets, but no additional risk assessment practices were identified.

Data storage is partially addressed in the framework and in practice. The framework expects data to be stored according to localization requirements, which most markets do well. However, the literature highlights the additional benefit of allocating storage resources based on business value of data. One market (CBS) applies a related practice: customization of data storage to fit clients' requirements.

The three sources are well-aligned on access controls. Both the framework and the markets address role-based restrictions and periodic reviews to ensure only necessary access is granted. In practice, some markets also have a process for granting ad hoc access for specific projects, a practice which could be added to the framework, as it adds flexibility.

Encryption appears to be sufficiently covered in the framework, which requires anonymization and minimization standards and that personal data usage is minimized. In practice, certain markets, such as DE, go a step further by using built-in anonymization features and automating the pseudonymization of sensitive data where possible. This is a useful best practice to add to the existing control in the framework.

Overall, while the core security and privacy measures are largely aligned, the partial coverage of risk assessments highlights a gap. Strengthening this area by expanding risk assessments beyond asset ratings can help identify needs for better protection measures.

Training and awareness

Training and awareness are recognized in the literature as essential for embedding DG into daily operations and building a culture of accountability and good data practices. Despite this, the framework currently does not include any explicit controls for training and awareness. However, the markets show clear recognition of its importance. Examples include onboarding sessions, mandatory training, reminders, flashcards, and initiatives to understand undesired data entry behavior.

Given the proven impact of training on DG success and the practical initiatives already present in some markets, this is an area where the framework could be expanded. Integrating training and awareness as a requirement would help ensure that everyone involved has the right understanding of data practices and contributes actively to improving data quality.

Roles and responsibilities

In the literature, roles such as a DG council, business data owners, data custodians, data stewards, and data users are commonly highlighted. These roles cover the strategic, tactical, and operational levels.

In practice, the markets reflect this layered approach by having specific teams and committees, such as DG teams (CBS), information security steering committees (CBS), and designated executive data users (NL). These roles fill similar functions to those described in the literature. Roles such as business data owners and data stewards are often not formally appointed in the markets. Instead, certain individuals can be identified as such, as their functions partly cover similar responsibilities.

The framework also recognizes the importance of roles and responsibilities, and defines the roles business data Owner and data owner (see Appendix A). However, these roles are not formalized as standalone controls within the framework. While this does establish some guidelines, it leaves room for improvement: a clearer and more comprehensive role structure could help clarify who are responsible or accountable for decision-making, execution, and issue reporting.

Overall, the literature, markets, and framework align on the need for clearly defined roles, but they differ in how explicitly and completely these roles are described and embedded. To close this gap, the framework could expand its role definitions and link them more directly to controls. This will provide an outline that the markets can adapt by appointing similar roles suited to their own organization.

Data architecture

Data architecture was not strongly emphasized in the reviewed literature. In practice, however, several markets have implemented more concrete measures related to data architecture. Examples include migrating data from multiple systems into a single platform, requiring critical processing through one system, standardizing reporting, and having a standardized change process for data. While the framework does cover the development of architectural guidelines, it does not fully reflect these practical best

practices. Therefore, adding these measures from the markets could strengthen the framework and help ensure consistency and integration across systems, which will improve data quality.

5.4 Model demonstration

This section demonstrates the proposed prioritization model using the results and findings obtained in the previous sections. Table 5.10 shows a demonstration for the UK.

There are two limitations, as a result of which we use the model with a simplified formula and an assumption.

First, as information about the complexity of improving controls is unavailable, we only consider the business value factors perceived business value (PBV) and perceived challenge (intensity). Therefore, we apply the following simplified version of the proposed formula, leaving out the complexity variable (C):

$$\text{PQ} = \text{Business value} - \text{Risk} = \frac{(PBV + PCI)}{2} - ML$$

Second, there are controls in the assessment results to which we did not assign a level, as explained in Section 4.2.1 and shown in 5.1. For these controls, we use the average maturity level calculated for the rest of the controls in the same category, following the categories of the AI DG Framework. This approach was chosen because it better reflects the strengths and weaknesses associated with a control than using the total average.

Following the criteria outlined for each variable in Section 4.3, each DG control is scored on PBV and PCI using coded insights from the interviews (Table 5.6 and Table 5.7). The maturity levels (ML) are taken from the maturity assessment (Table 5.1). The priority quotient (PQ) is calculated by subtracting the maturity level from the business value. To highlight the most important controls, Table 5.10 is sorted in descending order of PQ.

DG Control	**PBV**	**PCI**	**BV**	**ML**	**PQ**
Measuring consistency	5	5	5	*3.5*	1.5
Measuring uniqueness	5	5	5	*3.5*	1.5
Measuring and reporting data quality	5	4	4.5	3.5	1
Measuring freshness	5	3	4	*3.5*	0.5
Formulating completeness criteria	3	4	3.5	*3.5*	0
Defining accuracy metrics	5	4	4.5	*3.5*	0
Developing data transformation/enrichment standards	2	4	3	*3*	0
Developing data strategy and architectural guidelines	3	3	3	*3*	0
Developing data cleaning guidelines	4	3	3.5	2.5	-1
Defining master data	2	2	2	3	-1
Reporting lines for DG	1	1	1	2	-1
Documenting data preparation activities	2	1	1.5	*3*	-1.5
Developing data minimization, aggregation and anonymization standards	2	1	1.5	*3*	-1.5
Knowing (test-)data distribution	1	1	2	*3.5*	-1.5
Inventory data assets in a data catalog	2	1	1.5	3	-1.5
Rating data assets on CIA	2	1	1.5	3	-1.5
Classifying data according to GAIM (classification scheme)	1	1	1	*3*	-2
Developing third party data standards	1	1	1	*3*	-2
Documenting purpose of personal data processing	2	1	1.5	3.5	-2
Centrally documenting purpose of tools/applications	2	1	1.5	3.5	-2
Periodically reviewing access rights	2	1	1.5	3.5	-2
Storing data based on localization regulations	2	1	1.5	*3.6*	-2.1
Developing data sharing standards	1	1	1	3.5	-2.5

Categorizing data according to taxonomy	1	1	1	*3.5*	-2.5
Restricting access based on roles	2	1	1.5	4	-2.5
Documenting data retention policy	2	1	1.5	4	-2.5
Integrating roles and responsibilities into organizational structure	2	1	1.5	4	-2.5
Defining and documenting roles and responsibilities	1	1	1	4	-3

Table 5.10: Demonstration of prioritization model for the UK market

The demonstration of the prioritization model for the UK shows that controls related to data quality measurement and metrics stand out with high business value scores and low maturity levels, resulting in high priority for improvement. This suggests that while stakeholders recognize the value of high-quality data, as seen in the PBV scores, formal measurement and monitoring processes are not yet mature.

