Year : 2021

BIG DATA AND ANALYTICS AS A NEW FRONTIER OF ENTERPRISE DATA MANAGEMENT

Fadler Martin

Fadler Martin, 2021, BIG DATA AND ANALYTICS AS A NEW FRONTIER OF ENTERPRISE DATA MANAGEMENT

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_19C4AEE4366A3

UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES

DÉPARTEMENT DES SYSTÈMES D'INFORMATION

BIG DATA AND ANALYTICS AS A NEW FRONTIER OF ENTERPRISE DATA MANAGEMENT

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès Sciences en systèmes d'information

par

Martin FADLER

Directrice de thèse
Prof. Christine Legner

Jury

Prof. Rafael Lalive, Président
Prof. Michalis Vlachos, expert interne
Prof. Boris Otto, expert externe
Prof. Patrick Mikalef, expert externe

LAUSANNE
2021

UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES

DÉPARTEMENT DES SYSTÈMES D'INFORMATION

BIG DATA AND ANALYTICS AS A NEW FRONTIER OF ENTERPRISE DATA MANAGEMENT

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès Sciences en systèmes d'information

par

Martin FADLER

Directrice de thèse
Prof. Christine Legner

Jury

Prof. Rafael Lalive, Président
Prof. Michalis Vlachos, expert interne
Prof. Boris Otto, expert externe
Prof. Patrick Mikalef, expert externe

LAUSANNE
2021

UNIL | Université de Lausanne
HEC Lausanne
Le Décanat
Bâtiment Internef
CH-1015 Lausanne

I M P R I M A T U R

Sans se prononcer sur les opinions de l'auteur, la Faculté des Hautes Etudes Commerciales de l'Université de Lausanne autorise l'impression de la thèse de Monsieur Martin FADLER, titulaire d'un bachelor en ingénierie économique de l'Université Technique de Clausthal, et d'un master en ingénierie économique de l'Université Technique de Berlin, en vue de l'obtention du grade de docteur ès Sciences en systèmes d'information.

La thèse est intitulée :

BIG DATA AND ANALYTICS AS A NEW FRONTIER OF ENTERPRISE DATA MANAGEMENT

Lausanne, le 21 septembre 2021

La Doyenne

Marianne SCHMID MAST

Members of the thesis committee

Prof. Christine LEGNER

Professor at the Faculty of Business and Economics (HEC) of the University of Lausanne, Switzerland.

Thesis supervisor

Prof. Michalis VLACHOS

Professor at the Faculty of Business and Economics (HEC) of the University of Lausanne, Switzerland.

Internal member of the thesis committee

Prof. Boris OTTO

Professor at the Technical University Dortmund, Germany.

External member of the thesis committee

Prof. Patrick MIKALEF

Associate Professor at the Norwegian University of Science and Technology.

External member of the thesis committee

Prof. Rafael LALIVE

Professor at the Faculty of Business and Economics (HEC) of the University of Lausanne, Switzerland.

President of the thesis committee

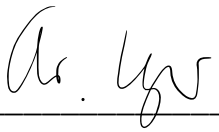
University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Martin FADLER

and have found it to meet the requirements for a doctoral thesis.
All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:  Date: September 7, 2020

Prof. Christine LEGNER
Thesis supervisor

University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Martin FADLER

and have found it to meet the requirements for a doctoral thesis.
All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature: _____ Date: 15.09.2021

Prof. Michalis VLACHOS
Internal member of the doctoral committee

University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Martin FADLER

and have found it to meet the requirements for a doctoral thesis.
All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature: _____ Date: 7.9.2021

Prof. Boris OTTO
External member of the doctoral committee

University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Martin FADLER

and have found it to meet the requirements for a doctoral thesis.
All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature: Patrick Mikalef Date: 08/09/2021

Prof. Patrick MIKALEF
External member of the doctoral committee

Acknowledgements

Firstly, I would like to thank my supervisor, Christine Legner. She gave me the opportunity to do my PhD in the first place. She has also influenced me with her passion for research and has always pushed me to my limits, so that I have learned much more than I expected in the last four years. Without her, my PhD journey would not have been as exciting and interesting. Secondly, I would like to thank my parents, Haidi and Georges Fadler. Without their great support and constant efforts to keep me on track in life, I would not have been able to start my doctoral studies at all. Thirdly, I would like to thank my fiancée, Polina Popova. She motivated me during difficult times and was extremely helpful in helping me persevere. Without her support, I probably would not have finished my doctoral studies. Fourthly, I would like to thank all my colleagues and friends. Without their help and necessary distraction, my doctoral journey would not have gone so smoothly.

UNIVERSITÉ DE LAUSANNE

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES

DÉPARTEMENT DES SYSTÈMES D'INFORMATION

BIG DATA AND ANALYTICS AS A NEW FRONTIER OF ENTERPRISE DATA MANAGEMENT

THÈSE DE DOCTORAT

Présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès Sciences en systèmes d'information

par
Martin FADLER

Directrice de thèse
Prof. Christine Legner

Jury
Prof. Rafael Lalive, président
Prof. Michalis Vlachos, expert interne
Prof. Boris Otto, experte externe
Prof. Patrick Mikalef, expert externe

LAUSANNE
2021

Members of the thesis committee

Prof. Christine LEGNER

Professor at the Faculty of Business and Economics (HEC) of the University of Lausanne, Switzerland.

Thesis supervisor

Prof. Michalis VLACHOS

Professor at the Faculty of Business and Economics (HEC) of the University of Lausanne, Switzerland.

Internal member of the thesis committee

Prof. Boris OTTO

Professor at the Technical University Dortmund, Germany.

External member of the thesis committee

Prof. Patrick MIKALEF

Associate Professor at the Norwegian University of Science and Technology.

External member of the thesis committee

Prof. Rafael LALIVE

Professor at the Faculty of Business and Economics (HEC) of the University of Lausanne, Switzerland.

President of the thesis committee

Doctoral Thesis

BIG DATA AND ANALYTICS AS A NEW FRONTIER OF ENTERPRISE DATA MANAGEMENT

Martin Fadler

Department of Information Systems,
Faculty of Business and Economics (HEC), University of Lausanne

Lausanne, 2021

"Difficulties strengthen the mind, as labor does the body."

– Seneca

Abstract

Big Data and Analytics (BDA) promises significant value generation opportunities across industries. Even though companies increase their investments, their BDA initiatives fall short of expectations and they struggle to guarantee a return on investments. In order to create business value from BDA, companies must build and extend their data-related capabilities. While BDA literature has emphasized the capabilities needed to analyze the increasing volumes of data from heterogeneous sources, EDM researchers have suggested organizational capabilities to improve data quality. However, to date, little is known how companies actually orchestrate the allocated resources, especially regarding the quality and use of data to create value from BDA. Considering these gaps, this thesis – through five interrelated essays – investigates how companies adapt their EDM capabilities to create additional business value from BDA. The first essay lays the foundation of the thesis by investigating how companies extend their Business Intelligence and Analytics (BI&A) capabilities to build more comprehensive enterprise analytics platforms. The second and third essays contribute to fundamental reflections on how organizations are changing and designing data governance in the context of BDA. The fourth and fifth essays look at how companies provide high quality data to an increasing number of users with innovative EDM tools, that are, machine learning (ML) and enterprise data catalogs (EDC).

The thesis outcomes show that BDA has profound implications on EDM practices. In the past, operational data processing and analytical data processing were two “worlds” that were managed separately from each other. With BDA, these "worlds" are becoming increasingly interdependent and organizations must manage the lifecycles of data and analytics products in close coordination. Also, with BDA, data have become the long-expected, strategically relevant resource. As such data must now be viewed as a distinct value driver separate from IT as it requires specific mechanisms to foster value creation from BDA. BDA thus extends data governance goals: in addition to data quality and regulatory compliance, governance should facilitate data use by broadening data availability and enabling data monetization. Accordingly, companies establish comprehensive data governance designs including structural, procedural, and relational mechanisms to enable a broad network of employees to work with data. Existing EDM practices therefore need to be rethought to meet the emerging BDA requirements. While ML is a promising solution to improve data quality in a scalable and adaptable way, EDCs help companies democratize data to a broader range of employees.

Table of Contents

Introductory Paper on

BIG DATA AND ANALYTICS AS A NEW FRONTIER OF ENTERPRISE DATA MANAGEMENT

Martin Fadler

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

Table of Contents

List of Figures

List of Tables

List of abbreviations

3Vs	Volume, Variety, and Velocity
AI	Artificial Intelligence
BD	Big Data
BDA	Big Data and Analytics
BI	Business Intelligence
BI&A	Business Intelligence and Analytics
DQ	Data Quality
EAP	Enterprise Analytics Platform
EDC	Enterprise Data Catalog
EDM	Enterprise Data Management
EDW	Enterprise Data Warehouse
ETL	Extract Transform Load
IS	Information System
IT	Information Technology
ML	Machine Learning
OLAP	Online Analytical Processing
OLTP	Online Transactional Processing

1 Introduction

Data emerge with an ever-growing volume, velocity, and variety (3Vs), primarily from large-scale enterprise systems, online social graphs, open data, and the ever-increasing penetration of digitized devices and applications (Baesens, Bapna, and Marsden et al. 2016). The resultant Big Data, as it is commonly known, promise significant value generation opportunities across industries (George et al. 2014; Philip Chen and Zhang 2014). Analytics *"is arguably the engine – key to doing something valuable with the data"* (Lycett 2013, p.383). While business intelligence and data mining have been around for decades, the emergence of Big Data and major advancements in machine intelligence have led to an explosion of opportunities for analyzing data in ways that have not been possible before (Agarwal and Dhar 2014, p.443-444). Big Data and Analytics (BDA)[1] not only enhances additional business value by improving internal business processes and decisions, but also by augmenting existing products and services or selling data offerings with innovative data-driven business models (Wixom and Ross 2017). These value promises have *"created an attitude of collecting data without a pre-defined purpose, promoting a bottom-up, inductive approach to big data collection, exploration, and analysis"* (Günther et al. 2017, p.195). In reality, however, BDA initiatives fall short of expectations and companies experience problems generating the expected returns (Grover et al. 2018; Shim and Guo 2015). The main reasons for this shortfall seem to be organizational. Actually, a recent survey shows that mainstream companies generally experience no problems implementing BDA technologies, but that 92.2% of them struggle with management challenges such as organizational alignment or business processes (Bean 2021).

To date, researchers have addressed the essential changes and gained a fundamental understanding of BDA capabilities (Akter et al. 2016; Gupta and George 2016; Mikalef and Pappas et al. 2018) and value creation mechanisms (Grover et al. 2018). However, they also emphasize the lack of research – especially using explorative studies – on how companies actually build BDA capabilities and manage related resources. Aaltonen and Tempini (2014) emphasize the prerequisite that "*whether data-based business opportunities can be realized depends on an organizational capability to harness the potential embedded in newly available digital data*" (p. 108). Consequently, companies must adapt their enterprise data management (EDM) capabilities – being a firm's ability to deliver data in the appropriate quality to different data consumers and to adapt it to changing business needs and directions (Mithas et al. 2011).

[1] The term Big Data Analytics is used interchangeably with Big Data and Analytics (e.g., Grover et al. 2018), but it can also denote a specific type of analytics that uses very large datasets.

However, existing EDM approaches mostly focus on data quality and master data management and do not yet embrace the requirements emerging from BDA.

Considering these gaps, this thesis – through five interrelated essays – investigates how companies adapt their EDM capabilities to create additional business value from BDA. The first essay lays the foundation of the thesis by investigating how companies extend their Business Intelligence and Analytics (BI&A) capabilities to build more comprehensive enterprise analytics platforms. The second and third essays contribute to fundamental reflections on how organizations are changing and designing data governance in the context of BDA. The fourth and fifth essays look at how companies provide high quality data to an increasing number of users, that is, machine learning to improve data quality and enterprise data catalogs to broaden data use.

The remainder of this introductory essay to the thesis is structured as follows: First, relevant literature is summarized to explain the changes induced by BDA and to justify the need to adapt EDM capabilities. Thereafter, an overview is presented of the thesis by relating the essays to one another, and by explaining the general research setting. This is followed by a detailed account, summarizing the motivation, research approach, contributions, and implications of each essay. Finally, a discussion is presented of the overall conclusions derived from the thesis.

2 Background

2.1 The Evolution of Big Data and Analytics

Since the early beginnings of electronic data, digital data have been analyzed to improve businesses' efficiency and effectiveness. Currently, analytics encompasses traditional approaches to business intelligence (BI) as well as new ways of analyzing Big Data, thereby enabling sophisticated Artificial Intelligence (AI) applications (Watson 2019). The field has evolved through three major phases, in conjunction with the availability of different data sources (Chen et al. 2012).

During the first phase, in order to support decision making, companies mostly collected structured data about business transactions through legacy systems, such as enterprise resource planning, and primarily stored it in relational database management systems (RDMBS) (Chen et al. 2012). The first generation Decision Support System (DSS) used a dedicated data repository and model basis (Sprague 1980) to calculate the key performance indicators and to deliver reports on historic data in structured formats. This application-centric architecture was subsequently replaced by new DSS applications such as executive information systems and dashboards/scorecards (Watson 2014). Enterprise Data Warehouses (EDWs) allowed companies to integrate data from multiple operational systems in a pre-defined structure using Extract Transform Load (ETL) tools and to support a wide variety of applications simultaneously, such as queries, online analytical processing (OLAP), simple graphical visualizations, or data mining (Chen et al. 2012). The establishment of a central repository for all enterprise data had the advantage of simplifying analytics delivery (Watson 2009). As early as 1989, Howard Dresner coined the term business intelligence – BI as *"a broad category of applications, technologies, and processes for gathering, storing, accessing, and analyzing data to help business users make better decisions"* (Watson 2009, p. 491). EDWs allowed companies to process data in real time, thereby supporting business performance management with better decision making at both a strategic/tactical level and an operational level (Watson 2009). Most of these technologies have been integrated into commercial BI platforms and are commonly used by companies (Chen et al. 2012).

The second phase was initiated by the emergence of the Internet and the Web. Ever since, companies can present their business online and directly interact with their customers through web applications. Accordingly, companies collect web-based data and user-generated content, primarily in the form of semi-structured HTML documents (incl. text, images, and videos), and

server log files (Watson 2014). This form of data captures opinions, sentiments, and business information (i.e., industry, product, customer, and company) and allows enterprises to understand customer needs and to identify business opportunities in a timelier manner (Chen et al. 2012). However, organizations must integrate web analytics and text analytics, such as natural language processing, social network analysis, and spatial-temporal analysis, into their existing BI platforms. Since analytics technologies have increased in importance, the term (business) analytics is often used in conjunction or interchangeably with BI (Davenport 2006). In accordance with Chen et al. (2012), this thesis uses the term Business Intelligence and Analytics (BI&A) to cover both expressions in the analytical-usage context.

In the contemporary third phase, with the emergence of smartphones and the ubiquity of sensors embedded in connected devices, data are collected on a more granular level than previously. The additional data allow enterprises to accurately trace and analyze their business operations, but it also requires them to rethink the way they manage data and deliver analytics products. Currently, data *"are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization technologies"* (Chen et al. 2012, p. 1166). Traditional EDWs cannot cope with these requirements, due to their lack of flexibility in terms of modifying data structures and dealing with multiple data formats (Jukić et al. 2015; Sivarajah et al. 2017). Companies are therefore extending their existing data infrastructures with data lakes in order to build more comprehensive platforms. Data lakes store data without a pre-defined structure, which allows them to explore and experiment with data in a more flexible and efficient way (Farid et al. 2016; Madera and Laurent 2016; Watson 2017) and to deliver analytics products with functionalities that clearly go beyond the mere aggregation and visualization of data, and that also comprise AI (Watson 2017).

Considering these technical developments, Big Data and Analytics (BDA) is used as an umbrella term that reflects the phenomenon of ever-increasing data volume, variety, and velocity (3Vs), as well as modern techniques and technologies to store, process, and analyze the 3Vs with BI&A.

2.2 Value Creation with Data

BDA provide companies various, important opportunities for data monetization. However, companies struggle to gain a return on increased investments, being an issue that raises general questions on how they can foster value creation from their investments in BDA resources.

Data's inherent value potential can be attributed to its unique property of being non-rivalrous, which means that it can be used for multiple purposes simultaneously (in contrast to physical resources) without losing its value (Wang 1998). Accordingly, data can generate business value in different ways. With contemporary data ecosystems, companies can monetize data in direct and indirect ways (Wixom and Ross 2017). First, companies use data to improve internal business processes and decision-making quality. In this regard, they use key performance indicators to improve decision-making quality or predictive models, for example, to understand future events and to trigger business processes more effectively. Second, companies use data to augment existing products and services (called "wrapping"). In this case, enterprises often provide dashboards or analytical services in conjunction with their physical offerings to enrich the overall value proposition and to differentiate it from their competitors. Third, companies sell their data offerings to new and existing markets. Here, they establish dedicated data-driven business models and deliver data services that often provide better scalability than their physical opponents.

In order to create business value, companies must invest in data-related resources, namely technological, human, and organizational resources (Gupta and George 2016; Mikalef et al. 2018). However, according to the resource-based view (RBV) (Penrose and Pitelis 1959; Wernerfelt 1984), the mere possession of resources does not lead to value creation. Instead, companies must manage their resources by "*structuring the firm's resource portfolio, bundling the resources to build capabilities, and leveraging those capabilities with the purpose of creating and maintaining value for customers and owners*" (Sirmon et al. 2007, p.273). Aaltonen and Tempini (2014) emphasize that "*whether data-based business opportunities can be realized depends on an organizational capability to harness the potential embedded in newly available digital data*" (p. 108). Thus, value creation is a complex process that entangles physical, human, and organizational elements to create synergistic effects (Kohli and Grover 2008). Accordingly, building the capabilities that enable value creation from data is considered a key sucess factor (Akter et al. 2016; Grover et al. 2018; Gupta and George 2016; Legner et al. 2020; Mikalef et al. 2018; Mithas et al. 2011; Zeng and Glaister 2018). Literature on BDA and EDM that explored corresponding capabilities emphasizes complementary aspects in managing data resources.