In contrast, access rights and roles show relatively high maturity and low priority, which indicates that foundational governance structures for accountability and permission management are relatively well-established and not an obstacle for the business.

Based on the highest PQs, the three recommended areas for improvement are:

- Measuring data quality (completeness, accuracy, freshness, consistency, uniqueness), formulating measurement criteria, and reporting on this.
- Developing data transformation/enrichment standards.
- Developing data strategy and architectural guidelines.

Overall, the prioritization model shows that the UK should focus its next improvement steps primarily on defining and standardizing data quality metrics and controls, developing data transformation/enrichment standards, and developing a data strategy and architectural guidelines.

Chapter 6

Discussion

This chapter will reflect on the findings from Chapter 5, and answer the four sub-questions formulated in Chapter 2. Limitations of this study are also included below in the discussions of the various sub-questions.

6.1 Reflection on maturity assessment

To answer RQ.1 **How can the maturity of AI-related data governance practices be assessed against regulatory requirements?**, we demonstrated a combined approach of structured document review, questionnaires, stakeholder interviews, and a mapping of gathered information to maturity levels described in the AI DG Framework. This provided us valuable insights into the current state of DG maturity across different markets of the HR firm.

Most markets demonstrate relatively strong maturity in security-related controls, which reflects that regulatory drivers and privacy requirements have successfully pushed these areas higher on the agenda.

In contrast, controls for clearly defining and documenting roles and responsibilities remain underdeveloped, with many markets relying on informal arrangements rather than formal appointments and up-to-date documentation. Similarly, tools such as data catalogs and registers are often incomplete or not consistently maintained, indicating a lack of formalized processes.

Another consistent challenge is the lack of a structured approach for measuring and monitoring data quality. Although good practices are carried out ad hoc, such as cleaning and integrating platforms, systematic measurement and reporting are rare, and integration challenges and duplicate data persist.

Among the assessed markets, CBS and DE generally show higher maturity, supported by stricter client requirements, dedicated DG/MDM teams, and stronger privacy and access controls. The NL, BE, and the UK score lower in most areas but share similar improvement needs.

Reflecting on the assessment process, the combination of document reviews and interviews was valuable for uncovering market-specific contexts and highlighting both maturity levels and practical improvement needs. Using gathered information and mapping them to multiple relevant controls ensured that each DG control could be assessed independently and with consistent supporting evidence.

However, several limitations of the process should be acknowledged. First, not all markets completed the questionnaire on time or in full, which reduced its usefulness and the depth of interviews. In addition, because the interviews served a dual purpose (assessing maturity and identifying practical challenges and good practices to support the improvement plan and framework refinement) the questions asked did not always align perfectly with every DG control's specific maturity criteria. This meant that some controls, especially those less relevant to immediate daily operations, could not be fully assessed.

In hindsight, this raises the question of whether a different approach, such as having each market self-assess all controls using the framework first, could have yielded more complete and comparable maturity scores. Interviews could then be used only to validate and contextualize those scores, and to inquire about practical needs and challenges.

Another challenge in our approach was that some maturity level descriptions were too advanced or specific for the current state of DG, for example requiring concrete measurement percentages without a supporting measurement process in place. This made it difficult to assign precise scores and underlines the need for frameworks to be realistic and flexible enough to capture both early-stage and more advanced practices.

Overall, despite these limitations, the method provided a solid starting point to assess DG maturity against regulatory requirements and to identify areas where practical improvements and framework adjustments are needed.

6.2 Reflection on business value analysis

To address RQ.2 **How can the business value of AI-related data governance practices be measured?**, we combined an inductive thematic coding approach of interviews with a focus on perceived benefits, recurring challenges, and best practices.

The results show that data quality is the most important business value driver, as it was linked to operational efficiency, client satisfaction, and a requirement for AI initiatives. Data architecture also plays a key role in improving data quality and decision-making. Other enablers include clear roles and responsibilities, good documentation, security and privacy, and training and awareness to prevent costly mistakes.

Importantly, these topics that are considered as the most important DG aspects, also emerged as the main challenges: poor data quality, fragmented systems, unclear ownership, and insufficient training and awareness were common pain points across markets. This reinforces that these areas are both the biggest value opportunities and the biggest gaps that need to be addressed.

This confirms that combining a more compliance-focused maturity assessment with open questions about challenges and benefits reveals meaningful insights about how DG practices translate into tangible business benefits and costs.

One challenge of this approach was that interviewees did not always know how to directly answer broad questions about which practices deliver the most value or which challenges are most critical. Instead, we noticed that they were better able to describe concrete examples when asked about a specific practice. This highlights that more targeted, practice-specific questions are more efficient in uncovering clear links between DG activities and business outcomes.

Overall, the qualitative coding approach worked well to gain insight into this connection. However, future investigations could approach this quantitatively in order to strengthen the link to measurable KPIs, such as error rates, duplicate records, or time saved. This will provide a more objective way to estimate how severe certain challenges are and measure the business value of individual practices more objectively.

6.3 Framework recommendations

To investigate RQ.3 **How can the current AI DG Framework be refined with best practices identified from literature and the maturity assessment?**, we compared the practices from literature, the five markets, and the AI DG Framework. The comparison shows that the framework already covers many important elements that are emphasized as good practice, such as clear data quality standards, metadata management elements, and access controls. This indicates that its foundation is largely relevant and usable in practice. However, the analysis also highlights that the controls could be made more complete and actionable by integrating practical best practices visible in the markets and by aligning with best practices from literature.

Below are the improvement points suggested to the HR firm's HR AI DG Framework, organized according to the different observations described in Section 5.3:

First, some topics are only partially covered or fragmented across multiple controls:

- Metadata strategy: develop a single, overarching metadata strategy describing objectives, roles, standards, and technologies to bring together elements that are now spread across controls.

- Data lineage: extend the control on data preparation to include end-to-end data flow documentation. Tools like BlueDolphin and data build tools can be added as examples.
- Data cleaning: expand the existing guidelines to include concrete processes for regular cleaning, supported where possible by automation.
- Risk assessments: extend beyond basic CIA ratings to cover broader risk assessments in order to take protective measures where needed.

Second, several practical best practices already applied by markets could be included in the framework as concrete examples or extensions of existing controls:

- Data quality: add practices such as implementing automated checks and duplicate handling.
- Anonymization and pseudonymization: incorporate best practices like built-in anonymization features and automated pseudonymization, as used in DE.
- Ad hoc access controls: add flexibility to role-based access by including processes for granting ad hoc access for specific projects.
- Localized taxonomy: recommend that centrally defined taxonomies can be adapted to fit local contexts, following the example of UK and DE.
- Data architecture: include best practices like migrating data to single platforms, requiring critical processes to run through a unified system, standardizing reporting, and implementing standardized change processes.