2.3 Big Data and Analytics Capabilities

BDA literature emphasizes value creation from data assets ("Big Data") through the use of statistical, processing, and analytics techniques to create value (Grover et al. 2018). The corresponding mechanisms have been investigated through different theoretical lenses with varying scope, including business intelligence systems (Trieu 2017), business analytics (Seddon et al. 2017) and, most recently, BDA (Grover et al. 2018). These studies reveal that value creation involves complex processes to obtain the actual value from initial investments in analytics resources. Grover et al. (2018) underline that "*converting IT investment in BDA to valuable capabilities is a dynamic process that involves identification of where, how, and what value will be created. Capabilities include the ability to both manage and analyze data to create new insights*" (p. 397). In the long term, companies can develop BDA as a strategic resource that is valuable, rare, hard to imitate, and organizationally embedded (VRIO-framework), and that sustains a competitive advantage (Grover et al. 2018; Gupta and George 2016; Mikalef et al. 2018). While consensus reigns about the importance of the BDA capability, the specific constituents of the BDA capability and how they are constructed are less clear (Akter et al. 2016; Grover et al. 2018; Gupta and George 2016; Mikalef and Pappas et al. 2018; Zeng and Glaister 2018). Akter et al. (2016) describe the BDA capability as a multi-level construct of interwoven BDA technology, talent, and management capabilities.[2]

The *BDA technology capability* refers to the flexibility of the BDA platform that enables BDA professionals to quickly develop, deploy, and support a firm's resources (Akter et al. 2016). Furthermore, Grover et al. (2018) highlight the need for a platform "*collecting, integrating, sharing, processing, storing, and managing big data*" (Grover et al. 2018, p. 399). They further emphasize that companies must actively manage a portfolio of different analytics methods (e.g., text or social media analytics). This portfolio encompasses descriptive, predictive, and prescriptive analytics techniques, aligned with the needs of business.

The *BDA talent capability* refers to the ability of a BDA professional to perform a task in the BDA environment (Akter et al. 2016). This ability depends on technical knowledge (e.g., database management), technology management knowledge (e.g., visualization tools and management and deployment techniques), business knowledge (e.g., the understanding of short-term and long-term goals), and relational knowledge (e.g., cross-functional collaboration using information) (ibid). Grover et al. (2018, p. 399) emphasize that "*without the right group of skilled big data experts (e.g., data scientists, big data engineers, and big data architects), it is impossible*

[2] By adapting the seminal IT capability model of Kim et al. (2012)

to develop and carry out a BDA strategy," and observe that this requirement remains one of the greatest challenges for firms. In addition to technical professionals, Zeng and Glaister (2018) emphasize that, in particular, it is the manager's ability to democratize, contextualize, and experiment with data in a collaborative process, and to build an effective organizational structure that makes an important difference in the value creation process (p.43).

The *BDA management capability* refers to the ability to handle routines (e.g., structures, policies, and decision making) in a structured way to manage BDA resources according to business needs and priorities (Akter et al. 2016; Kim et al. 2012). This capability encompasses BDA planning, investment, coordination, and control routines (Akter et al. 2016), and is executed by the BDA organization as an independent unit or is distributed to each key business function within an enterprise (Kim et al. 2012). Grover et al. (2018) emphasize: *"Without appropriate organizational structures and governance frameworks in place, it is impossible to collect and analyze data across an enterprise and deliver insights to where they are most needed"* (p. 417). Mikalef et al. (2018) underline the importance of data governance to account for data's growing strategic importance. Although parallels exist with research on BI, Phillips-Wren et al. (2015) argue that BDA implies changes to its management and governance that need to be further explored, as called for by several researchers (Abbasi et al. 2016; Goes 2014; Grover et al. 2018; Hassan 2019).

The presented BDA capability model (Akter et al. 2016) enables a systematic, albeit a rather abstract analysis of value creation and its business implications. Nevertheless, it is only a first structuring step to eventually gaining a more in-depth picture. The way companies actually transform their organizations and build a BDA capability remains unexplored (Akter et al. 2016; Grover et al. 2018; Gupta and George 2016; Mikalef et al. 2018). According to the research framework of Grover et al. (2018), the process of capability building has received considerably less attention than the process of capability realization. Consequently, "*little is known so far about the processes and structures necessary to orchestrate these resources into a firm-wide capability*" (Mikalef et al. 2018, p. 569). Due to BDA's strategic significance, Sivarajah et al. (2017) also call for "*further in-depth conceptual as well as empirical, especially case study and survey based research studies*" (p.15) as their literature review reveals that most studies are of an "*analytical nature*" (p.16).

2.4 Enterprise Data Management

While BDA emphasizes data's use for analytics, Enterprise Data Management (EDM)[3] draws attention to the "*ability to provide data and information to users with the appropriate levels of accuracy, timeliness, reliability, security, confidentiality, connectivity, and access and the ability to tailor these in response to changing business needs and directions*" (Mithas et al. 2011, p.238). EDM influences the development of performance management, customer management, and process management capabilities which are antecedents of superior organizational performance (ibid, p. 240). Thus, EDM has been attributed an enabler function with a focus on providing quality data as a prerequisite for value generation. The research field's perspective has evolved along with technological progress and the changing role of data in companies, from a database-centric view to a strategic view (Legner et al. 2020).

Research on EDM has emerged with the proliferation of the RBV in the 1980s. Goodhue et al. (1988) were among the first to address the problem of "*unmanaged*" data and emphasized the consideration of data as a valuable business resource. Several studies have typically regarded data "*as 'raw material' that needs to be processed to become information (Ahituv, 1989; Badenoch, Reid, Burton, Gibb, & Oppenheim, 1994) within the 'information lifecycle' (Levitan, 1982)*" (Otto 2015, p. 234). Analogous to other firm's assets, the quality management of data resources has been revealed to be a key driver of value creation and has become a dominant focus in the research discipline. In his seminal paper, Wang (1998) introduced the Total Data Quality Management Methodology (TDQM) that conceptualizes the delivery of high quality data products in organizations. Analogous to established concepts derived from manufacturing, raw data are transformed through information systems into data products which are delivered to data consumers, for example, customer account data. Data quality is defined as data that has "fitness for use" by data consumers and comprises multiple dimensions, namely intrinsic, contextual, representational, and accessibility (Wang and Strong 1996). Hence, "*high-quality data should be intrinsically good, contextually appropriate for the task, clearly represented, and accessible to the data consumer*" (ibid, p.6). While the aforesaid work has primarily proliferated into further studies on methods to assess data quality, other studies have investigated how companies build EDM capabilities to manage data quality enterprise-wide (Legner et al. 2020). Key themes in this stream are data governance and master data management (ibid). Master data management aims to improve the data quality of a company's core business entities, for example,

[3] The term enterprise data management (EDM) is used to emphasize the organizational character of data management practices and, when used with the same intent, is used interchangeably with the terms information management and corporate data management.

product and customer, and to create a consistent representation across operational systems (Otto 2015; Otto et al. 2012; Smith and McKeen 2008). These data types are regarded as a company's most valuable data resources as they provide the foundation to run business processes more efficiently, to improve business agility, and to enhance decision-making quality and compliance reporting. With this focus and the increasing strategic relevance of data, companies have essentially started to build more thorough organizational capabilities by establishing data governance. According to Weber et al. (2009): "*Data governance defines roles, and it assigns responsibilities for decision areas to these roles. It establishes organization-wide guidelines and standards for DQM [(data quality management)], and it assures compliance with corporate strategy and laws governing data*" (p. 2). Hoven (1999) already anticipated the role of the data steward as the "*person responsible for ensuring the effective and efficient management and utilization of the enterprise's data resources*" (pp. 88-89), which remains a core, contemporary data governance role. Other researchers have investigated different data governance mechanisms to build organizational structures that improve data quality enterprise wide and that manage data lifecycles in a consistent manner across systems (Khatri and Brown 2010; Korhonen et al. 2013; Otto 2011a, 2011b; Tallon et al. 2013; Weber et al. 2009).

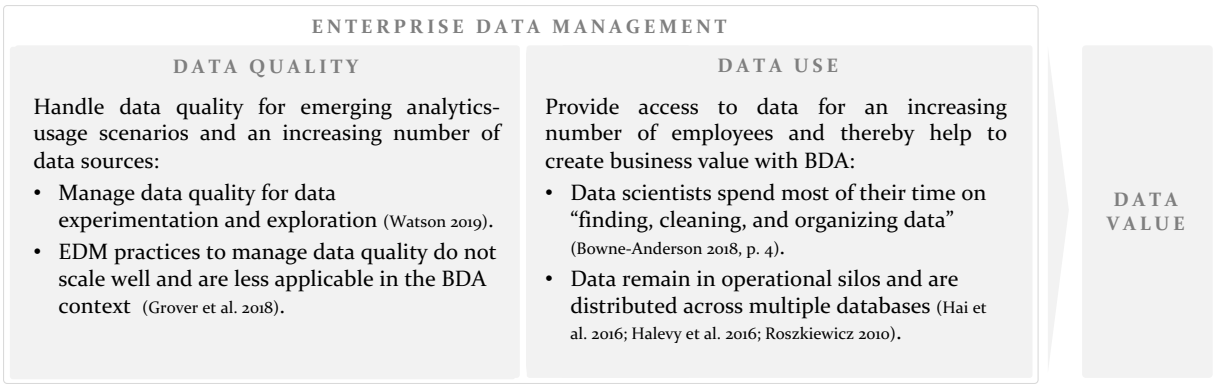


Figure 1. Research gap

In conclusion, prior literature on EDM focuses primarily on improving the quality of data residing in operational systems and, consequently, needs to be revisited to cope with BDA (see Figure 1). BDA comes with different data management principles than those used in traditional operational systems and BI environments. While the purpose of data is known in operational systems and BI environments and major efforts are invested in upfront data integration to make data fit for use, these efforts are kept to a minimum in data lakes in order to increase flexibility in data processing and to identify new, previously unknown purposes. This change has implications for and poses fundamental challenges to EDM. As a result, data scientists spend most of their time on, among others, "*finding, cleaning, and organizing data*" (Bowne-Anderson

2018, p. 4). A technical reason for this is that data often remain in operational silos and are distributed across multiple databases (Hai et al. 2016; Halevy et al. 2016; Roszkiewicz 2010). EDM must also account for the changing role of data and place a greater emphasis on the enablement of data monetization in addition to control and compliance. As the companies' demand for data increases, EDM must rethink the existing (mostly manual) practices and find novel approaches to manage data quality at scale. Grover et al. (2018) stress the importance of handling data quality, data integration, and data security to create a valuable BDA asset. They also call for adaptable and scalable approaches to improve data quality.

3 Dissertation Overview

3.1 Research Objectives

This dissertation addresses key requirements from BDA that need to be considered by EDM (see Figure 2). On the data provisioning side, EDM must handle data quality for an increasing number of data sources and emerging analytics-usage scenarios, for example, ML. On the data usage side, EDM must provide access to data for an increasing number of employees and thereby help to create business value with BDA. This thesis investigates these BDA requirements through five essays:

Essay I lays the foundation for the thesis and explores how enterprises build BI&A capabilities. Data and analytics governance are the topic of Essay II and Essay III. Essay II investigates how enterprises assign fundamental decision rights to data, that is, data ownership, in the BDA environment. Essay III extends this view and investigates how companies adapt their data governance designs, among others, to support data monetization. From a tool perspective, Essay IV and V look into new approaches to manage data quality and data use. Essay IV explores how ML techniques help to improve data quality and thereby reduce the high level of manual efforts. Essay V investigates enterprise data catalogs as emerging platforms for data democratization that supports technical as well as business professionals in finding, accessing, and using data.

Figure 2. Overview of essays

3.2 Research Setting

This thesis was conducted within a consortium research project (Österle and Otto 2010) (see Figure 3) in the EDM domain, the Competence Center Corporate Data Quality (CC CDQ). The CC CDQ brings together industry experts from about 20 multi-national corporations and a research team to work on problems with significant practical relevance. Consortium research is based on the collaborative practice research approach and involves close collaboration between researchers and practitioners to serve "*general knowledge interest as well as knowledge interests that are specific for the participating organisations*" (Mathiassen 2002, p.10). This collaboration requires that researchers generate and communicate results, which are subsequently evaluated by practitioners and further improved based on their feedback (Mathiassen 2002). Through this course of action, the research outcomes are not only specifically relevant to the involved practitioners, but they also contribute to the development of professional practices in general (ibid). Moreover, consortium research adopts a design science research paradigm (Hevner et al. 2004) to develop artifacts and to produce generalizable research outcomes with scientific rigor (Österle and Otto 2010).

Consortium research has been particularly successful in the interdisciplinary field of EDM that straddles the cross-section of computer science, information systems, and management research (Legner et al. 2020). In this setting, knowledge is accumulated in both research and practitioner communities and the close collaboration in the CC CDQ allows to develop relevant research results using rigorous research methods (ibid). In the CC CDQ, the practitioners are EDM experts and represent companies with varying EDM maturity and from diverse industries, for example, pharmaceuticals, automotive, fast-moving consumer goods, public transportation, chemicals, or sportswear. The companies also have varying levels of BDA maturity, and some of them have already established data lakes and used advanced analytics in production, whereas others are only beginning with and are in a pilot stage of BDA. For the thesis, this setting is ideal to determine how companies adapt their EDM capabilities to foster value creation from BDA, which is a complex socio-technical phenomenon that has not yet been intensively researched. In the consortium setting and from the available repertoire, the researcher must select approaches - including their corresponding methods - that suite the situation of inquiry. In the context of this thesis, mostly explorative, qualitative research approaches were chosen in order to account for the novelty of the underlying research problem. In all essays, especially to ground the artifact design, desk research was used to gather scientific and practitioner knowledge. Case studies were applied in Essays I – III to gain a fundamental understanding of how companies adapt their organizations. Focus groups with a selected group of experts were used in all essays

to discuss specific topics and evaluate intermediary research outcomes. Expert interviews were conducted in all essays to gain an in-depth understanding of individual company situations, which were complemented with the analysis of company specific documents. All research topics were addressed in plenary discussions, simultaneously involving more than 30 industry experts, mostly to present final research outcomes. In Essay V, the researchers were also partially involved in the projects, which allowed them to instantiate artifact versions and gather feedback in real-world settings. Furthermore, this essay used a survey to evaluate an intermediary artifact version.

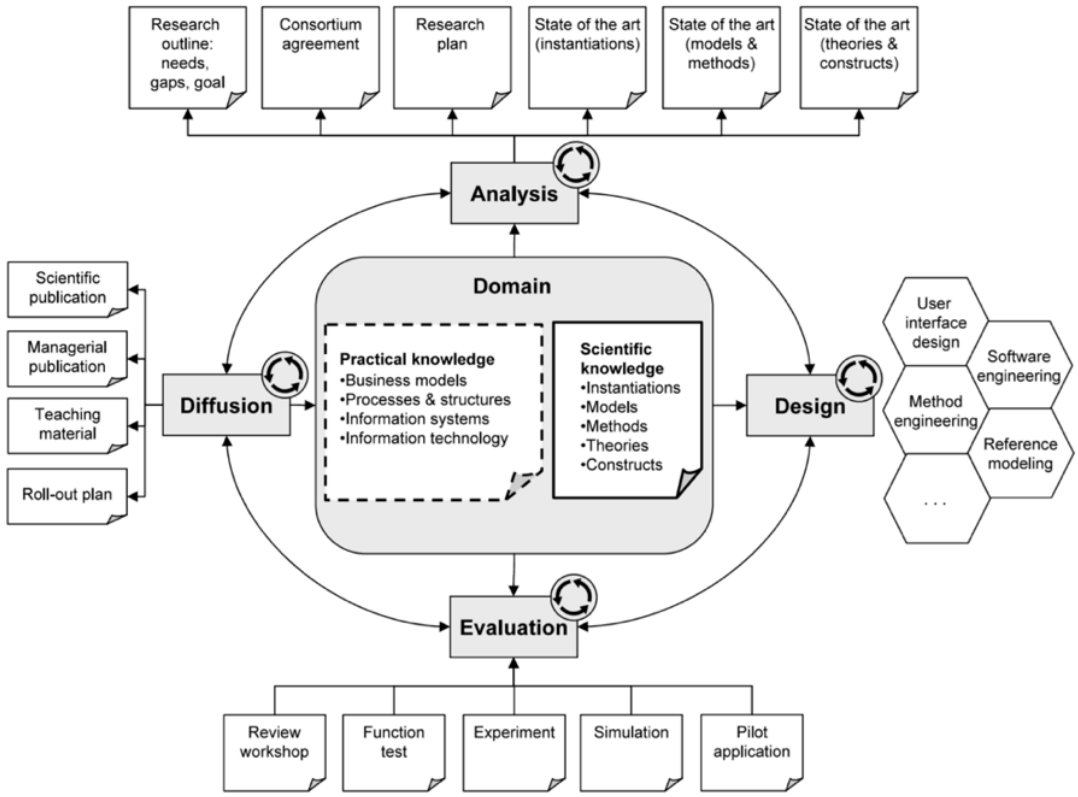


Figure 3. Consortium research overview (adapted from Österle and Otto 2010)

Table 1. Dissertation structure and essays

Essay	Research question(s)	Research method(s)	Key contributions	Publication status
Essay I: Building Business Intelligence and Analytics Capabilities - A Work System Perspective	How do enterprises build Business Intelligence and Analytics capabilities?	Multi-method research design: case studies (4 cases), expert interviews, document analysis, focus groups	Business Intelligence and Analytics capabilities and work systems	Presented at International Conference on Information Systems (2020)
Essay II: Data Ownership Revisited: Clarifying Data Accountabilities in Times of Big Data and Analytics	How do enterprises define data ownership in the context of Big Data and Analytics?	Multiple case study research: focus groups, expert interviews (4 cases)	Data ownership types and principles	a: First version presented at European Conference on Information Systems (2020) b: Extended version published in *Journal of Business Analytics (JBA) (2021)*
Essay III: Data Governance: From Master Data Quality to Data Monetization	RQI: How do companies design data governance using structural, procedural, and relational mechanisms? RQII: How do companies implement data governance to address the changing role of data?	Multiple case study research: focus groups, expert interviews, document analysis (9 cases)	Data governance mechanisms and archetypes	a: First version presented at European Conference on Information Systems (2021) b: Extended version for submission to an IS journal
Essay IV: Machine Learning Techniques for Enterprise Data Management: A Taxonomic Approach	RQI: Which elements describe ML techniques for enterprise data management? RQII: Which archetypes of ML for enterprise data management can be distinguished?	Taxonomy development: case studies, expert interviews, focus groups (Nickerson et al. 2013)	ML for data management taxonomy and archetypes	a: First version presented at Pre-ICIS SIGDSA (2019) b: Extended version for submission to an IS journal
Essay V: All Hands on Data: A Reference Model for Enterprise Data Catalogs	What are the main constituents of an Enterprise Data Catalog as emerging platforms for data democratization?	Design science research (Peffers et al. 2007): focus groups, expert interviews, document analysis, projects, surveys	Reference model for enterprise data catalogs	Extended version for submission to an IS journal

4 Essay Summary

4.1 Essay I: Building Business Intelligence and Analytics Capabilities – A Work System Perspective

4.1.1 Motivation

While prior research mainly focused on Business Intelligence and Analytics (BI&A) capabilities and explained the value creation steps (see 2.3), the focus was much less on building these capabilities. Among the few studies, Schüritz et al. (2017) investigate analytics competence centers to depict organizational design patterns, while Kettinger et al. (2019) explore information management capability building and develop guidelines for senior executives. Although both studies address BI&A capability building, they do so with a focus on partial aspects and with a different research aim. In conclusion, existing research on BI&A capability building is fragmented and lacks a clear theoretical framing to understand the constituents of the emerging analytics platforms and their value creation mechanisms.