Third, some important aspects are not yet formalized or included in the AI DG Framework:

- Training and awareness: introduce a dedicated control for training and awareness measures to embed DG practices into daily work and build a culture of accountability. Include examples from the markets, such as onboarding, reminders, and regular refresher sessions. Additionally, combine data quality reporting with concrete feedback loops that trace errors back to their source and raise awareness.
- Roles and responsibilities: expand the defined roles in the framework with other roles described in the literature (DG council, data owners, data custodians, data stewards, data users). Make these roles part of controls so that markets can appoint people to similar functions adapted to their local DG organization.

Finally, our maturity assessment process, highlighted that controls with level descriptions that required a measured implementation percentage (25%, 50%, 75%, 100%) were too difficult to measure in practice. Hence, this could be improved by revising the maturity level descriptions such that they reflect practical and helpful requirements, following the maturity phases described in the framework: ad hoc, managed, defined, measured, and optimized (as explained in Appendix A).

Overall, the framework's focus remains on ensuring foundational DG controls are in place, but these could be strengthened by the proposed changes. Doing so would not only improve completeness but also make the framework more usable and relevant for data needs in local markets.

6.4 Reflection on prioritization model

To explore RQ.4 **Which areas should be prioritized when implementing an AI-related data governance framework to maintain a risk-to-benefit balance?**, we proposed a prioritization model that compares the business value of DG practices with the remaining risk.

A strength of this model is that it translates qualitative findings into priorities considering multiple factors. By combining evidence from the maturity assessment, perceived business value, and data challenges, it supports more informed decision-making about where to invest effort first.

Another benefit is flexibility, as the model's main variables risk and business value can be defined and measured in different ways depending on an organization's context and completeness of data. For example, the business value scores could be derived from performance metrics, as long as they can be translated into a comparable scale.

However, the model also has limitations. The current value scoring method depends on subjective interpretation of interview data. Furthermore, using maturity as an indicator for risk may not fully reflect all possible risk aspects, such as financial consequences or reputational harm. Lastly, the model's usefulness depends on having clear differences between DG controls. If there are no clear perceived business value enablers or challenges identified, and maturity levels are relatively uniform across all controls, the model's ability to derive priority areas would be reduced.

Despite this, the model is useful for organizations with differing maturity stages in areas of DG, where detailed quantitative measures may not yet exist. It offers a practical method for balancing risk and benefit when implementing DG.

Future studies should focus on improving it with more formal risk criteria and use quantitatively measured insights for business value, for example, through PLS-SEM, as recommended in Section 4.3.

Chapter 7

Conclusion

7.1 Answers to the research questions

In this study, we explored how DG practices contribute to business value and proposed a model to help prioritize DG practices in organizations. The research began with a literature study, which reviewed important DG practices, ways to measure business value, and prioritization models. This provided a foundation for the practical part of this study: the improvement of the HR firm's AI DG Framework and the prioritization of DG practices.

Through questionnaires and interviews with IT- and business departments, we collected data from five markets where the HR firm operates in. This data was analyzed in several ways in order to answer the various research questions: a structured maturity assessment against the AI DG Framework, a thematic coding analysis, and a gap analysis.

The maturity assessment shows uneven maturity across controls of the AI DG Framework, with higher maturity in security and compliance-related controls and lower maturity in roles, data quality and data lineage. While the markets CBS and DE generally scored higher, all markets shared common improvement areas, such as that formalized processes are missing.

The thematic coding analysis revealed that DG practices relating to data quality measurement, metadata management, and data architecture were perceived as delivering the highest business value. On the other hand, controls linked to security and privacy, training and awareness, and roles were less emphasized by the markets, indicating less immediate value. Data quality and data architecture were also perceived as the biggest current challenges, along with roles and training and awareness. While these qualitative insights are valuable, future work should incorporate quantitative measuring methods for stronger evidence.

Following the background study and data analysis, we compared the AI DG Framework to DG best practices from literature and the five HR firm markets. While the framework already incorporates key DG elements, it can be extended with literature and market implementations to address relevant and actionable controls. Our suggestions include the modification of percentage-based controls and the addition of practices such as metadata strategy, data lineage, and training and awareness.

Our proposed prioritization model aims to balance risk and business value by prioritizing controls with the highest expected business value and lower maturity. Using insights from our previous analyses, we demonstrated the proposed model for the UK market. The model suggested a focus on data quality, data transformation/enrichment standards, and data strategy and architecture.

Overall, this research highlights the importance of aligning DG practices with both regulatory requirements and business needs while providing insights for improvement.

7.2 Contributions

The study contributes to both academics and the industry in several ways.

First, we demonstrated a mixed method of document reviews, questionnaires, interviews and framework mapping for assessing DG maturity in an organization.

Second, we identified DG areas that drive business value while linking them to operational and strategic benefits. This provides a foundation for finding the most important DG improvement points and a justification for investing in these areas.

Third, we proposed refinements to the AI DG Framework, integrating best practices from literature and practical implementations to improve its usability and completeness. This may be used by organizations with a similar context and DG foundation.

Finally, we introduced a prioritization model that balances risk and business value to aid organizations in prioritizing DG areas.

7.3 Future work

Although this study provides a foundation for assessing and improving DG in organizations, there are several suggestions for future work.

Our assessment method was effective, but future studies could consider alternative assessment methods. For example, self-assessments or built-in (maturity) scoring tools could provide consistent answers more efficiently.

Furthermore, the prioritization model could be enhanced with different risk factors and business value measurements. It should be tested in various organizational contexts to validate its robustness and usability. This should also be investigated for the refined AI DG Framework.

Lastly, as basic DG controls in organizations improve and adoption of AI increases, future work should incorporate and test more specific AIG controls.

By addressing these gaps, the field of DG and AIG can benefit from more complete, practical and validated frameworks and models, further aiding organizations in implementing regulatory compliant and valuable controls.

Bibliography

[1] Rene Abraham, Johannes Schneider, and Jan vom Brocke. Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, 49:424–438, 2019. ISSN 0268-4012. doi: https://doi.org/10.1016/j.ijinfomgt.2019.07.008. URL `https://www.sciencedirect.com/science/article/pii/S0268401219300787`.

[2] Raed AI Abueed and Mehmet Aga. Sustainable knowledge creation and corporate outcomes: Does corporate data governance matter? *Sustainability*, 11(20):5575, 2019.

[3] Adebunmi Okechukwu Adewusi, Ugochukwu Ikechukwu Okoli, Ejuma Adaga, Temidayo Olorunsogo, Onyeka Franca Asuzu, and Donald Obinna Daraojimba. Business intelligence in the era of big data: A review of analytical tools and competitive advantage. *Computer Science & IT Research Journal*, 5(2):415–431, 2024.

[4] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*. Harvard Business Press, 2022.

[5] Shahriar Akter, Samuel Fosso Wamba, and Saifullah Dewan. Why pls-sem is suitable for complex modelling? an empirical illustration in big data analytics quality. *Production Planning & Control*, 28(11-12):1011–1021, 2017.

[6] Ali Al-Badi, Ali Tarhini, and Asharul Islam Khan. Exploring big data governance frameworks. *Procedia computer science*, 141:271–277, 2018.

[7] Ibrahim Alhassan, David Sammon, and Mary Daly. Data governance activities: A comparison between scientific and practice-oriented literature. *Journal of enterprise information management*, 31(2):300–316, 2018.

[8] Wasif Ali and Huzaifa Arsalan. Ensuring data security and privacy: Strategies for targeted data discovery, data management systems, and private data access in educational settings, 05 2024.

[9] Judie Attard and Rob Brennan. Challenges in value-driven data governance. In *On the Move to Meaningful Internet Systems. OTM 2018 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part II*, pages 546–554. Springer, 2018.

[10] Amazon Web Services (AWS). How jpmorgan chase built a data mesh architecture to drive significant value to enhance their enterprise data platform, 2021. URL `https://aws.amazon.com/tr/blogs/big-data/how-jpmorgan-chase-built-a-data-mesh-architecture-to-drive-significant-value-to-enhance-their-e`

[11] James M Barker. *Data Governance: the missing approach to improving data quality*. University of Phoenix, 2016.

[12] Filipe Andrade Bernardi, Domingos Alves, Nathalia Crepaldi, Diego Bettiol Yamada, Vinícius Costa Lima, and Rui Rijo. Data quality in health research: integrative literature review. *Journal of medical Internet research*, 25:e41446, 2023.

[13] Teemu Birkstedt, Matti Minkkinen, Anushree Tandon, and Matti Mäntymäki. Ai governance: themes, knowledge gaps and future agendas. *Internet Research*, 33(7):133–167, 2023.

[14] Tor Broström and Karin Svahnström. Solar energy and cultural-heritage values. *World Renewable Energy Conference, vol. 8 (Low-Energy Architecture)*, 11 2011. doi: 10.3384/ecp110572034.

[15] Paul Brous, Marijn Janssen, and Rutger Krans. Data governance as success factor for data science. In *Conference on e-Business, e-Services and e-Society*, pages 431–442. Springer, 2020.

[16] Erik Brynjolfsson, Lorin M Hitt, and Heekyung Hellen Kim. Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*, 2011.

[17] Lai Kuan Cheong and Vanessa Chang. The need for data governance: a case study. *ACIS 2007 proceedings*, page 100, 2007.

[18] Carol Corrado, Charles Hulten, and Daniel Sichel. Intangible capital and us economic growth. *Review of income and wealth*, 55(3):661–685, 2009.

[19] Nadine Côrte-Real, Tiago Oliveira, and Pedro Ruivo. Assessing business value of big data analytics in european firms. *Journal of Business Research*, 70:379–390, 2017.

[20] Francesco Dallanoce. Explainable ai: A complete summary of the main methods, 2022. URL `https://medium.com/@dallanoce.fd/explainable-ai-a-complete-summary-of-the-main-methods-a28f9ab132f7`. Published on Medium.

[21] DataVersity. Creating and implementing a metadata strategy, 2022. URL `https://www.dataversity.net/creating-and-implementing-a-metadata-strategy/`.

[22] Paul R Daugherty and H James Wilson. *Human+ machine: Reimagining work in the age of AI*. Harvard Business Press, 2018.

[23] Patricia Gomes Rêgo De Almeida, Carlos Denner dos Santos, and Josivania Silva Farias. Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology*, 23(3): 505–525, 2021.

[24] Tanvi Desai, Felix Ritchie, Richard Welpton, et al. Five safes: designing data access for research. *Economics Working Paper Series*, 1601:28, 2016.

[25] Ren Bin Lee Dixon. A principled governance for emerging ai regimes: lessons from china, the european union, and the united states. *AI and Ethics*, 3(3):793–810, 2023.

[26] Yanqing Duan, John S Edwards, and Yogesh K Dwivedi. Artificial intelligence for decision making in the era of big data–evolution, challenges and research agenda. *International journal of information management*, 48:63–71, 2019.

[27] Nijs Jan Duijm. Recommendations on the use and design of risk matrices. *Safety science*, 76:21–31, 2015.

[28] Mohamed Z. Elbashir, Philip A. Collier, and Michael J. Davern. Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *International Journal of Accounting Information Systems*, 9(3):135–153, 2008. ISSN 1467-0895. doi: https://doi.org/10.1016/j.accinf.2008.03.001. URL `https://www.sciencedirect.com/science/article/pii/S1467089508000353`. Eighth International Research Symposium on Accounting Information Systems (IRSAIS).

[29] European Commission. Regulatory framework proposal on artificial intelligence, 2024. URL `https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai`. Accessed: 2025-02-18.

[30] Flevy. Case: Gaming company's strategic metadata management framework to overcome data challenges, 2025. URL `https://flevy.com/topic/metadata-management/case-gaming-companys-strategic-metadata-management-framework-overcome-data-challenges`.

[31] Flevy. Case: Transforming data management—an electronics manufacturer's strategic journey, 2025. URL `https://flevy.com/topic/information-architecture/case-transforming-data-management-an-electronics-manufacturers-strategic-journey`.

[32] Flevy. Case: Streamlining data governance—building material industry's metadata management, 2025. URL `https://flevy.com/topic/metadata-management/case-streamlining-data-governance-building-material-industry-metadata-management`.

[33] Flevy. Case: Metadata management initiative in a biotech firm for precision medicine, 2025. URL https://flevy.com/topic/metadata-management/case-metadata-management-initiative-biotech-firm-precision-medicine.

[34] Gartner. Create a business case for metadata management to best fulfill your data and analytics initiatives. Technical Report 3913692, Gartner, 2019. URL https://www.gartner.com/en/documents/3913692. Gartner Document ID: 3913692.

[35] Gartner. Definition: Enterprise metadata management (emm), 2025. URL https://www.gartner.com/en/information-technology/glossary/enterprise-metadata-management-emm.

[36] Kartikay Goyle, Quin Xie, and Vakul Goyle. Dataassist: A machine learning approach to data cleaning and preparation. In *Intelligent Systems Conference*, pages 476–486. Springer, 2024.

[37] Maximilian Grafenstein. Reconciling conflicting interests in data through data governance. an analytical framework (and a brief discussion of the data governance act draft, the data act draft, the ai regulation draft, as well as the gdpr). 2022.

[38] Varun Grover, Roger HL Chiang, Ting-Peng Liang, and Dongsong Zhang. Creating strategic business value from big data analytics: A research framework. *Journal of management information systems*, 35(2):388–423, 2018.

[39] Susie Gu. Exploring the role of ai in business decision-making and process automation. *International Journal of High School Research*, 6(3), 2024.