This essay proposes work system theory (WST) as a theoretical lens to study capability building in the context of BI&A. A work system is a "*system in which human participants and/or machines perform work (processes and activities) using information, technology, and other resources to produce specific product/services for internal or external customers*" (Alter 2013). The work system framework facilitates an understanding of how resources (participants, information, and technologies) are orchestrated (by means of processes/activities) to build capabilities (products/services for customers). Several researchers have used WST to analyze specific BI&A applications (e.g., Alter 2004; Heart et al. 2018; Marjanovic 2016), which enhances confidence in the use of the WST lens to systematically analyze how enterprises orchestrate their tangible and intangible BI&A resources and build BI&A capabilities.

4.1.2 Research objectives and approach

This essay aims to investigate BI&A capability building in enterprises and asks the following research question:

RQ: *How do enterprises build Business Intelligence and Analytics capabilities?*

This study applied a multi-method research design (Venkatesh et al. 2016) comprising expert interviews and focus groups. As part of the consortium research program, an expert group was formed in February 2019 encompassing 11 BI&A experts from seven high-profile European

companies to investigate Enterprise Analytics Platform (EAP) challenges over a period of one year (5 meetings). The experts are responsible for defining governance structures (including the definition of roles, responsibilities, and processes) and are familiar with the different BI&A initiatives in their respective companies. This constellation allowed the collection of unique field data from ongoing BI&A initiatives in European companies and the gaining of a broader understanding of the current state of BI&A in enterprises. Data collection was done through four case studies and five focus group meetings. Thereafter, the collected data were analyzed through the WST lens.

Case selection

For the case studies, four (of the seven) companies (see Table 2) with a relatively high, comparable BI&A maturity were selected following literal replication logic (Benbasat et al. 1987; Yin 2003). All four companies have a BI&A infrastructure that includes, beside an EDW, an established data lake, the incorporation of data scientist teams to perform data experiments, and the implementation of data governance mechanisms to control the value creation process.

Theoretical integration with work system theory

The study uses WST (Alter 2013) to integrate the collected data. WST consists of three components: the work system definition (see Motivation), the work system framework, and the work system lifecycle model (see Figure 4). While the work system framework provides a snapshot of a certain point in time, the work system lifecycle model describes how a work system evolves over time along four phases: *Initiation*, *Development*, *Implementation*, and *Operation and Maintenance*. In order to structure the processes and the activities, the snapshot is used as the main structure to integrate data about the general capability building process and the lifecycle.

Table 2. Case companies

Company	Industry	Revenue/ # Employees	Key informants	BI&A context
A	Consumer goods	$50–100 b / ~80 000	Data governance manager, enterprise data architect	Organization: central data and analytics management organization Infrastructure: central Big Data platform for the innovation and industrialization of analytics use cases
B	Public transportation	$1–50 b / ~35 000	Leader business information management, data governance manager, Big Data platform architect	Organization: central data management organization and central/decentralized data science team Infrastructure: corporate data lake for data exploration/experimentation and the operation of analytics use case
C	Industry products	$50–100 b / ~110 000	Project manager data lake	Organization: Central data management organization and advanced analytics group Infrastructure: Operation of multiple data lakes and data warehouses
D	Consumer goods	$1–50 b / ~30 000	Head of data and analytics, head of data governance	Organization: Central data and analytics management organization with high business intelligence maturity Infrastructure: Operation of a central enterprise data warehouse with extensions to undertake analytics

Data collection

For each case company, data was collected through semi-structured interviews and additional company material on their BI&A platform designs, role models, and organizational structures. Triangulation of primary (interviews) and secondary (company materials) sources ensured construct validity (Yin 2003).

The focus groups helped to evaluate and further improve the findings through BI&A work systems. This was done by reflecting, in the context of the literature and with a broader group of experts, on the four cases. The experts met five times in person between February 2019 and February 2020, to investigate in depth – during each meeting – an essential topic.

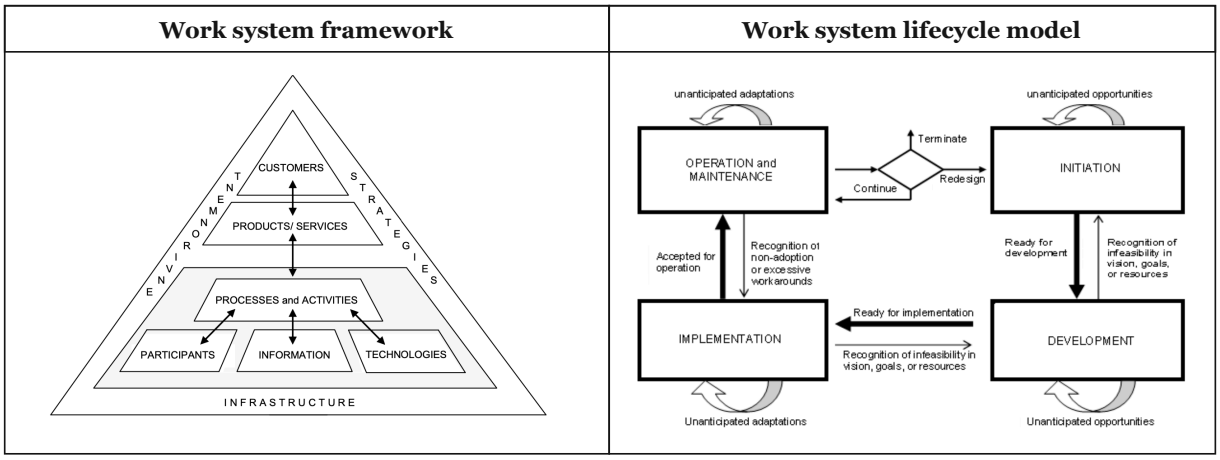


Figure 4. Work system framework and lifecycle (Alter 2013)

4.1.3 Research outcomes and contributions

Four BI&A capabilities (i.e., *Reporting*, *Data exploration*, *Analytics experimentation*, and *Analytics production*) are identified and the capability building process is explained through WST. Each capability and the building process are summarized as follows:

Reporting

The *Reporting* capability creates value through *Transparency and access* and *Continuous monitoring and proactive adaptation*, according to the value creation mechanisms defined by Grover et al. (2018). Hereby, the capability equips business users – periodically or in real time – with reports and digital dashboards summarizing the business transactions or machine data in the form of key performance indicators and visualizations (Chen et al. 2012; Watson 2009). The following describes the typical phases of building this BI&A capability: The *Initiation* phase starts with the specification of the business user's information needs (e.g., decisions to be made, delivery format, and frequency) by an analytics expert. In the *Development* phase, either the analytics expert or a data analyst develops the report by means of a BI tool, which is done in multiple iterations with the business user. However, depending on data availability, a data architect and a data engineer may be included in this step to onboard, at first, the data to the data warehouse. In the *Implementation* phase, the data engineer deploys the report, business users are trained to use the report, and the report is documented in a data catalog. In the *Operation and Maintenance* phase, a data engineer monitors the ETLs, while a data steward monitors the data quality in general and the report's use in particular.

Data exploration

The *Data exploration* capability creates value through *Discovery and experimentation* and *Continuous monitoring and proactive adaptation*, according to the value creation mechanisms

defined by Grover et al. (2018). Hereby, a business user gains access to data of their domain of interest, in a dedicated environment that allows in-depth and flexible data analysis "*without an a priori understanding of what patterns, information, or knowledge it might contain*" (Baker et al. 2009, p. 534). A typical way how this BI&A capability is built is as follows: In the *Initiation* phase, business users specify their data requirements with the support of an analytics expert. In the *Development* phase, the required data are onboarded to the data warehouse (in case it is not yet available) and loaded into a data mart. Here, a data architect models the required data and a data engineer implements the required ETLs. Next, data access is granted to the user for the BI tool and the data. In the *Implementation* phase, the business users are trained by data analysts and analytics experts to use the tool and the data. The analytics expert documents the data and training material in a data catalog. In the *Operation and Maintenance* phase, data are continuously pushed to the access tool, which is monitored by the data engineer. A data steward takes data quality measures and ensures that the data remain fit for use.

Analytics experimentation

The *Analytics experimentation* capability creates value through *Discovery and experimentation* and *Learning and crowd-sourcing*, according to the value creation mechanisms defined by Grover et al. (2018). The capability equips data scientists with a virtual sandbox environment to develop and test analytics algorithms in a virtual sandbox environment (Watson 2014). Accordingly, datasets are made available in their raw format without integrating them in a predefined structure that allows flexible data repurposing. A typical way of building this BI&A capability is as follows: In the *Initiation* phase, the analytics use case is specified and a business case is calculated by a team comprising an analytics expert (domain knowledge), a data architect (data knowledge), and a data scientist (analytics knowledge). In the *Development* phase, the data architect identifies and models the required data, and the data engineer onboards the data to the data lake (in case the data are not yet available). The newly onboarded data are documented in the data catalog by the relevant data stewards and data engineers. In the *Implementation* phase, the data engineer grants data access to the data scientist in a dedicated sandbox environment. In the *Operation and Maintenance* phase, the data scientist tests the feasibility of different algorithmic approaches in multiple iterations. However, a data engineer might change or onboard more data.

Analytics production

The analytics models that prove feasible are deployed and made accessible with the *Analytics production* capability. This capability creates value through *Prediction and optimization* and

Customization and targeting, according to the value creation mechanisms defined by Grover et al. (2018). Hereby, a business user accesses an analytics model (predictive/prescriptive) in business applications. The *Analytics production* capability enables the efficient and effective deployment of analytics models and ensures that they generate business value throughout their lifecycles (Watson 2019). A typical way of building this BI&A capability is as follows: In the *Initiation* phase, a system engineer, a data architect, the responsible data scientist, and an analytics expert review the analytics models and specify the requirements for the production. In the *Development* phase, the analytics model is optimized for production by a developer (e.g., an ML engineer), a system engineer designs the application architecture, a data architect provides the data models, and a data engineer implements the ETLs. Documentation and training material (if necessary) are created accordingly. In the *Implementation* phase, business users are trained to use the analytics model. In the *Operation and Maintenance* phase, business users use the analytics model in business applications. The data engineer continuously monitors the analytics model's quality. In the case of changes in the underlying data distribution (concept drift), the data scientist may need to newly optimize the model.

Interestingly, several commonalities are observed across the BI&A work system components (i.e., participants, technologies/infrastructures, and processes/activities). These commonalities are integration possibilities that lead to synergistic effects between the identified BI&A capabilities. This means that companies may lose synergies if they manage their existing BI environments (i.e., *Reporting* and *Data exploration*) and emerging Big Data infrastructures (i.e., *Analytics experimentation* and *Analytics production*) separately. The case analysis reveals that managing the four work systems as an integrated EAP can create benefits at organizational and infrastructure levels and help build superior BI&A capabilities.

Table 3. BI&A capabilities in case companies

BI&A capability		**Reporting**	**Data exploration**	**Analytics experimentation**	**Analytics production**
Value creation mechanism (Grover et al. 2018)		Transparency and access/Continuous monitoring and proactive adaptation	Discovery and experimentation/ Transparency and access	Discovery and experimentation/ Learning and crowd-sourcing	Prediction and optimization/ Customization and targeting
Customers		Business user	Business user	Data scientist	Business user
Products/Services		Periodically providing reports or real-time updating of dashboards summarizing business transactions in the form of key performance indicators and visualizations (Chen et al. 2012; Watson 2009).	Providing an environment to explore and make sense of data in a certain domain of interest, without an a priori understanding of what information it may contain on the issue being investigated (Alpar and Schulz 2016; Baker et al. 2009).	Providing a virtual sandbox environment to develop and test analytics use cases and to prove their feasibility (Watson 2014).	Providing an up-to-date analytics model in a business application (Watson 2019).
Exemplary processes/activities	**Initiation**	• Report specification	• Requirement specification	• Use case specification and prioritization	• Requirement specification
	Development	• Data onboarding • Report development	• Data onboarding • BI tool setup • Creation of training material	• Data onboarding • Configuration of sandbox environment • Documentation	• Analytics model optimization • Architecture design and implementation
	Implementation	• Deployment • Training • Documentation	• Access provisioning • Training • Documentation	• Access provisioning • Support in use case understanding	• Training of business users
	Operation and maintenance	• Monitoring of ETL, data quality and report use	• Monitoring of ETL and data quality.	• Analytics model development • Data engineering	• Monitoring of analytics model quality • Maintenance of analytics model
Exemplary participants		Business user Analytics expert Data architect		Data scientist Data engineer Data steward	
Exemplary information		Historic/Real time Predefined structure Domain knowledge		Raw format Reference data Pretrained models	
Exemplary technologies		BI tools Extract, transform, load Data catalog		Interactive computing tools Sandbox environment Monitoring tools	

4.1.4 Conclusion

While existing research on BI&A value creation mainly focuses on the mechanisms and the process steps, this essay sheds light on the inner mechanisms of BI&A capability building and the structure of emerging EAPs. The study proves that WST provides a systematic approach to explain capability building and to identify commonalities. With the four identified BI&A capabilities and their corresponding work systems, this essay provides a comprehensive understanding of the emerging EAPs' components and their synergistic effects when managing them in an integrated way.

From a practitioner perspective, the research outcomes help not only to understand the essential BI&A resources and their interplay, but also to analyze the current situation and to define an appropriate organizational and infrastructure setup for building and managing EAPs.

From an academic perspective, the research findings contribute to capability building as a prerequisite for value generation from BDA. Therefore, this essay addresses important questions outlined in the BDA research agenda by Grover et al. (2018), related to the creation of analytics capabilities, that is, the ability to integrate, disseminate, explore, and analyze Big Data. In this field, several promising research opportunities are identified, which relate to all four BI&A work systems and their integration into an EAP. For instance, the transition from *Analytics experimentation* to *Analytics production* remains a challenge in practice and requires an in-depth analysis.

4.2 Essay II: Data Ownership Revisited: Clarifying Data Accountabilities in Times of Big Data and Analytics

4.2.1 Motivation

Data ownership clarifies fundamental rights and responsibilities for data and has been discussed since the beginning of electronic data processing (Maxwell, 1989; Spirig, 1987; Van Alstyne et al., 1995; Wang et al., 1995). Studies have investigated data ownership for operational systems and data warehouses (Winter and Meyer 2001), where the purpose of data processing is known (e.g., *Reporting* capability in Essay I). While the assignment of data accountabilities is still required in the contemporary corporate environment, the emerging data lake infrastructures require a different approach to data governance (Chessell et al. 2018). So does the *Analytics experimentation* capability (see Essay I) that, for instance, assume that data are used for new, previously unknown purposes. With this change, the definition of accountabilities for data is more challenging because data flow across organizational units to match different data consumer' requirements in respect of data format, granularity, and quality. Such cross-unit data flows require effective coordination (Dinter, 2013; Winter, 2008). These developments raise the question how we need to reinterpret and apply data ownership concepts to cope with these emerging challenges in EAPs.

4.2.2 Research objectives and approach

This essay aims to understand how data ownership concepts change in the context of BDA and therefore addresses the following research question:

RQ: *How do enterprises define and adapt data ownership in the Big Data and Analytics context?*

In order to analyze such a complex phenomenon in its specific context, this study performed an extensive literature review and conducted explorative research based on multiple case studies (Benbasat et al. 1987; Yin 2003). The selected research design is ideal to answer *how* questions (Yin 2003) and to analyze rich information related to the adoption of BDA and the definition of data-related roles in enterprises. The investigation of multiple case studies ensures theory's robustness and the ability to draw generalizable conclusions (Benbasat et al. 1987; Yin 2003).

Case selection

Four (out of seven) companies (see Table 4) were selected from the expert group described in Essay I, following literal replication logic (Benbasat et al. 1987; Yin 2003). All four companies had

established a data lake and had introduced corresponding data and analytics roles, including the concept of data ownership. However, they stem from different industries.

Data collection

Within each case company, one semi-structured interview was conducted with the key informants from the company to understand the technological and organizational structures to manage BDA. The data were complemented by additional company material (e.g., BDA platform designs, role models, and organizational structures) to triangulate the collected data and to ensure construct validity (Yin 2003).

Table 4. Selected cases

Case name	Industry	Size	Key informants	Big Data and Analytics context
Company A	Fast-moving consumer goods	Revenue: $50B to $100B Employees: ~80 000	Manager: Data governance, Enterprise data architect	Organization: central data and analytics management organization Infrastructure: central Big Data platform for the innovation and industrialization of analytics use cases
Company B	Public transportation and mobility infrastructure	Revenue: $1B to $50B Employees: ~35 000	Leader: Business information management, Data governance manager, Big Data platform architect	Organization: central data management organization and central/decentralized data science team Infrastructure: corporate data lake for data exploration/experimentation and the operation of analytics use case
Company C	Manufacturing	Revenue: $1B to $50B Employees: ~90 000	Director: Data architecture and engineering, Project manager: Data platform	Organization: corporate data management organization and central platform team Infrastructure: central data platform to enable digital innovations and to scale the operation of data products
Company D	Healthcare and life science	Revenue: $1B to $50B Employees: ~50 000	Leaders: Head of Data Products and Solutions, Global Enterprise Data Strategy Lead	Organization: federated organization with a data and analytics center of excellence and staff in line of business Infrastructure: multiple data platforms serving specific analytics needs and an enterprise-wide data platform

Data analysis

The analysis was conducted in two steps. First, a within-case analysis was conducted to clarify the different data ownership types in each company. Second, a cross-case analysis was performed to identify common data ownership types and the responsibility of each.

4.2.3 Research outcomes and contributions

The study reveals that BDA leads to significant changes and extensions of data ownership. From the analysis of the literature and the cases, six propositions are identified that explicate three data ownership types and describe the implications of data repurposing. The propositions are summarized in the following:

Data ownership types

The first proposition explicates the different data ownership types that are prevalent in BDA environments.

Proposition 1: In the context of BDA, companies define data ownership at three levels: data source or dataset (data supply), data product (data demand), and data platform.

These three data ownership types are prevalent in the case companies (see Table 5) and explain the fundamental rights and responsibilities to manage data in BDA platforms. The succeeding propositions describe the responsibilities associated with each data ownership type.

Proposition 2: The data owner ensures compliant access to and use of data, not only in the source system, but also on the platform and in data products. This addition extends beyond the traditional scope of responsibility and requires one to manage more data dependencies.

The data owner plays a key business role in enterprises, since data quality remains one of the key challenges to create value from BDA (Abbasi et al. 2016; Grover et al. 2018; Wamba et al. 2015). However, BDA extends the responsibilities of data owners to also provide the input data for new data products. Therefore, as data suppliers, data owners must additionally handle this extended use of data in their area of responsibility. The counterparts of data owners are data product owners, who represent data consumers.

Proposition 3: The data product owner ensures business value of a data product over its lifetime, including use case portfolio management, development, maintenance, and user support. Depending on the data product's complexity, this role may require technical expertise; thus, this may be a shared role between business and IT.