[40] Joseph F Hair, Jeffrey J Risher, Marko Sarstedt, and Christian M Ringle. When to use and how to report the results of pls-sem. *European business review*, 31(1):2–24, 2019.

[41] Elena Huff and John Lee. Data as a strategic asset: Improving results through a systematic data governance framework. In *SPE Latin America and Caribbean Petroleum Engineering Conference*, page D031S013R001. SPE, 2020.

[42] IBM. The 6 pillars of data quality and how to improve your data, 2023. URL https://www.ibm.com/products/tutorials/6-pillars-of-data-quality-and-how-to-improve-your-data#6pillarsofdataquality. IBM Tutorial.

[43] Dama International. *DAMA-DMBOK: Data Management Body of Knowledge (2nd Edition)*. Technics Publications, LLC, Denville, NJ, USA, 2017. ISBN 1634622340.

[44] Ashraful Islam. Data governance and compliance in cloud-based big data analytics: A database-centric review. 2024.

[45] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly*, 37(3): 101493, 2020. ISSN 0740-624X. doi: https://doi.org/10.1016/j.giq.2020.101493. URL https://www.sciencedirect.com/science/article/pii/S0740624X20302719.

[46] Steven Ji-fan Ren, Samuel Fosso Wamba, Shahriar Akter, Rameshwar Dubey, and Stephen J Childe. Modelling quality dynamics, business value and firm performance in a big data analytics environment. *International Journal of Production Research*, 55(17):5011–5026, 2017.

[47] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.

[48] Sivananda Reddy Julakanti, Naga Satya KiranmayeeSattiraju, and Rajeswari Julakanti. Data protection through governance frameworks. *arXiv preprint arXiv:2502.10404*, 2025.

[49] Vijay Khatri and Carol V Brown. Designing data governance. *Communications of the ACM*, 53(1): 148–152, 2010.

[50] Kanaka Rakesh Varma Kothapalli. Exploring the impact of digital transformation on business operations and customer experience. *Global Disclosure of Economics and Business*, 11(2):103–114, 2022.

[51] Sanjay Krishnan and Eugene Wu. Alphaclean: Automatic generation of data cleaning pipelines. *arXiv preprint arXiv:1904.11827*, 2019.

[52] John Ladley. *Data governance: How to design, deploy, and sustain an effective data governance program*. Academic Press, 2019.

[53] Laura Lucaj, Patrick Van Der Smagt, and Djalel Benbouzid. Ai regulation is (not) all you need. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1267–1279, 2023.

[54] Luxafor. The eisenhower matrix: How to prioritize your tasks. `https://luxafor.com/the-eisenhower-matrix/`, 2023. URL `https://luxafor.com/the-eisenhower-matrix/`. Accessed: 2023-11-15.

[55] Nick Martijn, Joris Hulstijn, Mark de Bruijne, and Yao-Hua Tan. Determining the effects of data governance on the performance and compliance of enterprises in the logistics and retail sector. In Marijn Janssen, Matti Mäntymäki, Jan Hidders, Bram Klievink, Winfried Lamersdorf, Bastiaan van Loenen, and Anneke Zuiderwijk, editors, *Open and Big Data Management and Innovation*, pages 454–466, Cham, 2015. Springer International Publishing. ISBN 978-3-319-25013-7.

[56] Alfred Homère NGANDAM Mfondoum, Mesmin Tchindjang, Jean Valery, M Mfondoum, and I Makouet. Eisenhower matrix* saaty ahp= strong actions prioritization? theoretical literature and lessons drawn from empirical evidences. *Iaetsd Journal For Advanced Research In Applied Sciences. Retrieved from https://www. iaetsdjaras. org/gallery/3-february-880. pdf*, 2019.

[57] Patrick Mikalef, Maria Boura, George Lekakos, and John Krogstie. The role of information governance in big data analytics driven innovation. *Information & Management*, 57(7):103361, 2020.

[58] Oliver Müller, Maria Fay, and Jan Vom Brocke. The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *Journal of management information systems*, 35(2):488–509, 2018.

[59] Thu Nguyen, Hong-Tri Nguyen, and Tu-Anh Nguyen-Hoang. Data quality management in big data: Strategies, tools, and educational implications. *Journal of Parallel and Distributed Computing*, 200: 105067, 2025.

[60] Gabriel P Oliveira, Bárbara MA Mendes, Clara A Bacha, Lucas L Costa, Larissa D Gomide, Mariana O Silva, Michele A Brandão, Anisio Lacerda, and Gisele L Pappa. Assessing data quality inconsistencies in brazilian governmental data. *Journal of Information and Data Management*, 14 (1), 2023.

[61] Boris Otto. A morphology of the organisation of data governance. 2011.

[62] Boris Otto. Managing the business benefits of product data management: the case of festo. *Journal of Enterprise Information Management*, 25(3):272–297, 2012.

[63] Emmanouil Papagiannidis, Ida Merete Enholm, Chirstian Dremel, Patrick Mikalef, and John Krogstie. Toward ai governance: Identifying best practices and potential barriers and outcomes. *Information Systems Frontiers*, 25(1):123–141, 2023.

[64] Elena Parmiggiani and Miria Grisot. Data curation as governance practice. *Scandinavian Journal of Information Systems*, 32(1):Article 1, 2020. URL `https://aisel.aisnet.org/sjis/vol32/iss1/1`.

[65] Sunil Pathak, Venkataraghavan Krishnaswamy, and Mayank Sharma. Impact of it practices and business value of it measurement. *International Journal of Productivity and Performance Management*, 69(4):774–793, 2020.

[66] Arif Perdana, Hwee Hoon Lee, SzeKee Koh, and Desi Arisandi. Data analytics in small and mid-size enterprises: Enablers and inhibitors for business value and firm performance. *International Journal of Accounting Information Systems*, 44:100547, 2022. ISSN 1467-0895. doi: https://doi.org/10.1016/j.accinf.2021.100547. URL `https://www.sciencedirect.com/science/article/pii/S146708952100049X`.

[67] João Silva Pestana. Data governance valuation: A model for assessing the impact on organisations´ business. Master's thesis, Universidade NOVA de Lisboa (Portugal), 2023.

[68] David Plotkin. *Data stewardship: an actionable guide to effective data management and data governance*. Academic press, 2020.

[69] Elisabetta Raguseo and Claudio Vitari and. Investments in big data analytics and firm performance: an empirical investigation of direct and mediating effects. *International Journal of Production Research*, 56(15):5206–5221, 2018. doi: 10.1080/00207543.2018.1427900. URL `https://doi.org/10.1080/00207543.2018.1427900`.

[70] Gautam Ray, Jay B Barney, and Waleed A Muhanna. Capabilities, business processes, and competitive advantage: choosing the dependent variable in empirical tests of the resource-based view. *Strategic management journal*, 25(1):23–37, 2004.