The data product owner plays a key business-oriented role in creating value from analytics products. In contrast to BI environments, the creation of analytics products requires more resources (e.g., for development and deployment) and greater coordination efforts (e.g., by managing an analytics portfolio) (see also *Analytics experimentation* and *Analytics production capabilities* in Essay I). Therefore, this ownership type is important to translate business requirements into data products in an efficient and effective way. However, data products are usually dependent on data from different organizational units, which require additional coordination effort to manage these provider-consumer relationships. Therefore, the data platform owner plays an important role in facilitating this effort.

Table 5. Data ownership types in the context of BDA

Data owner type	Responsibilities	Support in cases	Exemplary statement
Data owner	Accountable for quality and lifecycle of data in the domain of responsibility.	A, B, C, D	*"[…] accountable for the overall integrity, data lifecycle, and data quality of data created in his ownership."* (A)
	Fulfils quality requirements for data in the domain of responsibility for data products.	A, D	*"Fulfils service-level agreements for data products."* (A)
	Ensures compliant access and use of data in the domain of responsibility by handling requests, providing access, and approving usage.	A, B, C, D	*"Controls reading access […] ensures compliant use through the provision of no-join policies […]."* (B)
Data platform owner	Ensures data quality on the platform by managing data pipelines to onboard and provision data.	A, C	*"Oversees the implementation and availability of data pipelines to onboard data to the data hub and to provision data to data solutions."* (C)
	Accountable for onboarding of valuable data according to a business need and potential.	B	*"Ensures that new and valuable data are onboarded to the data lake according to the business need and potential."* (B)
	Responsible for the development and operation of the data platform. Approves compliance of data products according to data platform standards.	B, C, D	*"Data platform owner prioritizes all the requirements coming from all areas and own the platform, and basically give the direction how the platform will develop further."* (D)
Data product owner	Ensures that a data product addresses a business need and generates business value over its lifetime.	A, D	*"He ensures business value of a data product over its lifetime."* (A)
	Accountable for a data product over its lifetime, including use case portfolio management, development, maintenance, and user support.	A, C, D	*"Accountable for a data application over its lifetime, which includes compliant implementation, maintenance of the data application, and support of users."* (C)
	Ensures compliant access and use of data products.	B	*"Manages access to data lab, app, or user home and is accountable for any activity […] on it over its lifetime."* (B)

Proposition 4: In BDA environments, the data platform owner role facilitates data supply (data owners) and data demand (data product owners). This activity ensures the availability of data on the platform for data exploration and experimentation, but also for the operation of data products.

There is usually one owner for each platform in an enterprise, while there are many data owners and data product owners. This role is important in coordinating the data demand (data owner) and supply (data product owner) to ensure the efficient and effective delivery of data products.

Implications of data repurposing

Data repurposing has implications for the assignment of data ownership and implies changes in responsibilities, which are described by two further propositions. The fifth proposition reflects on the implications at the data source level.

Proposition 5: With data repurposing, data's context of use deviates more often from its origin. Thus, new data owners may be assigned if the data creators are not able to cope with the additional data requirements.

Data repurposing might imply changes in a dataset's context of use and results in new data requirements. However, these deviations must be managed at the source level and require new responsibilities to maintain data requirements and to ensure compliant access and use. The data are then repurposed on the platform, which leads to further implications on the platform level, which are described by a sixth proposition.

Propositions 6: With data repurposing, the number of dependencies between datasets and data products are increasing. The data platform owner assumes additional responsibilities for maintaining transparency and contractual agreements between data owners and data product owners.

Data repurposing is performed on the data platform. Thereby, new dependencies are created that need to be managed in addition to the data source requirements. For instance, Google engineers warn against the high technical debt of ML systems due to the increase of data dependencies (Sculley et al. 2015).

4.2.4 Conclusion

This essay reaffirms the importance of the data ownership concept but concludes that it should be reconsidered in emerging BDA environments, that is, EAPs. Some of the established principles for operational systems and data warehouses remain valid in these environments; most importantly, the need for a clear distinction between the owner on the data supply side (*data owner*) and the owner on the data demand side (*data product owner*). However, BDA environments pose specific challenges due to data repurposing and the characteristics of advanced analytics products, which lead to an extension of responsibilities and an additional data ownership type.

The research outcomes contribute, in general, to the data and analytics governance literature and, in particular, to research on decisions rights and IS ownership. This study extends the prevailing view on data ownership by integrating the data platform owner type and reflecting

on the implications of data repurposing. Based on Grover et al.'s (2018) research framework, the foundation of BDA governance is laid to facilitate the value creation process and to explain decision rights according to Tiwana et al.'s (2013) IT governance cube.

From a practitioner perspective, the data ownership types are useful to align them with existing data governance designs and to derive further roles and responsibilities.

From a scientific perspective, researchers can use the data ownership types to investigate structural data governance mechanisms (e.g., definition of decision areas and distribution of decision rights) and the identified data repurposing challenges to further extend the literature on BDA management.

4.3 Essay III: Data Governance: From Master Data Quality to Data Monetization

4.3.1 Motivation

Companies must not only adapt their fundamental rights and responsibilities when building BDA capabilities (see Essay II), but also their overarching data governance designs to foster the alignment of strategic objectives with business and IT stakeholders and with control value creation at an enterprise level. A rich body of knowledge exists on IT governance that provides a thorough analysis of different mechanisms. Here, IT governance mechanisms are distinguished by their organizational purposes and classified into structural (definition of decision rights and responsibilities), procedural (formulation of decision making), and relational (communication, knowledge sharing, and alignment) mechanisms (De Haes and Van Grembergen 2004; Peterson 2004). However, research on data governance is generally scarce. Furthermore, it does not reflect the strategic role that data currently plays and focuses mainly on specific (mostly structural) mechanisms. Previous studies have explored data governance in the context of data warehouses (Rifaie et al. 2009; Watson et al. 2004), master data and data quality management (Khatri and Brown 2010; Otto 2011c, 2011b; Weber et al. 2009), and data lifecycle management (Tallon et al. 2013). Hence, literature on data governance must be extended to include the changing role of data in enterprises (Grover et al. 2018). This lack requires an understanding of governance designs – encompassing structural, procedural, and relational governance mechanisms – with the same strategic view considered by researchers investigating these mechanisms for IT artifacts.

4.3.2 Research objectives and approach

This essay aims to understand how companies design data governance with data's changing role and addresses the following research question:

RQ I: *How do companies design data governance using structural, procedural, and relational mechanisms?*

RQ II: *How do companies implement data governance to address the changing role of data?*

To answer this research question, this study uses multiple exploratory case studies to investigate a diverse set of nine multinational companies in terms of their industries, strategic contexts, data scope, and experience with data governance (Benbasat et al. 1987; Yin 2003). This research design is a promising approach because governance designs are contingent on a variety of internal and external factors (Sambamurthy and Zmud 1999).

Case selection

Theoretical sampling (Eisenhardt and Graebner 2007) was applied to a selection of nine enterprises (see Table 6) that are part of the consortium research program. These enterprises have diverse characteristics in terms of their industries, strategic contexts, data scope, and experience with data governance. The sample diversity allows an analysis of differences and commonalities in data governance designs and the deriving of generalizable patterns (Dubé and Paré 2003).

Table 6. Case companies

Company	Industry	Revenue/Employees	Main contact
A	Public transportation and freight, mobility infrastructure	$1B–$50B / ~35 000	Product owner data strategy
B	Manufacturing, chemicals	$1B–$50B / ~5 000	Head of Corporate Data Management
C	Packaging, food processing	$1B–$50B / ~25 000	Head of Data Management and BI
D	Manufacturing, automotive	$1B–$50B / ~90 000	Vice-President: Data and Analytics Governance
E	Consumer goods	$50B–$100B / ~350 000	Master Data Lead
F	Manufacturing, automotive	$1B–$50B / ~150 000	Head of Master Data Management
G	Pharmaceuticals	$1B–$50B / ~70 000	Global Data Lead-Enterprise Solution
H	Consumer goods, retail	$1B–$50B / ~30 000	Vice-President: Data and Analytics
I	Consumer goods, retail	$100B–$150B / ~450 000	Head of Data Management

Data collection

Primary data was collected through semi-structured interviews of 1.5 hours duration with key informants who form part of the central data organization and who have a mandate for enterprise-wide data governance. The interviews were complemented with an analysis of additional company documents (e.g., the company's data strategy or data role models) and publicly available information (e.g., news articles or financial reports). Triangulation of the information gathered through the primary and secondary sources ensured construct validity (Yin 2003).

Data analysis

The collected data were analyzed in two steps. First, a within-case analysis was conducted by coding the interviews with identified governance mechanisms as the analysis framework, which was further extended during the coding process. The final set of identified data governance mechanisms provides answers to the first research question. Second, in order to answer the second research question, a cross-case analysis was executed to identify commonalities and

differences in the implementation of data governance mechanisms. Through this analysis, patterns were derived and grouped into archetypes. The results were presented and discussed in two focus groups comprising the interviewees and other data governance experts. Both focus groups confirmed the identified data governance mechanisms and archetypes.

4.3.3 Research outcomes and contributions

The immediate research outcomes form a set of structural, procedural, and relational data governance mechanisms and three data governance archetypes (see Table 7) that characterize typical ways of governing data according to the implemented governance mechanisms: (1) improve master data quality, (2) enable enterprise-wide data management, and (3) coordinate the network to enable data monetization. The archetypes are summarized using the three identified governance mechanisms.

Improve master data quality

Enterprises (i.e., case companies B and G) in this governance archetype establish mechanisms to improve data quality for master data in a few data domains, among others, customers, products, and finance. These data domains represent the most relevant data objects and define distinct areas of responsibility. While this structuring approach is also used in the other data governance archetypes, Archetype I has distinct governance characteristics. From a structural governance perspective, responsibilities are primarily centralized, although the data lifecycle is mainly managed through the business functions. Accordingly, the central data team assumes operational responsibility to gather business requirements, create data quality measures, monitor data quality, and support projects facing data quality issues. A few decentralized roles, which are often assumed by dedicated teams or shared service centers, directly manage the data lifecycle in the business functions. From a procedural governance perspective, the primary focus of the central team is the planning and management of investments in data quality and infrastructure improvements, which are driven either by the IT budget or by the budgets of business stakeholders. Data quality is proactively measured, and the data lifecycle is managed for core business objects. From a relational governance perspective, data standards and compliance are communicated to the business functions by the central data organization. The central data team aligns and collaborates with IT stakeholders, mostly in regular meetings or in collaboration with the business stakeholders involved in projects. Knowledge is shared on the compliant use of data.

Enable enterprise-wide data management

Enterprises (i.e., case companies E, F, H, I) in this data governance archetype have an enterprise-wide focus on data management and consider a diverse set of data domains and data types. In accordance with this extended scope, Archetype II adapts the implemented governance mechanisms. From a structural governance perspective, the central data team has a wider array of responsibilities. Beside the responsibility of managing data quality, it also assumes responsibility for data strategy and data access/availability. Accordingly, more responsibilities are decentralized. In contrast to the previous archetype, responsibilities for the gathering of business requirements and maintaining data are mainly decentralized to business functions. From a procedural governance perspective, the strategy and planning process is enterprise-wide and focuses investments, not only to improve data quality but also to enhance overall access and availability. Hereby, business cases establish new data domains that are calculated to further strengthen the data management capability. From a relational governance perspective, the relational mechanisms are more intensively established than in the first archetype. For instance, roles and responsibilities are communicated, and regular meetings and steering committees foster collaboration and alignment between data and business professionals.

Table 7. Data governance archetypes

	DATA GOVERNANCE ARCHETYPES		
	Archetype I	**Archetype II**	**Archetype III**
	Improve master data quality	***Enable enterprise-wide data management***	***Coordinate the network to enable data monetization***
	CASE COMPANIES		
	B and **G**	**E**, **F**, **H**, and **I**	**A**, **C**, and **D**
DATA STRATEGY			
Objectives	Improve data quality to enable business processes/reporting	Improve data quality to enable business processes/reporting, broaden data access/availability	Improve data quality, broaden data access/availability, monetize data
Scope	Narrow scope on master and reference data and few data domains	Broad scope on any data type and increasing number of data domains	Broad scope on any data type including analytical data and stable number of data domains
STRUCTURAL MECHANISMS			
Organizational structure	Central/Decentral	Central/Federated	Federated
Steering and oversight	Small data organization with essential data roles	Dedicated boards Large data organization with data roles, including assigned roles to business stakeholders	Dedicated boards Large data organization of data and analytics roles, manages as an extended network
PROCEDURAL MECHANISMS			
Strategic planning	Some uncoordinated data strategy planning activities, investments in data quality improvements and infrastructure	Emerging data strategy planning process, investments in data quality improvements and infrastructure, business case analysis for new data domains	Data strategy planning and control process, pro-active identification, and management of data monetization opportunities
Data governance design and control	Ad hoc creation of standards and data models for master data	Data governance framework and process for data modeling and architecture design	Data and analytics data governance framework, unified data architecture
Operational data management	Data quality monitoring and support, uncoordinated data lifecycle management	Data quality monitoring and support, coordinated data lifecycle management	Data quality and use monitoring and support, and data lifecycle management in business functions
RELATIONAL MECHANISMS			
Alignment and collaboration with business	Mostly through procedures or extended boards	Boards and collocation	Boards and collocation
Alignment and collaboration with IT	Collocation with 1-2 data roles in IT functions	Collocation with an extended array of responsibilities for data-related aspects in IT function	Collocation or even combined with a focus on delivering data and analytics products
Data knowledge sharing and use	Few communities for master data Training compliant access and use	Regular updates Emerging community management Training in data quality methods	Regular updates Community management for data and analytics Training in data literacy

Coordinate the network to enable data monetization

Enterprises (i.e., companies A, C, D) in this data governance archetype understand data as a strategic asset and a major driver of their digital transformation. These companies integrate their

data and analytics organization, through which they promote synergies and seamlessly manage data quality and usage. Accordingly, they implement data governance mechanisms that enable data monetization in various ways. From a structural governance perspective, the central data team mostly undertakes strategic responsibilities. The operational responsibilities are delegated to a coordinated network of decentral-organized data roles. Therefore, companies establish the role of the Chief Data Officer to strengthen alignment and manage data monetization activities across the enterprise. From a procedural governance perspective, the focus of all processes is to find new monetization opportunities that align with the enterprise-wide and domain perspectives. From a relational governance perspective, the coordination of the extended network of data professionals becomes a key concern. Alignment and collaboration occur on both an operational level (through communities) and a strategic level (through boards). Communication and knowledge sharing are primarily channeled through data communities, which comprise key data users and which are actively coordinated as virtual networks.

4.3.4 Conclusion

This study describes how data governance designs change to account for the strategic role that data play in today's enterprises and integrates requirements emerging from BDA. In addition, it provides a thorough analysis of structural, procedural, and relational data governance mechanisms. The three data governance archetypes reveal that data evolve beyond the data quality and operational aspects shown in previous studies (e.g., Otto, 2011; Tallon, Ramirez and Short, 2013). They are also key to manage data as a strategic asset (Legner et al. 2020) and to leverage data's monetization opportunities (Wixom and Ross 2017). As a result, relational mechanisms gain in importance by coordinating a broad network of professionals who monetize data in various ways. Moreover, these mechanisms are required to foster the alignment and collaboration of data and analytics teams with business and IT stakeholders.

From a practitioner's perspective, the data governance archetypes assist in designing data governance initiatives and managing data as a strategic asset. From an academic perspective, this essay advances the scientific field of IT governance in general and data governance in particular. With data's changing role in enterprises, data governance mechanisms must be adapted. Therefore, relational and procedural governance mechanisms play an essential role in the coordination of decentralized data roles and the creation of additional business value. These developments call for research that investigate these changes in greater depth.

4.4 Essay IV: Machine Learning Techniques for Enterprise Data Management: A Taxonomic Approach

4.4.1 Motivation

Data quality (DQ) is considered a significant factor for organizational success (Otto and Österle 2015). A study reveals that poor DQ can cost a company between 15 – 25 % of its annual revenue (Redman 2017) and top management recognizes DQ as a critical factor when implementing AI (Pyle and José 2015), business analytics, or self-service business intelligence (BARC 2018). The prevalent traditional (rule-based) DQ management approaches that are primarily used by EDM are considered tedious and do not scale well with increasing amounts of data (Stonebraker and Ilyas 2018). An early study found that from 33 "dirty data" types at least 24 of them required intervention by a domain expert and only nine could be handled automatically (Kim et al. 2003). Nevertheless, recent advances in BDA techniques, especially ML (including deep learning), provide means to handle DQ more efficiently. First, evidence on the significant opportunities of ML to support EDM exists in research and practice. For instance, *Data Tamer* (commercialized under the name *Tamr*) is an ML-based data curation system, developed at the Massachusetts Institute of Technology, which was able to reduce data curation costs in three real world examples by about 90% (Stonebraker et al. 2013). More recent studies show how ML outperforms other approaches to predict missing values (called data imputation) (Wu et al. 2020) or to create "golden" records from duplicated data (called record fusion) (Heidari et al. 2020). Most of these studies focus on specific data curation tasks and are of a pure technical nature. Thus, a comprehensive overview of how EDM can benefit from ML is currently non-existent but is arguably beneficial for practice and research in order to gain a systematic understanding of this dynamic and rapidly evolving field.

4.4.2 Research objectives and approach

This essay aims to comprehensively understand how ML can be used in EDM practices. Therefore, this study addresses the following research questions:

RQ I: *Which elements describe machine learning techniques for Enterprise Data Management?*

RQ II: *Which archetypes of machine learning for Enterprise Data Management can be distinguished?*

To answer these research questions, this study's objective is to develop a taxonomy that helps to classify and describe ML techniques for EDM. Taxonomies are *"systems of groupings which are derived conceptually or empirically"* (Nickerson et al. 2013, p.3). They are ideal to structure a dynamic field that has not yet been intensively researched, and in which concepts remain in disorder. This situation prevails in this study: ML techniques for individual data curation tasks exist and continue to be developed, but they remain disordered and have not been linked with the larger context of EDM.

The taxonomy was developed following the method suggested by Nickerson et al. (2013).

Data collection

As a foundation for the development of a taxonomy, cases that describe ML techniques for EDM were collected from academic literature, focus groups and expert interviews, and market analysis (see Table 8).

Table 8. Sources considered for taxonomy development

Sources	Applied method	ML techniques
Academic literature	Literature review to identify ML techniques suggested for EDM in research.	29 academic cases
Focus groups and expert interviews	Focus groups and expert interviews to identify ML techniques that companies have started to explore and use for EDM	12 practitioner cases
Market analysis	Screening of tools and suites that offer ML techniques for EDM	19 applications

Taxonomy development

The collected cases were used to develop the taxonomy that provides an answer to the first research question. According to Nickerson et al. (2013), the taxonomy's purpose and its meta-characteristics must be defined first. The purpose is to describe and classify ML techniques for EDM; the meta-characteristics are the EDM context (i.e., the specific situation in which the ML

technique is applied) and the ML application (i.e., the concrete way of applying ML to support EDM). The taxonomy was developed in four iterations. First, a conceptual-to-empirical iteration was performed to ground the taxonomy in scientific literature. Thereafter, three iterations of empirical-to-conceptual were performed in which the dimensions and characteristics were further adapted, based on the examination of a sub-set of the collected cases. After the taxonomy reached its final status, it was ex post evaluated by assessing its robustness through the calculation of intercoder reliability, using the classifications of two researchers.

Archetypes identification

To address the second research question, the taxonomy was used to classify the 60 cases and identify archetypes (i.e., typical application scenarios of ML in EDM) through qualitative clustering.

4.4.3 Research outcomes and contributions

The research outcomes are a taxonomy to classify ML techniques for EDM and nine typical application scenarios (archetypes) to use ML in data management. These outcomes are summarized as follows:

Taxonomy

The taxonomy comprises nine dimensions which are structured along the chosen meta-characteristics (see Table 9). The *EDM context* meta-characteristic encompasses four dimensions: the *Data production process* (as the high-level process), the *Data domain* (as the data's use context), the *Data curation task* (as the specific activity performed to curate data), and the *DQ impact* (as the benefit from data curation). The *ML application* meta-characteristic comprises five dimensions: the *Input data* (data type to train the ML model), the *Learning strategy* (the way how the ML model is trained), the *Learning goal* (the output of an ML model), the *Model type* (the ML model architecture), and *Task impact* (the benefit of using the ML model from a task perspective).

Table 9. Taxonomy of machine learning techniques in Enterprise Data Management

	Dimension	Characteristics				References
EDM context	**Data domain**	Party	Location	Thing		(Cleven and Wortmann 2010)
	Data production processes	Acquire and create[1]	Unify and maintain[1]	Protect and retire[1]	Discover and use[1]	(Strong, Lee, and Wang 1997)
	Data curation tasks	Data cleaning	Entity resolution	Data trans-formation	Data integration	(Ilyas and Chu 2015) (Rahm and Do 2000) (Rahm and Bernstein 2001) (Shvaiko and Euzenat 2005) (Mukkala et al. 2015) (Köpcke and Rahm 2010) (Elmagarmid et al. 2007)
		Metadata discovery		Data archiving	Data enrich-ment[1]	
	DQ impact	Intrinsic	Contextual	Represent-ational	Accessibil-ity	(Wang and Strong 1996)
ML application	**Input data**	Structured[1]	Semi-structured[1]	Unstructured[1]		(Li et al. 2008)
	Learning strategy	Supervised	Semi-supervised	Unsupervised		(James et al. 2013)
	Learning goal	Classification	Regression	Clustering		(Fayyad et al. 1996) (James et al. 2013)
		Summar-ization	Dependency modelling	Change and deviation detection		
	Model type	Shallow[1]		Deep[1]		(LeCun et al. 2015)
	Task impact	Substitution	Augmen-tation	Assemblage		(Rai et al. 2019)
Legend: [1] characteristic added in the empirical-to-conceptual iteration						

Archetypes

Based on the classification of the complete case base with the taxonomy, nine archetypes were identified that each represent a homogenous group of ML techniques supporting a distinct data production process (see Table 10). Each data production process is subsequently briefly described, along with a description of the corresponding archetypes.

In the data production process *Acquire and create,* data are sourced externally, created manually, or created automatically by a machine. Three typical scenarios of ML use in this process can be distinguished: First, ML supports the manual data entry, which is often prone to wrong or incomplete data entries. Here, ML ensures that data is entered into systems in the expected quality, for example, by adjusting the sequence of form elements dynamically (Chen et al. 2010) or by predicting the correct form values (Ali and Meek 2009). Second, ML automates the manual transformation of data from a source into a target format, which is often time consuming and prone to errors. For instance, ML can extract text segments from free text (Hu et al. 2017) or translate text from one language into another (DeepL 2018; Wu et al. 2016). Third, ML

automatically enriches data, which is also a time-consuming task. For instance, ML assigns products to the correct product category or commodity codes.

Table 10. Archetypes of machine learning techniques for Enterprise Data Management

Processes	Archetype	Description	Classification
Acquire and create	1. Support manual data entry Case IDs: 1 - 7	Learn data entry patterns to prefill values and adapt sequence of form elements for faster data entry and higher DQ by minimizing the risks of invalid/wrong data entries, blanks, or typos.	CURATION TASK: data cleaning DQ IMPACT: intrinsic DATA INPUT: structured STRATEGY: supervised, unsupervised GOAL: dependency modelling TASK IMPACT: augmentation
	2. Automated transformation of data Case IDs: 8 - 19	Learn how to transform data from a source to a target format, that is, extract structured data from text, photos, or videos, translate text or automatically generate text from structured data.	CURATION TASK: data transformation DQ IMPACT: representational DATA INPUT: structured/unstructured STRATEGY: supervised, unsupervised GOAL: clustering TASK IMPACT: substitution
	3. Support data enrichment Case IDs: 20 - 24	Learn to classify records and documents to enhance further processing and analysis.	CURATION TASK: data enrichment DQ IMPACT: contextual DATA INPUT: structured/unstructured STRATEGY: supervised GOAL: classification TASK IMPACT: augmentation
Unify and maintain	4. Support data cleaning Case IDs: 25 - 33	Learn to detect and correct data errors from existing datasets and user feedback to accelerate reactive data cleaning.	CURATION TASK: data cleaning DQ IMPACT: intrinsic DATA INPUT: structured STRATEGY: supervised GOAL: classification TASK IMPACT: assemblage
	5. Support data matching Case IDs: 34 – 42	Learn to identify similar data entities to reduce the number of duplicates and enhance data unification.	CURATION TASK: entity resolution DQ IMPACT: intrinsic, representational DATA INPUT: structured STRATEGY: supervised GOAL: classification TASK IMPACT: augment., assemblage
	6. Support data integration Case IDs: 43 - 47	Learn to link data and tables from heterogenous sources, based on semantic and syntactic similarities, to accelerate data integration and discovery.	CURATION TASK: data integration DQ IMPACT: representational DATA INPUT: structured/semi-structured STRATEGY: supervised GOAL: classification TASK IMPACT: augmentation
	7. Automatic derivation of data quality rules Case IDs: 48 – 51	Learn the dependencies between data attributes to extract and discover new DQ rules, in order to facilitate proactive data management.	CURATION TASK: metadata discovery DQ IMPACT: intrinsic DATA INPUT: structured STRATEGY: unsupervised GOAL: summarization TASK IMPACT: augmentation
Protect and retire	8. Automatic detection of sensitive and out-of-date data across systems Case IDs: 52 - 56	Learn to identify sensitive data and detect life-cycle events, for example, when data needs to be retired to reduce the risk of non-compliance with data protection regulations.	CURATION TASK: metadata discovery DQ IMPACT: accessibility DATA INPUT: structured/unstructured STRATEGY: supervised GOAL: classification TASK IMPACT: augmentation
Discover and use	9. Support the discovery of relevant data Case IDs: 57 - 60	Learn data usage patterns and deep representations of data and tables to make dataset recommendations, in order to enhance discovery and use.	CURATION TASK: data integration, enrichment DQ IMPACT: contextual DATA INPUT: structured/semi-structured STRATEGY: super-, unsupervised GOAL: classif., regress., clustering TASK IMPACT: augmentation

In the data production process *Unify and maintain*, data are maintained according to business requirements. Four typical scenarios of ML use in this process can be distinguished: First, ML supports data cleaning activities, which are usually required in any process involving data. For

instance, ML cannot only help to detect but also to repair data by predicting types of data repairs (Volkovs et al. 2014). Second, ML assists in finding duplicated data records (data matching) and resolving them to a unique entity. This is a typical problem when different systems are merged. For instance, ML can resolve textual and also erroneous data records (Mudgal et al. 2018). Third, ML helps to integrate data across systems, which is necessary when pursuing analytical use cases. For instance, ML finds semantic mappings between schemas (Doan et al. 2001). Fourth, ML suggests additional data quality rules for proactive data quality management (Hipp et al. 2001).

In the data production process *Protect and retire*, data are protected when they contain sensitive information or removed once they are out of date or do not comply with regulations. Here, ML detects sensitive data across systems and classifies them into different protection levels.

In the data production process *Discover and use*, data are made accessible for data consumers. Here, ML links interrelated datasets from heterogenous systems (Fernandez et al. 2018).

4.4.4 Conclusion

While previous studies mainly developed ML solutions for data curation, this study's outcomes link this technical perspective to EDM in order to identify different usage scenarios and to understand the socio-technical implications in a systematic way. Based on a classification of an extensive case base with the taxonomy, nine usage scenarios of ML in EDM are identified, which suggest ways to improve DQ in a more scalable way than do traditional, rule-based approaches. In conclusion, ML supports both reactive and proactive approaches to EDM. However, the use of ML has implications for general work processes. Manual data maintenance efforts are shifted from data custodians (in a reactive mode) to data collectors (in a proactive mode) so that data are correctly entered into systems at the point of entry. In addition, the analyzed cases reveal that ML applications often do not automate processes entirely, but that suggestions are provided on the basis of which a human user reacts to reach a desired system status in an interactive way.

From an academic perspective, the research outcomes can help researchers to position their work and to identify opportunities for further research. While some of the identified archetypes build on a rich body of research that has evolved over the last decades (e.g., entity matching or data integration), other archetypes can be considered novel and open interesting, new research fields (e.g., data discovery or sensitive data detection).

From a practitioner standpoint, the taxonomy and archetypes can help companies to identify areas where ML can support existing EDM processes.

4.5 Essay V: All Hands on Data: A Reference Model for Enterprise Data Catalogs

4.5.1 Motivation

In the realm of BDA, companies must coordinate an increasing number of data provider-consumer relationships (see Essay II) and broaden networks of professionals to monetize data in various ways (see Essay III). Here, data access and search form the main bottlenecks to leveraging data and creating additional business value. Data democratization has become an essential concept in overcoming these obstacles and making data available for a broader range of employees (Awasthi and George 2020). Prior studies have not yet investigated the means and, more specifically, the platforms that support data democratization. In practice, however, companies establish enterprise data catalogs (EDC) as integrated platforms to support technical as well as business professionals in finding, accessing, and using data. These platforms are considered an integral component of the future enterprise IT landscapes (Belissent et al. 2019) and EAPs (see Essay I). Nevertheless, the EDC market is dynamic and solutions come with varying functionalities (Goetz et al. 2020; Sallam et al. 2020; Zaidi et al. 2017). In research, the EDC concept has yet not been reflected and, in practice, companies struggle to choose and implement an EDC solution that aligns with their data democratization strategy. Therefore, making sense of the EDC concept can open interesting, new research avenues while providing fundamental insights into the possibilities of democratizing data in enterprises.

4.5.2 Research objectives and approach

This essay aims to understand the EDC concept in the broader context of data democratization initiatives and addresses the following research question:

RQ: *What are the main constituents of an Enterprise Data Catalog as an emerging platform for data democratization?*

The objective of this research is to develop a reference model which helps to clarify the EDC concept and to understand its main constituents. A reference model is defined "*as a normative construction (or artifact) created by a modeler who describes a system's universal elements and relationships as a recommendation, thus creating a center of reference*" (Ahlemann and Riempp 2008, p.89). In data management, reference models have been used extensively to accumulate knowledge and to provide essential guidance (Legner et al. 2020). They represent a particular approach to accelerate the development of enterprise-specific models (Fettke and Loos 2003, p. 35) and are therefore ideal to fulfill the research goals of this study.

The EDC reference model was developed in three iterations with close collaboration of companies, following the research method defined by Peffers et al. (2007). Data were collected through literature reviews, participation in EDC projects, and focus groups and interviews with data management experts from 13 large international companies that are part of the consortium research program.

4.5.3 Research outcomes and contributions

The EDC reference model comprises multiple levels (Frank 2014). The first level is the reference model architecture "*to decompose the overall problem domain into smaller manageable units and provide a high-level overview of the reference model*" (Ahlemann and Riempp 2008, p. 92). The architecture was modeled on the basis of an analysis of related concepts that address data democratization in different contexts (i.e., digital library and data space) and of predominant IS architecture conceptualizations (Chang et al. 2007; Scheer 2001; Scheer and Schneider 2006). The architecture structures three views (organization, function, and data) and their relation to each other (see Figure 5). The second level includes the views which describe an EDC from an organization, function, and data perspective, through their domain-specific elements. Each view is summarized as follows:

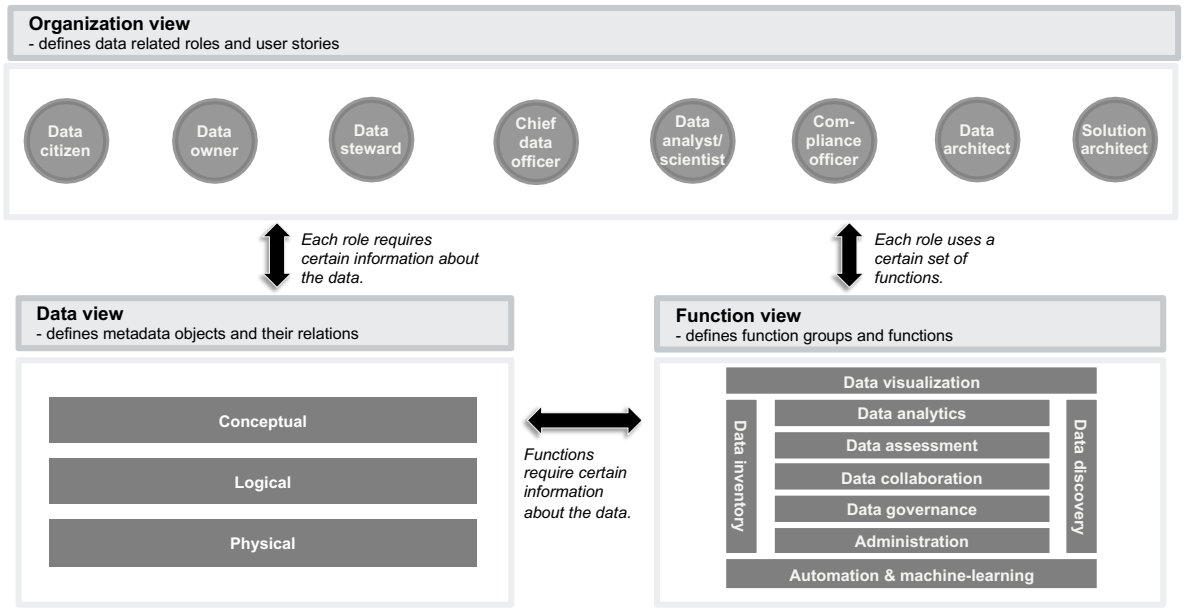


Figure 5. Enterprise data catalog reference model architecture

Organization view

The organization view reflects on the specific needs of different user communities, which is a paramount requirement for data democratization platforms. Eight user roles are identified that belong to data collector, data consumer, and data custodian role categories (Lee and Strong 2004) and serve different purposes regarding data supply, demand, and curation (Borgman 2003; Lord et al. 2004). Accordingly, different user stories are distinguished for each role. Each user story describes typical reasons why a role interacts with the EDC and defines the related function groups/functions and metadata objects. On the supply side, data collectors (e.g., the data architect, solution architect, or data steward) bring data to the EDC that are of relevance for data consumers. They register and document enterprise data resources, for example, a data steward produces a dataset of the past purchases that each customer has done in the previous year. On the demand side, data consumers (e.g., the data citizen, data analyst/scientist, compliance officer, or chief data officer) use data to pursue their daily tasks. They primarily use the EDC to find, access, and understand data, for example, a data scientist searches for a reference dataset (e.g., school holidays) that can be used in a predictive modelling use case. On the curation side, data custodians (e.g., data steward, data owner) take care of the data that has been registered in the EDC to ensure that it is fit for use by data consumers. They manage access, assess data quality, and document data.

Function view

The function view comprises the different functionalities that are needed to support the different user communities in using the EDC. They are hierarchically decomposed – using function trees (Scheer 2001, pp. 21-38) – into two-layers of function groups and functions. The function groups *Data inventory* and the *Data discovery* are essential to manage data supply (e.g., data registration or metadata management) and demand (e.g., search or data delivery). The other function groups comprise functions to support individual user roles in *Data governance* (e.g., workflow or role and responsibility management), *Data assessment* (e.g., data quality or data usage), *Data analytics* (e.g., data query or data story), and *Administration* (e.g., configuration or user management), which are complemented with function groups for *Visualization* (e.g., graphs or diagrams), *Automation and ML* (e.g., automated scanning or recommendation), and *Data collaboration* (e.g., tagging or rating).

Data view

The data view structures the different metadata objects and attributes in the form of a metadata model (Kerhervé and Gerbé 1997). This model describes data (including relationships) from business and system-oriented perspectives to make it accessible to different user communities. According to data modelling guidelines (Batini et al. 1986; Tsichritzis and Klug 1978), the metadata model comprises three layers : conceptual, logical, and physical. The conceptual layer serves the business understanding and is composed of a business process, business terminology, analytics, and a governance view. The logical layer represents a system-agnostic view and abstraction between the physical data storage layer and the conceptual layer (Kumpati 1988). It documents data domains, related business objects/attributes, and applications. The physical layer is the implementation view on data and documents how data are organized and stored in enterprise systems. It documents systems, interfaces, data structures, and data objects/attributes.

4.5.4 Conclusion

This study contributes to the ongoing academic debates on the democratization of data and provides an academic conceptualization of the EDC concept. The EDC reference model anchors emerging platforms for data democratization in enterprises to related concepts, including the digital library and the dataspace (overall concept), metadata management (data view with metadata objects), and data governance (organization view with user roles). The EDC integrates existing metadata management approaches (i.e., data dictionary, business glossary, and metadata repository) to address a broader range of professionals who work with data. However, the EDC concept goes beyond these earlier concepts because it contains rich functionalities that enable the different user roles to not only find and access data but also to collaborate and use data according to their specific needs. EDCs complement EAPs and are an important constituent of future enterprise application landscapes. Moreover, they support companies to govern data on an enterprise-level.

From an academic viewpoint, the findings contribute to the ongoing debates on data democratization, research on data management (Legner et al. 2020), and the emerging EAPs (see Essay I; Hyun et al. 2020).

From a practical perspective, the reference model assists companies in comparing and selecting an EDC solution, but also in providing guidance during the implementation of the solution.

5 Discussion

5.1 Summary and contributions

Data have always played an important role in organizations, but its role and significance in enterprises has changed significantly in the past decade. This shift raises questions about how companies can build the necessary capabilities to leverage their data in innovative ways and create long-term strategic value. This thesis contributes with fundamental reflections on how companies adapt their EDM capabilities and respond to research calls on the potential of capability building to create value from BDA (Akter et al. 2016; Grover et al. 2018; Gupta and George 2016; Mikalef et al. 2018).

BI&A capabilities/Enterprise analytics platform

The first essay lays the foundation for this thesis. Using work systems theory as theoretical lens, it contributes to understanding the resource orchestration processes and capability building for BDA. From four case studies and intense exchanges with experts in focus group meetings, four BI&A capabilities are identified that are prevalent in companies and discussed in the literature: *reporting*, *data exploration*, *analytics experimentation*, and *analytics production*. For each BI&A capability, patterns are determined in the form of a work system with its specific components. Moreover, commonalities among the work systems are detected that suggest managing them as a unified enterprise analytics platform (EAP) to generate synergistic effects.