[71] Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016, 2016.

[72] Ishak Firdauzi Ruslan, Muhammad Fahmi Alby, and Muharman Lubis. Applying data governance using dama-dmbok 2 framework: The case for human capital management operations. In *Proceedings of the 8th International Conference on Industrial and Business Engineering*, pages 336–342, 2022.

[73] Felipe F Salerno and Antonio Carlos G Maçada. Analysis of the relationship between data governance and data-driven culture, 2022.

[74] David Sammon and Tadhg Nagle. The data value map: a framework for developing shared understanding on data initiatives. 2017.

[75] Mark Saunders. Research methods for business students. *Person Education Limited*, 2009.

[76] Secoda. How panasonic manages data at scale, 2024. URL `https://www.secoda.co/customers/how-panasonic-manages-data-at-scale`.

[77] Brayan V Seixas, Dean A Regier, Stirling Bryan, and Craig Mitton. Describing practices of priority setting and resource allocation in publicly funded health care systems of high-income countries. *BMC health services research*, 21(1):90, 2021.

[78] Shafaq Siddiqi, Roman Kern, and Matthias Boehm. Saga: a scalable framework for optimizing data cleaning pipelines for machine learning applications. *Proceedings of the ACM on Management of Data*, 1(3):1–26, 2023.

[79] Tomáš Sobotík. How to ensure data quality with great expectations, 2021. URL `https://medium.com/snowflake/how-to-ensure-data-quality-with-great-expectations-271e3ca8b4b9`. Published on Medium (Snowflake Engineering Blog).

[80] Araz Taeihagh. Governance of artificial intelligence. *Policy and society*, 40(2):137–157, 2021.

[81] Arzu Teymurova. Metadata management in data governance: A comprehensive scientific analysis. 04 2025.

[82] Her Majesty's Treasury. The magenta book: guidance notes for policy evaluation and analysis. *London: HM Treasury (Magenta Book Background Papers*, 2007.

[83] U.S. Department of Defense. Dod data strategy. Technical report, U.S. Department of Defense, October 2020. URL `https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DOD-DATA-STRATEGY.PDF`. Document ID: 2002514180.

[84] Michael Joseph Walsh, John McAvoy, and David Sammon and. The data governance journey in practice: Insights from case study research. *Information Systems Management*, 0(0):1–18, 2025. doi: 10.1080/10580530.2025.2477459. URL `https://doi.org/10.1080/10580530.2025.2477459`.

[85] Serge-Lopez Wamba-Taguimdje, Samuel Fosso Wamba, Jean Robert Kala Kamdjoug, and Chris Emmanuel Tchatchouang Wanko. Influence of artificial intelligence (ai) on firm performance: the business value of ai-based transformation projects. *Business process management journal*, 26(7): 1893–1924, 2020.

[86] Jingran Wang, Yi Liu, Peigong Li, Zhenxing Lin, Stavros Sindakis, and Sakshi Aggarwal. Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality. *Journal of the Knowledge Economy*, 15(1):1159–1178, 2024.

[87] Damian Whittard, Felix Ritchie, Ruthie Musker, and Michael Rose. Measuring the value of data governance in agricultural investments: A case study. *Experimental Agriculture*, 58:e8, 2022.

Appendix A

AI Data Governance Framework

This chapter presents a shortened version of the AI Data Governance Framework developed by the HR firm in this study, introduced in Section 3.2.

The framework is designed as a capability maturity model. It includes 28 DG controls (best practices), grouped into five categories: roles & responsibilities, data lineage, data quality & purpose, data architecture & platform, and security.

For each control the following is described: the control objective, a description, the specific control, and criteria for five levels of maturity. In the shortened version below, we omit the controls and summarize the descriptions. The structure is as follows:

- The controls are divided over six sections according to their category in the framework.
- The first line in bold indicates the control objective.
- The second line contains the summarized description.
- The numbered list represents the maturity level criteria, with the numbers corresponding to the maturity level.

To provide clarification on the capability maturity model design, we explain the meaning of each maturity level, which is linked to a risk level:

1. Ad hoc (high risk): processes are performed as needed and only at the project level. Data issues are fixed reactively rather than proactively through improved processes. Data is not considered a strategic resource.
2. Managed (medium risk): processes are planned and executed within policy guidelines, but the tools and skills for managing data are still inadequate. Data management is taken more seriously.
3. Defined (minimum requirement): standard processes help provide consistent data quality to meet business needs and regulatory compliance. Management oversight has been introduced along with monitoring and feedback loops.
4. Measured (benchmark): process metrics are judged against agreed upon variances. Data is treated as an asset and every is concerned with its accuracy and timeliness. Applications are written to capture data issues that are resolved quickly to avoid fines or reputational damage.
5. Optimized (exceeds benchmark): process performance is continuously improved through feedback from various sources. Data is regarded as a critical asset and vital element for our business to operate.

A.1 Responsibilities

Responsibilities are clearly defined for all stakeholders.
Clearly defining roles and responsibilities will help ensure that data is managed efficiently, compliantly and effectively.

1. There are no formal roles & responsibilities in place.

2. Roles & responsibilities are informally assigned but not documented.
3. Roles & responsibilities are formally assigned, but not consistently enforced.
4. Roles & responsibilities are formally defined, documented and partially enforced.
5. Roles & responsibilities are fully defined, documented & enforced (KPI).

The responsibilities have been appointed to certain functions within the organization.
Assigning responsibilities to specific functions will ensure that individuals can effectively carry out their designated tasks.

1. No responsibilities have been formally appointed.
2. Responsibilities are appointed informally and inconsistently.
3. Responsibilities are formally appointed but not fully integrated into the org structure.
4. Responsibilities are formally appointed, documented, and partially integrated into the org structure.
5. Responsibilities are fully appointed, documented and integrated into the org structure.

Clear reporting lines on the data governance responsibilities exist.
Clear reporting lines for data governance will increase transparency and accountability.

1. No clear reporting lines exist for data governance responsibilities.
2. Reporting lines for data governance are informally understood but not documented.
3. Reporting lines for data governance are documented but not fully enforced.
4. Reporting lines are documented and partially enforced with regular reviews.
5. Clear and consistent reporting lines are fully documented, enforced and integrated into the organizational structure.

The framework also defines two roles, but these are not included in the form of a control:

- Business data owner: a business executive who is able to drive business priorities and has extensive business knowledge on subject area. This person is accountable for definitions.
- Data owner: a data savvy person who has extensive business knowledge on subject area or of specific system. This person is responsible for alignment with the data, definitions, and data quality rules.

A.2 Data lineage

Master data management ensures data quality. For each data category (such as Talent data, Client data, People data) it has been defined what the master data is.
Master Data Management (MDM) is a comprehensive process and set of technologies that aims to create a single, consistent, and accurate view of an organization's critical business data, often referred to as "master data."