Data and analytics governance

The second and third essays contribute to the debates about data and analytics governance (Grover et al. 2018), as well as the *"Centralized and decentralized big data capability structures,"* referred to by Günther et al. (2017).

Essay II explains how companies define and assign their fundamental rights and responsibilities for data in the context of BDA environments, and identifies three ownership types, namely *data owner*, *data product owner* and *data platform owner*. The data ownership types help assigning the decision rights for governing the content of IT artifacts according to Tiwana et al. (2013)'s IT Governance Cube. The study confirms that data ownership remains an important concept, but BDA requires changes in fundamental responsibilities of EDM to enable data repurposing and handle an increasing number of data provider-consumer relationships.

The third essay provides a more holistic understanding of how enterprises adapt their governance mechanisms to account for the more significant strategic role of data (Grover et al.

2018; Legner et al. 2020). It defines a set of structural, procedural and relational data governance mechanisms that are commonly implemented by companies. It also identifies three different archetypes that characterize typical ways of governing data, that are (1) *improve master data quality*, (2) *enable enterprise-wide data management*, and (3) *coordinate the network to enable data monetization*. Data governance extends beyond structural governance mechanisms and the mere definition of data-related roles and responsibilities; especially, relational, and procedural governance mechanisms are found to be important to decentralize data responsibilities and coordinate the increasing network of data professionals and data users in large organizations.

Enterprise Data Management tools

The fourth essay investigates ML and its role in managing data quality in an adaptable and scalable way (Grover et al. 2018; Zhu et al. 2014). The developed taxonomy and identified archetypes help in understanding the EDM problem context and the different usage scenarios of ML through a socio-technical perspective. This essay concludes that ML indeed facilitates EDM processes, but that it also fundamentally changes the EDM practices. ML is therefore a promising way to overcome the obstacle of managing data quality to create value from BDA, albeit it requires data science skills.

The fifth essay explores ways to democratize data with EDCs and contributes to the application of the FAIR principles in the enterprise context (Wilkinson et al. 2016). It develops a reference model to conceptualize this emerging concept and its constituents by means of three architecture views, that are (1) *organization view*, (2) *function view*, and (3) *data view*. The findings show that EDCs facilitate data democratization for a broad audience within organizations and go beyond other metadata management solutions by enriching data documentation functionalities with data usage functionalities. Such platforms are therefore a key factor to create value from BDA and a core component of EAPs.

5.2 Theoretical implications

Transactional and analytics systems – from co-existence to collaboration

In the past, operational data processing (OLTP) and analytical data processing (OLAP) were two separated "worlds" that were managed separately from each other. The data flow between these worlds was mainly unidirectional, i.e. data were mirrored by or transferred from operational systems to data warehouses. Data quality issues were resolved at the source, for example, by curating master data. With BDA, however, the data flow is increasingly bidirectional and

requires stronger collaboration between the two data "worlds." Data are onboarded to data lakes without a pre-defined purpose and are then changed to fit, for example, the use contexts of an advanced analytics model. Not only does the data provisioning require more interventions of analytics experts and domain experts to ensure the right data understanding, but also new data quality requirements, for example, that a label required for training an ML model be returned to the source. Moreover, advanced analytics products, for instance, often enhance operational systems to run business processes more efficiently. As the data flow is increasingly bi-directional, companies must manage data lifecycles and analytics-product lifecycles in correspondence. The thesis outcomes (see Essay I and II) provide fundamental insights into how data and analytics can be managed in unison. However, further research is required on managing data quality within continuously changing use contexts, for instance.

The "emancipation" of data governance from IT governance

IS research has for a long time viewed data and information as an integral component of IT artifacts, especially when investigating IT management and governance (Tiwana et al. 2013). However, data must now be viewed as a distinct value driver as it requires specific mechanisms to foster value creation from BDA. For instance, Essay II shows that fundamental rights and responsibilities for data are different from, for example, the more technical-oriented accountabilities of IT applications or infrastructure (Winkler and Wessel 2018). Wixom and Watson (2010) already suggested that BI organizations should emerge as separate organizations, other than IT, because the operational characteristics of delivering reports, for instance, differ from usual IT processes that manage the underlying infrastructure. With BDA, data have become the long-expected, strategically relevant resource (Otto 2015). This thesis informs the management of BDA and forms the basis of data governance designs that consider structural, procedural, and relational mechanisms. It also shows that BDA also requires to extend governance goals: in addition to data quality and regulatory compliance, governance should facilitate data use by broadening data availability and enabling data monetization. Hence, subsequent studies must consider data governance not as an instrument of control, but as a strategic instrument that facilitates strategic alignment and value creation from BDA (Grover et al. 2018).

Rethinking EDM practices

EDM is faced with an ever-increasing demand for data in the enterprise. Existing EDM practices therefore need to be rethought to meet the emerging BDA requirements for data quality and data use. On the data quality side, ML is a promising solution to improve data quality in a more scalable and adaptable way (see Essay IV). However, further studies are needed to investigate learning-based solutions in greater detail. So far, ML techniques have been proposed mainly in more technically oriented research, but not in IS research to understand also the ways for adopting this new technology through a socio-technical lens. On the data use side, the responsibility of EDM extends to making data available and accessible to a larger number of employees. EDCs are a key platform that help democratize data (see Essay V). However, more research is needed that explores the various means and challenges of data democratization. For example, in the past, data were documented primarily for technical reasons, but now they must also be discoverable and understandable by less technical users. Subsequent studies must therefore place more emphasis on these new user groups. Eventually, data only creates value if it is used. Hence, the more employees can use data, the higher the benefits for the company.

5.3 Practical implications

Companies can create value from their data assets only by managing data quality and data use (e.g., analytics) in correspondence. Establishing the corresponding management structures and awareness for data require time and commitment over a longer period. Usually, there are many different data initiatives in companies that first need to be consolidated. The thesis outcomes can support practitioners in consolidating these initiatives and formulating a comprehensive EDM approach. First, by gaining a unified view of the analytics initiatives and managing the corresponding capabilities as an integrated enterprise analytics platform (see Essay I). Second, by adapting or establishing data governance designs with a focus on data availability and data monetization (see Essay II and Essay III). Third, by showing practitioners options to run data quality more efficiently with ML support (see Essay IV). Fourth, by supporting the practitioners in setting up an enterprise data catalog to facilitate data use (see Essay V). Thus, the thesis outcomes can support companies in setting up corresponding programs that provide them with a long-term directive to become more data driven.

5.4 Broadening the perspective in future research

This thesis has investigated data-related capabilities from a company internal perspective. However, value from data is also increasingly created outside the enterprise context along the

value chain by exchanging data between the involved parties (Otto 2019). For instance, the European Union is currently establishing a federated data infrastructure that "*strengthens the ability to both access and share data securely and confidently*" (GAIA-X, 2021). The findings of this thesis should therefore be expanded in future research to integrate also the data ecosystem perspective. Data ecosystems require different mechanisms to facilitate the data exchange between data providers and consumers with semi-automatically negotiation, execution, and monitoring of data usage agreements (Jarke et al. 2019). For this purpose, the International Data Space Association has developed a reference architecture to build platforms that facilitate data sharing and stimulate innovation. It "*includes as main components the so-called IDS Connector – a software component that annotates data to be exchanged with usage policies –, a broker, identity management, and a clearing house for data exchange and sharing transactions*" (ibid, p. 550). Generally, maintaining ownership of data in data ecosystems is more challenging because access to data is contracted out to external parties. With this particular focus, other studies have suggested technical methods to facilitate this process and preserve data ownership with "*watermarking*" techniques (Agrawal and Kiernan 2002; Heckel and Vlachos 2017; Zoumpoulis et al. 2013). These methods introduce a unique noise into a dataset so that the origin is always traceable, making it possible to maintain existing rights even when datasets are shared across organizational boundaries.

The outcomes of this thesis can contribute to the data ecosystem perspective in multiple ways but must be revisited for this particular scenario. For instance, the identified data ownership types in Essay II might also be applicable beyond the enterprise context: the data owner is a data creating party, the data product owner is a data consuming party, and the data platform owner maintains the federated data infrastructure and ensures the trustworthiness and integrity of the data exchange. Further investigations are nevertheless necessary to include the responsibilities for this specific scenario. Also, the findings of Essay V, the enterprise data catalog reference model, can potentially be extended to include the data eco-system requirements. As data are shared and offered publicly on the market, a data registry that allows to search, access, and eventually buy data would be beneficial to facilitate the data exchange and monetization.

5.5 Limitations

This thesis comes with limitations. The companies that have been investigated in the essays (except Essay IV) represent large, multi-national corporations. Their particular challenges are often associated with their complex organizational structures and system landscapes. These

issues might not occur in small and medium-sized organizations. Therefore, the results might not be transferrable.

All essays use qualitative, explorative research designs. The outcomes are therefore explanatory for certain phenomena, but do not provide an indication whether or not certain implementations are more successful than others. Arguably, the outcomes are factors that are linked to value creation and that most likely influence organizational performance in a positive way, for example, making data accessible and searchable. However, it is not implied that they guarantee a positive effect on revenue. The evaluation of the performance impacts requires further studies.

6 References

Aaltonen, A., and Tempini, N. 2014. "Everything Counts in Large Amounts: A Critical Realist Case Study on Data-Based Production," *Journal of Information Technology* (29:1), pp. 97–110. (https://doi.org/10.1057/jit.2013.29).

Abbasi, A., Sarker, S., and Chiang, R. 2016. "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the Association for Information Systems* (17:2).

Agarwal, R., and Dhar, V. 2014. "Editorial - Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research," *Information Systems Research* (25:3), pp. 443–448. (https://doi.org/10.1287/isre.2014.0546).

Agrawal, R., and Kiernan, J. 2002. "Chapter 15 - Watermarking Relational Databases," in VLDB '02: Proceedings of the 28th International Conference on Very Large Databases, P. A. Bernstein, Y. E. Ioannidis, R. Ramakrishnan, and D. Papadias (eds.), San Francisco: Morgan Kaufmann, pp. 155–166. (https://doi.org/10.1016/B978-155860869-6/50022-6).

Ahlemann, F., and Riempp, G. 2008. "RefModPM: A Conceptual Reference Model for Project Management Information Systems," *Wirtschaftsinformatik* (50:2), pp. 88–97.

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., and Childe, S. J. 2016. "How to Improve Firm Performance Using Big Data Analytics Capability and Business Strategy Alignment?," *International Journal of Production Economics* (182), pp. 113–131. (https://doi.org/10.1016/j.ijpe.2016.08.018).

Ali, A., and Meek, C. 2009. "Predictive Models of Form Filling," Technical Report No. MSR-TR-2009-1, Technical Report, Microsoft Research, p. 8.

Alpar, P., and Schulz, M. 2016. "Self-Service Business Intelligence," *Business & Information Systems Engineering* (58:2), pp. 151–155.

Alter, S. 2004. "A Work System View of DSS in Its Fourth Decade," *Decision Support Systems* (38:3), pp. 319–327.

Alter, S. 2013. "Work System Theory: Overview of Core Concepts, Extensions, and Challenges for the Future," *Journal of the Association for Information Systems* (14:2), pp. 72–121.

Awasthi, P., and George, J. J. 2020. "A Case for Data Democratization," in *Proceedings of the 26th Americas Conference on Information Systems (AMCIS)*, Virtual Conference, August 10, p. 23.

Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., and Zhao, J. L. 2016. "Transformational Issues of Big Data and Analytics in Networked Business," *MIS Quarterly* (40:4), pp. 807–818. (https://doi.org/10.25300/MISQ/2016/40:4.03).

Baker, J., Jones, D., and Burkman, J. 2009. "Using Visual Representations of Data to Enhance Sensemaking in Data Exploration Tasks," *Journal of the Association for Information Systems* (10:7), pp. 533–559.

BARC. 2018. "Top Business Intelligence Trends 2019 | What 2,700 BI Professionals Think," *BI Survey*. (https://bi-survey.com/top-business-intelligence-trends, accessed November 27, 2018).

Batini, C., Lenzerini, M., and Navathe, S. B. 1986. "A Comparative Analysis of Methodologies for Database Schema Integration," *ACM Computing Surveys* (18:4), pp. 323–364. (https://doi.org/10.1145/27633.27634).

Bean, R. 2021. "Why Is It So Hard to Become a Data-Driven Company?," *Harvard Business Review*. (https://hbr.org/2021/02/why-is-it-so-hard-to-become-a-data-driven-company).

Belissent, J., Leganza, G., and Vale, J. 2019. "Determine Your Data's Worth: Data Plus Use Equals Value," Consortium Report, Consortium Report, Forrester Research, February 5. (https://www.forrester.com/report/Determine+Your+Datas+Worth+Data+Plus+Use+Equals+Value/-/E-RES127541).

Benbasat, I., Goldstein, D. K., and Mead, M. 1987. "The Case Research Strategy in Studies of Information Systems," *MIS Quarterly* (11:3), pp. 369–386.

Borgman, C. L. 2003. From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World, MIT Press.
Bowne-Anderson, H. 2018. "What Data Scientists Really Do, According to 35 Data Scientists," *Harvard Business Review Digital Articles*, pp. 2–5.
Chang, T.-H., Fu, H.-P., Ou, J.-R., and Chang, T.-S. 2007. "An ARIS-Based Model for Implementing Information Systems from a Strategic Perspective," *Production Planning & Control* (18:2), Taylor & Francis, pp. 117–130. (https://doi.org/10.1080/09537280600913447).
Chen, H., Chiang, R. H., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact.," *MIS Quarterly* (36:4), pp. 1165–1188.
Chen, K., Chen, H., Conway, N., Hellerstein, J. M., and Parikh, T. S. 2010. "USHER: Improving Data Quality with Dynamic Forms," in *Proceedings of IEEE 26th*, Long Beach, CA, p. 12.
Chessell, M., Scheepers, F., Strelchuk, M., Starre, R. van der, Dobrin, S., and Hernandez, D. 2018. "The Journey Continues: From Data Lake to Data-Driven Organization," Redbooks.
Cleven, A., and Wortmann, F. 2010. "Uncovering Four Strategies to Approach Master Data Management," in *2010 43rd Hawaii International Conference on System Sciences*, Honolulu, Hawaii, USA: IEEE, pp. 1–10. (https://doi.org/10.1109/HICSS.2010.488).
Davenport, T. H. 2006. "Competing on Analytics," *Harvard Business Review*. (https://hbr.org/2006/01/competing-on-analytics).
De Haes, S., and Van Grembergen, W. 2004. "IT Governance and Its Mechanisms," *Information Systems Control Journal* (1), pp. 27–33.
DeepL. 2018. "DeepL." (https://www.deepl.com/home, accessed November 27, 2018).
Dinter, B. 2013. "Success Factors for Information Logistics Strategy — An Empirical Investigation," *Decision Support Systems* (54:3), pp. 1207–1218. (https://doi.org/10.1016/j.dss.2012.09.001).
Doan, A., Domingos, P., and Halevy, A. Y. 2001. "Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach," in *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, SIGMOD '01, New York, NY, USA: ACM, pp. 509–520. (https://doi.org/10.1145/375663.375731).
Dubé, L., and Paré, G. 2003. "Rigor in Information Systems Positivist Case Research: Current Practices, Trends, and Recommendations," *MIS Quarterly* (27:4), pp. 597–636.
Eisenhardt, K., and Graebner, M. 2007. "Theory Building from Cases: Opportunities and Challenges," *The Academy of Management Journal* (50:1), pp. 25–32.
Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. 2007. "Duplicate Record Detection: A Survey," *IEEE Transactions on Knowledge and Data Engineering* (19:1), pp. 1–16. (https://doi.org/10.1109/TKDE.2007.250581).
Fadler, M., and Legner, C. 2020. "Building Business Intelligence & Analytics Capabilities - A Work System Perspective," in *Proceedings of the 41st International Conference on Information Systems (ICIS)*, Hyderabad, India, December 13, p. 2615. (https://aisel.aisnet.org/icis2020/governance_is/governance_is/14).
Farid, M., Roatis, A., Ilyas, I. F., Hoffmann, H.-F., and Chu, X. 2016. *CLAMS: Bringing Quality to Data Lakes*, ACM Press, pp. 2089–2092.
Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. 1996. *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), Menlo Park, CA, USA: American Association for Artificial Intelligence, pp. 1–34. (http://dl.acm.org/citation.cfm?id=257938.257942).
Fernandez, R. C., Ilyas, I. F., Madden, S., Stonebraker, M. O. M., and Tang, N. 2018. "Seeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery," *ICDE*, p. 12.
Fettke, P., and Loos, P. 2003. "Classification of Reference Models: A Methodology and Its Application," *Information Systems and E-Business Management* (1:1), pp. 35–53. (https://doi.org/10.1007/BF02683509).

Frank, U. 2014. "Multilevel Modeling: Toward a New Paradigm of Conceptual Modeling and Information Systems Design," *Business & Information Systems Engineering* (6:6), pp. 319–337. (https://doi.org/10.1007/s12599-014-0350-4).

GAIA-X. 2021. "GAIA-X: A Federated Data Infrastructure for Europe." (https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html, accessed August 30, 2021).

George, G., Haas, M. R., and Pentland, A. 2014. "Big Data and Management," *Academy of Management Journal* (57:2), pp. 321–326. (https://doi.org/10.5465/amj.2014.4002).

Goes, P. 2014. "Big Data and IS Research," *MIS Quarterly* (38:3), pp. iii–viii.

Goetz, M., Leganza, G., and Hennig, C. 2020. "Now Tech: Machine Learning Data Catalogs, Q4 2020," Consortium Report, Consortium Report, Forrester Research. (https://www.forrester.com/report/Now%20Tech%20Machine%20Learning%20Data%20Catalogs%20Q4%202020/-/E-RES157529).

Goodhue, D. L., Quillard, J. A., and Rockart, J. F. 1988. "Managing the Data Resource: A Contingency Perspective," *MIS Quarterly* (12:3), Management Information Systems Research Center, University of Minnesota, pp. 373–392. (https://doi.org/10.2307/249204).

Grover, V., Chiang, R. H. L., Liang, T.-P., and Zhang, D. 2018. "Creating Strategic Business Value from Big Data Analytics: A Research Framework," *Journal of Management Information Systems* (35:2), pp. 388–423.

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., and Feldberg, F. 2017. "Debating Big Data: A Literature Review on Realizing Value from Big Data," *The Journal of Strategic Information Systems* (26:3), pp. 191–209. (https://doi.org/10.1016/j.jsis.2017.07.003).

Gupta, M., and George, J. F. 2016. "Toward the Development of a Big Data Analytics Capability," *Information & Management* (53:8), Big Data Commerce, pp. 1049–1064. (https://doi.org/10.1016/j.im.2016.07.004).

Hai, R., Geisler, S., and Quix, C. 2016. "Constance: An Intelligent Data Lake System," in *Proceedings of the 2016 International Conference on Management of Data*, New York, NY, USA: ACM, pp. 2097–2100. (https://doi.org/10.1145/2882903.2899389).

Halevy, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., and Whang, S. E. 2016. "Goods: Organizing Google's Datasets," in *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, San Francisco, California, USA: ACM Press, pp. 795–806. (https://doi.org/10.1145/2882903.2903730).

Hassan, N. R. 2019. "Where Are We Headed in Business Analytics? A Framework Based on a Paradigmatic Analysis of the History of Analytics," in *ICIS 2019 Proceedings* (Vol. 6), Munich, p. 17.

Heart, T., Ben-Assuli, O., and Shlomo, N. 2018. "Using the Work System Theory to Bring Big Data Analytics to the Inpatient Point of Care," in *ICIS 2018 Proceedings*, San Francisco, CA, pp. 1–9.

Heckel, R., and Vlachos, M. 2017. "Private and Right-Protected Big Data Publication: An Analysis," in Proceedings of the 2017 SIAM International Conference on Data Mining (SDM), Proceedings, Society for Industrial and Applied Mathematics, pp. 660–668. (https://doi.org/10.1137/1.9781611974973.74).

Heidari, A., Michalopoulos, G., Kushagra, S., Ilyas, I. F., and Rekatsinas, T. 2020. "Record Fusion: A Learning Approach," *ArXiv:2006.10208 [Cs, Stat]*. (http://arxiv.org/abs/2006.10208).

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75–105.

Hipp, J., Guntzer, U., and Grimmer, U. 2001. "DATA QUALITY MINING," *DMKD*, p. 6.

Hoven, J. van den. 1999. "Information Resource Management: Stewards of Data," *Information Systems Management* (16:1), Taylor & Francis, pp. 88–90. (https://doi.org/10.1201/1078/43187.16.1.19990101/31167.13).

Hu, M., Li, Z., Shen, Y., Liu, A., Liu, G., Zheng, K., and Zhao, L. 2017. *CNN-IETS: A CNN-Based Probabilistic Approach for Information Extraction by Text Segmentation*, presented at the

Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, June 11, pp. 1159–1168. (https://doi.org/10.1145/3132847.3132962).
Hyun, Y., Kamioka, T., and Hosoya, R. 2020. "Improving Agility Using Big Data Analytics: The Role of Democratization Culture," *Pacific Asia Journal of the Association for Information Systems* (12:2). (https://doi.org/10.17705/1pais.12202).
Ilyas, I. F., and Chu, X. 2015. "Trends in Cleaning Relational Data: Consistency and Deduplication," *Foundations and Trends® in Databases* (5:4), pp. 281–393. (https://doi.org/10.1561/1900000045).
James, G., Witten, D., Hastie, T., and Tibshirani, R. 2013. *An Introduction to Statistical Learning*, (Vol. 103), Springer Texts in Statistics, New York, NY: Springer New York. (https://doi.org/10.1007/978-1-4614-7138-7).
Jarke, M., Otto, B., and Ram, S. 2019. "Data Sovereignty and Data Space Ecosystems," Business & Information Systems Engineering (61:5), pp. 549–550.
Jukić, N., Sharma, A., Nestorov, S., and Jukić, B. 2015. "Augmenting Data Warehouses with Big Data," *Information Systems Management* (32:3), pp. 200–209.
Kerhervé, B., and Gerbé, O. 1997. "Models for Metadata or Metamodels for Data?," in *Proceedings of the 2nd IEEE Metadata Conference*, Silver Spring, Massachusetts, USA, September.
Kettinger, W. J., Zhang, C., and Li, H. 2019. "Information Management Capabilities in the Digital Era: The Senior Manager's Perspective," in *ICIS 2019 Proceedings*, pp. 1–15.
Khatri, V., and Brown, C. V. 2010. "Designing Data Governance," *Communication of the ACM* (53:1), pp. 148–152.
Kim, G., Shin, B., and Kwon, O. 2012. "Investigating the Value of Sociomaterialism in Conceptualizing IT Capability of a Firm," *Journal of Management Information Systems* (29:3), Routledge, pp. 327–362. (https://doi.org/10.2753/MIS0742-1222290310).
Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., and Lee, D. 2003. "A Taxonomy of Dirty Data," *Data Mining and Knowledge Discovery* (7), pp. 81–99.
Kohli, R., and Grover, V. 2008. "Business Value of IT: An Essay on Expanding Research Directions to Keep up with the Times," *Journal of the Association for Information Systems* (9:1), pp. 23–39.
Köpcke, H., and Rahm, E. 2010. "Frameworks for Entity Matching: A Comparison," *Data & Knowledge Engineering* (69:2), pp. 197–210. (https://doi.org/10.1016/j.datak.2009.10.003).
Korhonen, J. J., Melleri, I., Hiekkanen, K., and Helenius, M. 2013. "Designing Data Governance Structure: An Organizational Perspective," *Journal on Computing* (2:4), pp. 11–17.
Kumpati, M. 1988. Database Management System with Active Data Dictionary.
LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep Learning," *Nature* (521:7553), pp. 436–444. (https://doi.org/10.1038/nature14539).
Lee, Y., and Strong, D. 2004. "Knowing-Why about Data Processes and Data Quality," *Journal of Management Information Systems* (20:3). (http://www.jstor.org/stable/40398639).
Legner, C., Pentek, T., and Otto, B. 2020. "Accumulating Design Knowledge with Reference Models: Insights from 12 Years of Research on Data Management," *Journal of the Association for Information Systems* (21:3).
Li, G., Ooi, B. C., Feng, J., Wang, J., and Zhou, L. 2008. "EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-Structured and Structured Data," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data - SIGMOD '08*, Vancouver, Canada: ACM Press, p. 903. (https://doi.org/10.1145/1376616.1376706).
Lord, P., Macdonald, A., Lyon, L., and Giaretta, D. 2004. "From Data Deluge to Data Curation," in *Proceedings of the 3rd UK E-Science All Hands Meeting (AHM)*, pp. 371–375.
Lycett, M. 2013. "'Datafication': Making Sense of (Big) Data in a Complex World," *European Journal of Information Systems* (22:4), pp. 381–386. (https://doi.org/10.1057/ejis.2013.10).
Madera, C., and Laurent, A. 2016. "The Next Information Architecture Evolution: The Data Lake Wave," in *Proceedings of the 8th International Conference on Management of Digital*

EcoSystems, MEDES, New York, USA: ACM, pp. 174–180. (http://doi.acm.org/10.1145/3012071.3012077).
Marjanovic, O. 2016. "Improvement of Knowledge-Intensive Business Processes Through Analytics and Knowledge Sharing," in *ICIS 2016 Proceedings*, Dublin, Ireland, December 11, pp. 1–19.
Mathiassen, L. 2002. "Collaborative Practice Research," *Information Technology & People* (15). (https://doi.org/10.1108/09593840210453115).
Maxwell, B. 1989. "Beyond 'Data Validity': Improving the Quality of HRIS Data," *Personnel* (66:4).
Mikalef, P., Pappas, I. O., Krogstie, J., and Giannakos, M. N. 2018. "Big Data Analytics Capabilities: A Systematic Literature Review and Research Agenda," *Information Systems and E-Business Management* (16:3), pp. 547–578.
Mithas, S., Ramasubbu, N., and Sambamurthy, V. 2011. "How Information Management Capability Influences Firm Performance," *Management Information Systems Quarterly* (35:1), pp. 237–256.
Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., and Raghavendra, V. 2018. "Deep Learning for Entity Matching: A Design Space Exploration," in *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18*, Houston, TX, USA: ACM Press, pp. 19–34. (https://doi.org/10.1145/3183713.3196926).
Mukkala, L., Arvo, J., Lehtonen, T., and Knuutila, T. 2015. "Current State of Ontology Matching," No. 4, University of Turku Technical Reports, University of Turku.
Nickerson, R. C., Varshney, U., and Muntermann, J. 2013. "A Method for Taxonomy Development and Its Application in Information Systems," *European Journal of Information Systems* (22:3), pp. 336–359. (https://doi.org/10.1057/ejis.2012.26).
Österle, H., and Otto, B. 2010. "Consortium Research: A Method for Researcher-Practitioner Collaboration in Design-Oriented IS Research," *Business & Information Systems Engineering* (2:5), pp. 283–293. (https://doi.org/10.1007/s12599-010-0119-3).
Otto, B. 2011a. "Data Governance," *Business & Information Systems Engineering* (3:4), pp. 241–244.
Otto, B. 2011b. "Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers," *Communications of the Association for Information Systems,* (29).
Otto, B. 2011c. "A Morphology of the Organisation of Data Governance," in *Proceedings of the 19th European Conference on Information Systems (ECIS)*, Helsinki, Finland, June 9.
Otto, B. 2015. "Quality and Value of the Data Resource in Large Enterprises," *Information Systems Management* (32). (https://doi.org/10.1080/10580530.2015.1044344).
Otto, B., Hüner, K. M., and Österle, H. 2012. "Toward a Functional Reference Model for Master Data Quality Management," *Information Systems and E-Business Management* (10:3), pp. 395–425. (https://doi.org/10.1007/s10257-011-0178-0).
Otto, B., and Österle, H. 2015. *Corporate Data Quality Prerequisite for Successful Business Models*. (http://nbn-resolving.de/urn:nbn:de:101:1-201511272018 6).
Otto, B. 2019. "Interview with Reinhold Achatz on "Data Sovereignty and DataEcosystems,'" Business & Information Systems Engineering (61:5), pp. 635–636.
Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 45–77. (https://doi.org/10.2753/MIS0742-1222240302).
Penrose, E., and Pitelis, C. 1959. *The Theory of the Growth of the Firm*, (4th Revised edition.), Oxford ; New York: Oxford University Press, USA.
Peterson, R. 2004. "Crafting Information Technology Governance," *Information Systems Management* (21:4), pp. 7–22.

Philip Chen, C. L., and Zhang, C.-Y. 2014. "Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data," *Information Sciences* (275), pp. 314–347. (https://doi.org/10.1016/j.ins.2014.01.015).

Phillips-Wren, G., Iyer, L., Kulkarni, U., and Ariyachandra, T. 2015. "Business Analytics in the Context of Big Data: A Roadmap for Research," *Communications of the Association for Information Systems* (37:1).

Pyle, D., and José, C. S. 2015. "An Executive's Guide to Machine Learning | McKinsey." (https://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning, accessed November 27, 2018).

Rahm, E., and Bernstein, P. A. 2001. "A Survey of Approaches to Automatic Schema Matching," *The VLDB Journal* (10:4), pp. 334–350. (https://doi.org/10.1007/s007780100057).

Rahm, E., and Do, H. H. 2000. "Data Cleaning: Problems and Current Approaches," *IEEE Data Eng. Bull.* (23:4), pp. 3–13.

Rai, A., Constantinides, P., and Sarker, S. 2019. "Editor's Comments - Next-Generation Digital Platforms: Toward Human–AI Hybrids," *MIS Quarterly* (43:1), pp. iii–ix.

Redman, T. C. 2017. "Seizing Opportunity in Data Quality," *MIT Sloan Management Review*. (https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/, accessed November 18, 2018).

Rifaie, M., Alhajj, R., and Ridley, M. 2009. "Data Governance Strategy: A Key Issue in Building Enterprise Data Warehouse," in *Proceedings of the 11th International Conference on Information Integration and Web-Based Applications & Services (IiWAS)*, Kuala Lumpur Malaysia, December 14, pp. 587–591.

Roszkiewicz, R. 2010. "Enterprise Metadata Management: How Consolidation Simplifies Control," *Journal of Digital Asset Management* (6:5), pp. 291–297. (https://doi.org/10.1057/dam.2010.32).

Sallam, R., Sicular, S., den Hamer, P., Kronz, A., Schulte, W. R., Brethenoux, E., Woodward, A., Emmott, S., Zaidi, E., Feinberg, D., Beyer, M., Greenwald, R., Idoine, C., Cook, H., De Simoni, G., Hunter, E., Ronthal, A., Tratz-Ryan, B., Heudecker, N., Hare, J., and Clougherty Jones, L. 2020. "Top 10 Trends in Data and Analytics, 2020," Consortium Report, Consortium Report, Gartner Research. (https://www.gartner.com/en/doc/718161-top-10-trends-in-data-and-analytics-2020).

Sambamurthy, V., and Zmud, R. W. 1999. "Arrangement for Information Technology Governance: A Theory of Multiple Contingencies," *MIS Quarterly* (23:2), pp. 261–290.

Scheer, A.-W. 2001. *ARIS — Modellierungsmethoden, Metamodelle, Anwendungen*, (4th ed.), Berlin Heidelberg: Springer-Verlag. (https://www.springer.com/la/book/9783540416012).

Scheer, A.-W., and Schneider, K. 2006. "ARIS — Architecture of Integrated Information Systems," in *Handbook on Architectures of Information Systems*, International Handbooks on Information Systems, P. Bernus, K. Mertins, and G. Schmidt (eds.), Berlin, Heidelberg: Springer, pp. 605–623. (https://doi.org/10.1007/3-540-26661-5_25).

Schüritz, R., Brand, E., Satzger, G., and Bischhoffshausen, J. 2017. "How to Cultivate Analytics Capabilities within an Organization? - Design and Types of Analytics Competency Centers," in *Proceedings of the 25th European Conference on Information Systems (ECIS)*, Guimarães, Portugal, June 5, pp. 389–404.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D. 2015. "Hidden Technical Debt in Machine Learning Systems," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), Montreal, Canada, pp. 2503–2511.

Seddon, P. B., Constantinidis, D., Tamm, T., and Dod, H. 2017. "How Does Business Analytics Contribute to Business Value?: How Does Analytics Contribute to Business Value?," *Information Systems Journal* (27:3), pp. 237–269.

Shim, J. P., and Guo, C. 2015. "Big Data and Analytics: Issues, Solutions, and ROI," *Communications of the Association for Information Systems* (37), pp. 797–810.
Shvaiko, P., and Euzenat, J. 2005. "A Survey of Schema-Based Matching Approaches," in *Journal on Data Semantics IV* (Vol. 3730), S. Spaccapietra (ed.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 146–171. (https://doi.org/10.1007/11603412_5).
Sirmon, D. G., Hitt, M. A., and Ireland, R. D. 2007. "MANAGING FIRM RESOURCES IN DYNAMIC ENVIRONMENTS TO CREATE VALUE: LOOKING INSIDE THE BLACK BOX," *Academy of Management Review*, p. 21.
Sivarajah, U., Kamal, M. M., Irani, Z., and Weerakkody, V. 2017. "Critical Analysis of Big Data Challenges and Analytical Methods," *Journal of Business Research* (70), pp. 263–286.
Smith, H. A., and McKeen, J. D. 2008. "Developments in Practice XXX: Master Data Management: Salvation Or Snake Oil?," *CAIS* (23:4).
Spirig, J. 1987. "Compensation: The up-Front Issues of Payroll and HRIS Interface," *Personnel* (66:100), pp. 124–129.
Sprague, R. H. 1980. "A Framework for the Development of Decision Support Systems," *MIS Quarterly* (4:4), pp. 1–26.
Stonebraker, M., Bruckner, D., and Ilyas, I. F. 2013. "Data Curation at Scale: The Data Tamer System," in *CIDR Proceedings 2013*.
Stonebraker, M., and Ilyas, I. F. 2018. "Data Integration: The Current Status and the Way Forward," *IEEE Technical Committee on Data Engineering*, Data Engineering Bulletin, p. 7.
Strong, D., Lee, Y., and Wang, R. 1997. "Data Quality in Context," *Communications of the ACM* (40:5).
Tallon, P. P., Ramirez, R. V., and Short, J. E. 2013. "The Information Artifact in IT Governance: Toward a Theory of Information Governance," *Journal of Management Information Systems* (30:3), pp. 141–178.
Tiwana, A., Konsynski, B., and Venkatraman, N. 2013. "Special Issue: Information Technology and Organizational Governance: The IT Governance Cube," *Journal of Management Information Systems* (30:3), pp. 7–12.
Trieu, V.-H. 2017. "Getting Value from Business Intelligence Systems: A Review and Research Agenda," *Decision Support Systems* (93), pp. 111–124.
Tsichritzis, D., and Klug, A. 1978. "The ANSI/X3/SPARC DBMS Framework Report of the Study Group on Database Management Systems," *Information Systems* (3:3), pp. 173–191. (https://doi.org/10.1016/0306-4379(78)90001-7).
Van Alstyne, M., Brynjolfsson, E., and Madnick, S. 1995. "Why Not One Big Database? Principles for Data Ownership," *Decision Support Systems* (15:4), pp. 267–284.
Venkatesh, V., Brown, S., and Sullivan, Y. 2016. "Guidelines for Conducting Mixed-Methods Research: An Extension and Illustration," *Journal of the Association for Information Systems* (17:7), pp. 435–494. (https://doi.org/10.17705/1jais.00433).
Volkovs, M., Fei Chiang, Szlichta, J., and Miller, R. J. 2014. "Continuous Data Cleaning," in *2014 IEEE 30th International Conference on Data Engineering*, Chicago, IL: IEEE, March, pp. 244–255. (https://doi.org/10.1109/ICDE.2014.6816655).
Wamba, S. F., Akter, S., Edwards, A., Chopin, G., and Gnanzou, D. 2015. "How 'Big Data' Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study," *International Journal of Production Economics* (165), pp. 234–246. (https://doi.org/10.1016/j.ijpe.2014.12.031).
Wang, R. Y. 1998. "A Product Perspective on Total Data Quality Management," *Communications of the ACM* (41:2), pp. 58–65.
Wang, R. Y., Storey, V. C., and Firth, C. P. 1995. "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering* (7:4), pp. 623–640. (https://doi.org/10.1109/69.404034).
Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), pp. 5–33.