1. No master data is defined.
2. Master data definitions are not documented and applied ad hoc.
3. Master data definitions have been documented but are not applied consistently for all data sources.
4. Master data definitions have been documented and are applied consistently for all data sources.
5. Master data definitions have been documented and are applied consistently for the entire IT & data landscape.

Preparation - The data preparation activities need to be known.
By carefully preparing the data, businesses can ensure that they are making decisions based on accurate, reliable, and relevant information. Data preparation consists of transforming raw data into a format that can be analysed and used for decision-making.

1. No data preparation activities are known or logged.
2. 25% of the activities done to prepare data are known and logged.
3. 50% of the activities done to prepare data are known and logged.
4. 75% of the activities done to prepare data are known and logged.
5. 100% of the activities done to prepare data are known and logged.

Cleaning - To ensure we identify and correct errors in the data, or know how to handle missing values, data cleaning protocols have been developed and implemented.
Removing errors, inconsistencies, and duplicate data from the raw dataset. This ensures that the data is accurate and reliable.

1. No data cleaning guidelines have been developed or implemented.
2. Basic data cleaning guidelines exist, but are not yet standardised.
3. Data cleaning guidelines are standardised and followed by most teams.
4. Data cleaning guidelines are followed by all teams.
5. Data cleaning guidelines are regularly reviewed and followed by all teams.

Transformation/Enrichment - When data is being transformed/enriched, all transformations must be documented. This ensures traceability and ensures correct understanding of our data.
Adding additional information to the dataset from other sources. This helps to create a more comprehensive and informative dataset.

1. No data transformation/enrichment standards exist.
2. Basic data transformation/enrichment standards exist, but are not regularly followed.
3. The data transformation/enrichment standards are documented and followed by some teams.
4. The data transformation/enrichment standards are documented and followed by all teams.
5. The data transformation/enrichment standards are regularly reviewed and updates are followed by all teams.

Classification - All data must be classified according the classification scheme based on sensitivity and importance to the business.
The data classification framework categorizes information by sensitivity and value to enable proper security controls, ensure regulatory compliance, and support risk-based data handling. This enhances protection, reduces risks, and strengthens governance practices.

1. No data is classified.
2. 25% of the data has been classified.
3. 50% of the data has been classified.
4. 75% of the data has been classified.
5. 100% of the data has been classified.

Minimization - Where possible personal data usage is minimised and if needed personal data risks are mitigated via aggregation or anonymization.
Personal data minimization, aggregation and anonymization are crucial for legal compliance, risk reduction, building trust, and ethical data use. Only essential personal data should be collected, copied and used (data minimization). Where possible data should be used to reveal trends, not individuals (aggregation). And if possible we should remove identifying details for privacy (anonymization).

1. No data minimization, aggregation and anonymization standards exist.
2. Basic data minimization, aggregation and anonymization standards exist, but are not regularly followed.
3. The data minimization, aggregation and anonymization standards are documented and applied by some teams. No personal data is used in non-production environments.
4. The data minimization, aggregation and anonymization standards are documented and applied by all teams. No personal data is used in non-production environments and sensitive personal data fields are hidden in production environments.
5. The data minimization, aggregation and anonymization standards are regularly reviewed and updates are applied by all teams. No personal data is used in non-production environments and sensitive personal data fields are hidden in production environments.

Sharing and Visibility - Data sharing in alignment with our data classification and appropriate business and legal requirement.
Data sharing internally, with clients, with suppliers and with other stakeholders is only done in alignment with our data classification and appropriate business and legal requirements.

1. No data sharing standards exist.
2. Basic data sharing standards exist, but are not consistently followed.
3. The data sharing standards are documented and applied by some teams and embedded in authorization models of some systems.
4. The data sharing standards are documented, applied by all teams, embedded in authorization models of all systems and enforced via Data Loss Prevention technology.
5. The data minimization, aggregation and anonymization standards are regularly reviewed and updates are applied by all teams, embedded in authorization models of all systems and enforced via Data Loss Prevention technology.

Third Party Sources - Data ingestion of third parties.
Before data is obtained or ingested from third parties, it is validated whether the data is obtained in line with our data quality, legal, security and data protection standards.

1. Third party data is added to our datasets without any checks or limitations upfront.
2. Third party data is ad hoc checked before added to our datasets based on non documented standards.
3. Third party data is most times checked before added to our datasets based on documented standards.
4. Third party data is always checked before added to our datasets based on documented standards.
5. Third party data is always checked before added to our datasets and periodically reviewed based on documented standards.

A.3 Purpose

The original purpose for which the personal data will be used needs to be clear and documented according to the DP policy (the data mapping in OneTrust).
We are required to know the original purpose for which the (sensitive) personal data has been captured and where (which source system) this was done. To determine which data quality standards need to be met, the use case and the data elements required for it need to be clear.

1. The processing activity has not been documented yet.
2. This processing activity has been partially documented, but is not complete.
3. This processing activity is documented (as a one-off) but is not maintained.
4. This processing activity is documented; maintenance is infrequent but does occur.
5. The processing activity is finalised in OneTrust. This is continuously being maintained in an effective and efficient manner to ensure accuracy.

The original purpose of the application or tool that uses data (personal AND non-personal) needs to be clear and documented in our asset register (as per the security policy).
To determine which data quality standards need to be met, the use case and the data elements required for it need to be clear.

1. For 0% of the tools/applications purposes have been documented in our asset list.
2. For 25% of the tools/applications purposes have been documented in our asset list.
3. For 50% of the tools/applications purposes have been documented in our asset list.
4. For 75% of the tools/applications purposes have been documented in our asset list.
5. For 100% of the tools/applications purposes have been documented in our asset list.

A.4 Master data quality

Data categories (such as Talent data, Client data, HR data) have been distinguished following (to be defined) global guidelines.
As we have a lot of data as a company, it is important to divide the data into certain categories. This will help in managing, using and protecting the data and also help appointing data owners.

1. Data is not categorised.
2. 25% of the data has been categorised using local categorization.
3. 50% of the data of key processes has been categorised using globally aligned definitions.
4. 75% of the data of key processes has been categorised using globally aligned definitions.
5. 100% of the data has been categorised using globally aligned definitions.

Completeness - The level of completeness of the data required, needs to be defined and be measurable.
It is clear which items are necessary for the initiatives; to be able to tell which dimensions of completeness are required. It is possible to know the level of completeness per data set.

1. No criteria to measure completeness have been formulated for any data set, nor is this measured.
2. Criteria to measure completeness have been formulated and are measurable for 25% of the data sets.
3. Criteria to measure completeness have been formulated and are measurable for 50% of the data sets.
4. Criteria to measure completeness have been formulated and are measurable for 75% of the data sets.
5. Criteria to measure completeness have been formulated and are measurable for 100% of the data sets.

Accuracy - The level of accuracy of the data required, needs to be defined and measured.
The accuracy of critical data elements must be ensured. Meaning the data must be free from errors, not be inconsistent nor be misinterpreted.

1. No accuracy metrics have been defined, nor is this measured.
2. Accuracy metrics have been defined and are measured for 25% of the datasets.
3. Accuracy metrics have been defined and are measured for 50% of the datasets.
4. Accuracy metrics have been defined and are measured for 75% of the datasets.
5. Accuracy metrics have been defined and are measured for 100% of the datasets.

Freshness - Up-to-dateness of the data required is determined and measured.
Per use case it can differ how recent the data set needs to be. Therefore it must be possible to know when the data has last been updated and the history must be available.

1. It is unknown when the data has last been updated.
2. It is known for 25% of the data when it has last been updated.

3. It is known for 50% of the data when it has last been updated.
4. It is known for 75% of the data when it has last been updated.
5. It is known for 100% of the data when it has last been updated.

Consistency - All critical data needs to be consistent or relatable across all sources in scope for the intended use case.
With consistency we aim to unify our data to ensure the same data values are maintained across different locations. For global use cases, alignment with the Globally Aligned Information Model is required.

1. It is unknown which percentage of data is consistent.
2. It is known for 25% of the data if it is consistent.
3. It is known for 50% of the data if it is consistent.
4. It is known for 75% of the data if it is consistent.
5. It is known for 100% of the data if it is consistent.

Deduplicated - To maintain a clean database, improve data quality and improve efficiency, we need to make sure we deduplicate our data.
Data needs to be unique. Each data entity should have a unique identifier. When 100% of the data has been identified with a unique identifier this should result in there not being duplicates.

1. It is unknown which percentage of the data is unique.
2. 25% of the data is confirmed unique.
3. 50% of the data is confirmed unique.
4. 75% of the data is confirmed unique.
5. 100% of the data is confirmed unique.

(Test)- data sets need to comply with specified data quality standards within this framework. The distribution of the data sets is known and documented.
To be able to train and develop any AI- or machine learning models, specific requirements for the datasets need to be formulated up front. The properties of each (test)data set need to be visible, to ensure models can be trained adequately. The (test)data sets must be tested for potential bias.

1. The data distribution is unknown for all necessary (Test)- datasets.
2. The data distribution is known and available for 25% of the necessary (Test)- datasets.
3. The data distribution is known and available for 50% of the necessary (Test)- datasets.
4. The data distribution is known and available for 75% of the necessary (Test)- datasets.
5. The data distribution is known and available for 100% of the necessary (Test)- datasets.

Regular reports are created to measure the quality of our data.
When aiming for a certain level of quality by taking certain measures, it is important to monitor if the measures indeed lead to the required quality.

1. Data quality is not a topic that is being reported on.
2. Data quality is being reported on incidentally.
3. Data quality is being tracked and reported regularly, but without set goals or benchmarks.
4. Data quality targets are established and tracked, with periodic reporting and corrective actions.
5. Aiming for a certain measure of data quality is a KPI and is reported on annually.

A.5 Data architecture & platform

Scalability accessibility, integration and harmonisation of data use is enabled via our data architecture.
Our architecture ensures data is scalable, easily accessible and harmonised. Leveraging approved technology we can efficiently handle structured and unstructured data, ensuring adaptability and growth readiness.

1. There is no data strategy, nor data architectural guidelines.
2. There is a data strategy but no architectural guidelines exist.
3. The data strategy and architectural guidelines exist and are applied by some teams.
4. The data strategy and architectural guidelines are embedded in all teams' way of working and are enforced.
5. The data strategy and architectural guidelines are regularly reviewed.

Data assets are known via our data catalog.
Our data catalog is an organised inventory of data assets within the organization, providing metadata, descriptions, and other relevant information about the data. It helps users discover, understand, and access data efficiently, promoting data governance and informed decision-making.

1. A data asset inventory does not exist.
2. Data assets are known and 25% are in a data catalog.
3. Data assets are known and 50% are in a data catalog.
4. Data assets are known and 75% are in a data catalog.
5. Data assets are known and 100% are in a data catalog.

A.6 Security

Our CIA rating is applied to data assets.
By understanding the CIA rating of our data, we can implement appropriate security measures to protect our valuable assets, private risks, and ensure our continued success.

1. It is unknown what CIA rating data assets have.
2. 25% of the data assets have a CIA rating.
3. 50% of the data assets have a CIA rating.
4. 75% of the data assets have a CIA rating.
5. 100% of the data assets have a CIA rating, which are periodically reassessed.

Data should only be accessible to authorized people for a specific purpose, to ensure confidentiality and compliance.
Access controls regulate who can view, modify, or use specific data or resources within the organization. They are essential for protecting sensitive information, ensuring compliance with regulations, and maintaining data integrity.

1. Access to data is not restricted.
2. Access to data is restricted, but not following a documented RBAC model and our data classification.
3. Access to 50% of our data is restricted, following a RBAC model and data classification.
4. Access to 75% of our data is restricted, following a RBAC model and data classification.
5. Access to 100% of our data is restricted, following a RBAC model and data classification.

To ensure compliance and avoid misuse of data, access rights are regularly reviewed.
We know who has access to which data and for which purpose. There is a process in place to revoke access.

1. No documentation exists on who has access to data and no standard procedure is in place to revoke access rights.
2. It is known, but not centrally documented, who has access to data. Revoking rights is an ad hoc, manual activity.
3. It is known and documented who has access to data. There is a process in place for revoking access to data.
4. Regular reviews are done on who has access to data and if this is still accurate. There is a process in place for revoking access to data.
5. Reviewing who has access and if needed revoking their access rights is incorporated in an automated process which takes place every quarter.

Data is kept no longer than needed.
Based on the retention policy data is not kept or used longer than required based on the purpose it is used for and Redundant, Obsolete, or Trivial data is destroyed when required to minimise storage costs, improve data quality, improve processes and lower security, legal and data protection risks.

1. Data retention policy is not documented.
2. Data retention policy is documented but not implemented.
3. Data retention policy is documented and via manual processes implemented.
4. Data retention policy is documented and via automated processes implemented.
5. Data retention policy is documented and via automated processes implemented and periodically reviewed.

Storage - To comply with regulatory requirements, data must be stored and processed within specified geographical boundaries.
This means the data must be stored in a specific geographic location, but also processing of that data (incl. possible data sharing) must be subject to strict controls. This to prevent unauthorised cross-border data transfers.

1. Storing of data does not take any geographical boundaries into account.
2. 25% of data is stored and processed based on localization regulations.
3. 50% of data is stored and processed based on localization regulations.
4. 75% of data is stored and processed based on localization regulations.
5. 100% of data is stored and processed based on localization regulations.