Watson, H. 2014. "Tutorial: Big Data Analytics: Concepts, Technologies, and Applications," *Communications of the Association for Information Systems* (34:1).
Watson, H. J. 2009. "Tutorial: Business Intelligence – Past, Present, and Future," *Communications of the Association for Information Systems* (25:487–510), p. 26.
Watson, H. J. 2017. "Preparing for the Cognitive Generation of Decision Support," *MIS Quarterly* (16:3).
Watson, H. J. 2019. "Update Tutorial: Big Data Analytics: Concepts, Technology, and Applications," *Communications of the Association for Information Systems*, pp. 364–379.
Watson, H. J., Fuller, C., and Ariyachandra, T. 2004. "Data Warehouse Governance: Best Practices at Blue Cross and Blue Shield of North Carolina," *Decision Support Systems* (38:3), pp. 435–450.
Weber, K., Otto, B., and Österle, H. 2009. "One Size Does Not Fit All-A Contingency Approach to Data Governance," *J. Data and Information Quality* (1:1), pp. 1–27.
Wernerfelt, B. 1984. "A Resource-Based View of the Firm," *Strategic Management Journal* (5:5), pp. 171–180.
Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* (3:1), pp. 1–9. (https://doi.org/10.1038/sdata.2016.18).
Winkler, T. J., and Wessel, M. 2018. "A Primer on Decision Rights in Information Systems: Review and Recommendations," in *Proceedings of the 39th International Conference on Information Systems (ICIS)*, San Francisco, California, USA, December 13.
Winter, R. 2008. "Enterprise-Wide Information Logistics: Conceptual Foundations, Technology Enablers, and Management Challenges," in *Proceedings of ITI 2008*, , June, pp. 41–50. (https://doi.org/10.1109/ITI.2008.4588382).
Winter, R., and Meyer, M. 2001. "Organization of Data Warehousing in Large Service Companies - A Matrix Approach Based on Data Ownership and Competence Centers," *Journal of Data Warehousing* (6:4), pp. 23–29.
Wixom, B., and Ross, J. 2017. "How to Monetize Your Data," *MIT Sloan Management Review* (58:3).
Wixom, B., and Watson, H. 2010. "The BI-Based Organization:," *International Journal of Business Intelligence Research* (1:1), pp. 13–28. (https://doi.org/10.4018/jbir.2010071702).
Wu, R., Zhang, A., Ilyas, I. F., and Rekatsinas, T. 2020. "Attention-Based Learning for Missing Data Imputation in HoloClean," in *Proceedings of the 3rd MLSys Conference*, Austin, Texas, p. 19.
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. 2016. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *ArXiv:1609.08144 [Cs]*. (http://arxiv.org/abs/1609.08144).
Yin, R. 2003. Case Study Research: Design and Methods, Third Edition, Applied Social Research Methods Series, Vol 5, London, UK: Sage Publications, Inc.
Yin, R. K. 2003. *Case Study Research - Design and Methods*, (3rd ed.), Sage Publications.
Zaidi, E., De Simoni, G., Edjlali, R., and Duncan, A. D. 2017. "Data Catalogs Are the New Black in Data Management and Analytics," Consultancy Report, Consultancy Report, Gartner,

December 13. (https://www.gartner.com/doc/reprints?id=1-4MKJU2Y&ct=171220&st=sb&submissionGuid=12d68804-ceec-454e-b412-a66bdff38e2e).
Zeng, J., and Glaister, K. W. 2018. "Value Creation from Big Data: Looking inside the Black Box," *Strategic Organization* (16:2), SAGE Publications, pp. 105–140. (https://doi.org/10.1177/1476127017697510).
Zhu, H., Madnick, S., Lee, Y., and Wang, R. 2014. "Data and Information Quality Research: Its Evolution and Future," in *Computing Handbook, Third Edition*, H. Topi and A. Tucker (eds.), Chapman and Hall/CRC, pp. 16-1-16–20. (https://doi.org/10.1201/b16768-20).
Zoumpoulis, S. I., Vlachos, M., Freris, N. M., and Lucchese, C. 2014. "Right-Protected Data Publishing with Provable Distance-Based Mining," IEEE Transactions on Knowledge and Data Engineering (26:8), pp. 2014–2028. (https://doi.org/10.1109/TKDE.2013.90).

Building Business Intelligence & Analytics Capabilities - A Work System Perspective

Martin Fadler and Christine Legner

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

Presented at International Conference on Information Systems 2020

Abstract: Although enterprises believe that they can achieve a competitive advantage with big data and AI, their analytics initiatives' success rate still lags behind expectations. Existing research reveals that value creation with business intelligence and analytics (BI&A) is a complex process with multiple stages between the initial investments in BI&A resources and ultimately obtaining value. While prior research mostly focused on value generation mechanisms, we still lack a thorough understanding of how enterprises actually build BI&A capabilities. We explain the process in our research using work system theory (WST). Based on case studies and focus groups, we identify four prevalent BI&A capabilities: reporting, data exploration, analytics experimentation, and analytics production. For each identified BI&A capability, we derive patterns for BI&A resource orchestration, using the WST lens. Our findings complement the BI&A value creation research stream by providing insights into capability building.

Keywords: Business Intelligence, Analytics, Big Data Analytics, Capability Building, Resource Orchestration, Work System Theory

Table of Contents

List of Figures

List of Tables

1 Introduction

Digital applications and connected devices create an ever-increasing amount of data (Chen et al., 2012), a phenomenon known as big data. This phenomenon, combined with major breakthroughs in data infrastructure technologies and with artificial intelligence (AI) proliferation, allows enterprises to identify new data monetization opportunities, which have resulted in improved existing business processes, while simultaneously innovating and creating new business models (Wixom and Ross, 2017). Although enterprises believe that they can achieve a competitive advantage with big data and AI (Ransbotham et al., 2019; Ransbotham and Kiron, 2017), their analytics initiatives' success rate still lags behind expectations and many struggle to obtain a return on investment (Davenport and Bean, 2019; Grover et al., 2018; Shim and Guo, 2015). This lag raises fundamental questions about how investments in analytics create business value and reflects the ongoing debate on information systems' (IS) overall business value (Kohli and Grover, 2008; Schryen, 2013).

In the context of business intelligence and analytics (BI&A), researchers have analyzed value creation by using different theoretical lenses with various focuses and scopes, including business intelligence systems (Trieu, 2017), business analytics (Seddon et al., 2017), and big data analytics (Grover et al., 2018). These studies identify how BI&A creates organizational and strategic value for enterprises, but also reveal that value creation materializes by means of complex processes from initial investments in analytics resources in order to obtain the actual value. While this stream of research provides important insights into the relationship between resources, capabilities, and value creation, its focus is mainly on the value generation mechanisms. With reference to the research framework by Grover et al. (2018) research framework, the capability building process has received far less attention than the capability realization process. Consequently, "*little is known so far about the processes and structures necessary to orchestrate these resources into a firm-wide capability*" (Mikalef et al. 2018, p. 569).

To address this gap, we ask the following research question:

RQ: *How do enterprises build business intelligence and analytics capabilities?*

In line with literature, we use BI&A as unifying term to designate "*the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions*" (Chen et al. 2012, p. 1166). We consider big data analytics (BDA), resulting from increasing data volumes

and variety, evolving data infrastructure technologies and the proliferation of artificial intelligence, as evolution of BI&A or "BI&A 3.0" (Chen et al., 2012).

To explain how BI&A capabilities are built, our research uses a work system theory (WST) lens (Alter, 2013), which *"replaces the prevailing system-as-technical-artifact perspective with a genuine system perspective for focusing on IT-reliant systems in organizations"* (p. 74). Based on a multi-method research approach (Venkatesh 2016) comprising case studies and focus groups, we identify four prevalent capabilities in the BI&A context: *reporting*, *data exploration*, *analytics experimentation*, and *analytics production*. In respect of each identified BI&A capability, we derive BI&A resource orchestration patterns by analyzing data collected from the field through a WST lens. Our findings complement the research stream on BI&A value creation with insights into capability building. The findings also contribute to addressing the questions outlined in the research agenda by Grover et al. (2018) with regard to building BDA capabilities, i.e. the ability to integrate, disseminate, explore, and analyze big data. Our results not only inform the academic research community, but are also relevant for practitioners, who can use the identified BI&A work systems to define roles, processes, and technologies, thereby laying the foundation for value generation with BI&A.

The remainder of this paper is structured as follows: first, we summarize the existing research on BI&A and identify the research gap. Second, we outline our multi-method research design and the WST's theoretical lens. Third, we use WST to identify BI&A capabilities and the corresponding capability building process. Afterwards, we discuss our findings, which leads to our conclusion and outlook for future research.

2 Background

Since electronic data processing's early beginnings, digital data have been analyzed to improve businesses' efficiency and effectiveness. The field has evolved continuously, currently encompassing traditional approaches to business intelligence (Chen et al., 2012), as well as innovative ways of analyzing big data and enabling AI (Davenport, 2018). In recent years, researchers have created a fundamental understanding of BDA as an emerging field. This understanding has resulted in studies clarifying BDA concepts, technologies, and applications (Watson, 2014; Watson, 2019), as well as integrating BDA into the broader BI&A field (Chen et al., 2012).

Two perspectives dominate the IS literature on BI&A: the first stream sheds light on the evolution of BI&A in enterprises, with a focus on key concepts, applications, and technologies. The second stream aims to explain value generation in the BI&A context. While these findings identify different stages, ranging from investments in BI&A resources to obtaining value, they provide few insights into the way enterprises orchestrate their resources and build BI&A capabilities.

2.1 Evolution of Business Intelligence and Analytics

Since its early applications in the 1970s in the form of decision support systems (DSS), the BI&A field has not stopped evolving. The first DSS generation used a dedicated data repository and model basis (Sprague, 1980) to calculate the key performance indicators and deliver reports on historic data in structured formats. This application-centric architecture was subsequently replaced by new DSS applications, such as executive information systems, and dashboards/scorecards (Watson, 2014). Data warehouses allowed companies to integrate data from multiple operational systems in a pre-defined structure and to support a wide variety of applications simultaneously, such as queries, online analytical processing (OLAP) or data mining. Establishing a central repository for all enterprise data had the advantage of simplifying the BI&A delivery (Watson, 2009). As early as in 1989, Howard Dresner coined the term business intelligence (BI) as "*a broad category of applications, technologies, and processes for gathering, storing, accessing, and analyzing data to help business users make better decisions*" (Watson 2009, p. 491). Enterprise data warehouses allowed companies to process data in real-time and thereby support decision-making not only at strategic/tactical level, but also at operational level (Watson, 2009).

The advent of the internet made new data sources available through web applications. Mining social graphs or even, for instance, their customers' opinions allowed enterprises to significantly improve their understanding of their environment. Since analytics capabilities have gained increasing importance, the term (business) analytics is often used in conjunction or interchangeably with business intelligence (Chen et al., 2012; Davenport, 2006). With the emergence of smartphones and the ubiquity of sensors embedded in connected devices, data are collected on a more granular level than before. This change allows enterprises to accurately trace and analyze their business operations, but also requires them to rethink the way they manage data and deliver BI&A. Today, data "*are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization technologies*" (Chen et al. 2012, p. 1166). Traditional enterprise data warehouses cannot cope with big data requirements, due to their lack of flexibility in terms of modifying data structures and dealing with multiple data formats (Jukić et al., 2015; Sivarajah et al., 2017). Companies are therefore extending their existing data infrastructures to build more comprehensive enterprise analytics platforms. The latter comprise data lakes, which store data in a raw format without a pre-defined structure, to enable data exploration and experimentation (Farid et al., 2016; Madera & Laurent, 2016; Watson, 2017) as well as analytics products with capabilities that clearly go beyond the mere aggregation and visualization of data and also comprise artificial intelligence (Watson, 2017). In this context, experts argue that machine learning applications' strong reliance on data might lead to high technical debts (Sculley et al., 2015).

2.2 Value Generation through Business Intelligence and Analytics

Even though enterprises understand that they can achieve a competitive advantage with big data and advanced analytics (Ransbotham and Kiron, 2017), their analytics initiatives' success rate lags behind expectations and many struggle to obtain a return on investment (Davenport and Bean, 2019; Grover et al., 2018; Shim and Guo, 2015). This struggle raises questions about how BI&A investments create value and reflects the general debate on IS' business value (Kohli and Grover, 2008; Schryen, 2013). Different models have been proposed (see Table 11) to explain the value creation of business intelligence systems (Trieu, 2017), business analytics (Seddon et al., 2017), and big data analytics (Grover et al., 2018).

Trieu (2017) introduce a BI business value framework to integrate findings from the fragmented literature and to guide researchers. Based on the seminal IT business value process model by Soh and Markus (1995), she shows that BI creates value in a chain of required conditions ranging

from BI investments to BI assets, to BI impacts. She distinguishes three core processes: (1) the BI conversion process, which converts BI investments to BI assets through BI management and additional investments in non-BI resources. (2) The BI use process: once BI investments have been converted, BI assets can be used to generate BI impacts. BI effective/ineffective use patterns affect this generation process's performance. (3) Competitive process: this process transforms the BI impacts into organizational performance, which affects the firm's competitive position. Trieu (2017) extends the adapted Soh and Markus (1995) view by means of the findings by Schryen (2013) and Melville et al. (2004) by including context/environmental factors (firm, industry, country) and latency effects, which affect the BI business value generation process.

Based on an analysis of 16 models from the literature, Seddon et al. (2017) derive a business analytics success model comprising process and variance models. The process model builds on the observation that "prime drivers of business value from business analytics are actions driven by new insights and improved decision making" (ibid p. 244). The enabling technology and analytical specialists are the analytical resources used "by people in many parts of the organization" to generate insights and make decisions. Decisions lead to value-creating actions that either change or use the existing organizational resources and lead to organizational benefits from the analytics' use. The variance model provides a complementary view of the process model, and comprises a long-term organizational benefits model and a short-term project model. The short- and long-term organizational benefits depend on various factors. In the short-term project model (S), these factors include (S1) BA tools' functional fit, (S2) readily available high-quality data, (S3) analytical people, and (S4) overcoming organizational inertia. In the long-term model (L), the benefits depend on (L1) the analytics leadership, (L2) enterprise-wide analytics orientation, (L3) well-chosen targets, (L4) the extent to which evidence-based decision making is embedded in the organization's "DNA," and (L5) the on-going business analytics improvement projects.

Table 11. Prior studies on BI&A value generation

	BI business value framework (Trieu, 2017)	**Business analytics success model** (Seddon et al., 2017)		**BDA value creation framework** (Grover et al., 2018)
Scope	Business intelligence systems	Business analytics		Big data analytics
Business value	Organizational performance	Organizational benefits from the use of analytics from the senior management perspective		Strategic business value
Theory type	Integrated process and variance model	Process model	Variance model	Integrated process and variance model
Value generation approach	BI conversion process BI use process Competitive process	Business analytics' value creation process paths (P1, P2, P3)	Long-term organizational benefits model (L) Short-term model (S): factors driving benefits from each project	Capability building process Capability realization process
Constructs	Process: - BI investments - BI assets - Business impact Variance: - Environmental factors - Latency effects	Analytical resources Use analytical resources Insight(s) Decision(s) Value-creating actions Organizational resources	Long term: (L1) analytic leadership, (L2) enterprise-wide analytics orientation, (L3) well-chosen targets, (L4) the evidence-based decision making's extent, and (L5) on-going business analytics' improvement projects Short term: (S1) BA tools' functional fit, (S2) readily available high-quality data, (S3) analytical people, and (S4) overcoming organizational inertia	Process: - Big data infrastructure - BDA capabilities - Value creation mechanisms - Value targets Variance: - Moderating factors
Theoretical background	IT value models of Soh and Markus (1995), Schryen (2013) and Melville et al. (2004)	16 models of factors affecting organizational benefits from business analytics, e.g. Davenport et al.'s DELTA model of business analytics success factors		Resource based view Dynamic capabilities IT value models of Soh and Markus (1995), and Melville et al. (2004)

Grover et al. (2018) investigate how BDA creates strategic value from the resource-based view's lens. In their study, they focus on descriptive, predictive and prescriptive analytics, as well as on an analytics portfolio comprising text, predictive, audio, video, social media, geographic, streaming, and graph analytics. These authors suggest a conceptual framework that builds on dynamic capabilities' general framing and on IT value models that Soh and Markus (1995), as well as Melville et al. (2004), proposed. According to this framework, value is created through two main processes: building BDA capabilities and realizing BDA capabilities. Building BDA

capabilities involves investment in data, technological, and human resources to establish a BDA infrastructure (big data asset, analytics portfolio, and human talent). The latter activity leverages this BDA infrastructure to develop valuable BDA capabilities, i.e. the "ability to integrate, disseminate, explore, and analyze big data" (p. 398). The realization of BDA requires six distinct value creation mechanisms "*that mediate the linkage between BDA capabilities and value targets*": (1) transparency and access, (2) discovery and experimentation, (3) prediction and optimization, (4) customization and targeting, (5) learning and crowd-sourcing, and (6) continuous monitoring and proactive adaptation. The value targets could result in functional or symbolic strategic value.

2.3 Research Gap

The presented models help us understand how investments in BI&A resources create business value in terms of strategic and organizational performance. These models also reveal that value creation is a complex process with multiple stages between the initial investments in BI&A resources and eventually obtaining actual value. However, prior research mostly focused on value generation mechanisms, but do not explain how enterprises actually structure and deploy their BI&A resources to build BI&A capabilities. According to Grover et al. (2018), the latter remains an important research topic, because "*without appropriate organizational structures and governance frameworks in place, it is impossible to collect and analyze data across an enterprise and deliver insights [in]to where they are most needed*" (p. 417). Moreover, "little is known so far about the processes and structures necessary to orchestrate these resources into a firm-wide capability" (Mikalef et al. 2018, p. 569). Among the few studies, Schüritz et al. (2017) analyze analytics competence centers to identify organizational design patterns, while Kettinger et al. (2019) investigate how to build an information management capability to develop guidelines for senior executives. Although both studies explain BI&A capability building, they focus only on partial aspects and follow a different research aim. We conclude that existing research remains fragmented and without a clear theoretical framing.

Work system theory (WST) is a promising lens for studying capability building in the context of BI&A. A work system is a "*system in which human participants and/or machines perform work (processes and activities) using information, technology, and other resources to produce specific product/services for internal or external customers*" (Alter 2013, p.75). The work system perspective therefore helps us understand how resources (participants, information, technologies) are orchestrated (by means of processes/activities) to build capabilities (products/services for customers). Several researchers have applied WST to specific BI&A

applications. Alter (2004), for instance, analyzes a decision support system to demonstrate the WST perspective's usefulness. Heart et al. (2018) use the WST to design and implement a big data analytics tool for improving clinical decisions. Marjanovic (2016) investigates BI&A-supported, knowledge-intensive business processes by means of the WST lens. We conclude that using the WST lens is a promising approach to systematically analyze how enterprises orchestrate their tangible and intangible BI&A resources, and build BI&A capabilities.

3 Methodology

We use a multi-method research design (Venkatesh 2016) to investigate *how enterprises build their BI&A capabilities*. Our research activities started in February 2019, when we formed an expert group to investigate BI&A challenges as part of a multi-year research program on data management. Over a period of one year, we worked closely with 11 BI&A experts from seven high-profile European companies. All of the experts represent large corporations from a diversity of industries with ongoing initiatives regarding leveraging BI&A. These experts are responsible for establishing governance structures (including the definition of roles, responsibilities, and processes) and have a broad overview of the BI&A in their respective company. This setup provided us with unique access to field data from ongoing BI&A initiatives in European companies.

We collected data by means of four case studies and five focus group meetings, which we then analyzed through the WST lens. Using the two different qualitative data collection procedures (case studies and focus group meetings) allowed us to gain a broader understanding of the current state of BI&A in enterprises. The four case studies specifically allowed us to study companies with a comparable maturity. After reflecting on the four cases with a broader group of experts and in the context of the literature, we generalized our findings in the form of BI&A work systems.

3.1 Case Studies

From discussions in the expert group, we selected four (of the seven) companies for a detailed investigation of their BI&A environment and management approach (see Table 12). These four case companies have an enterprise data warehouse and an enterprise data lake as a BI&A infrastructure; they also have data scientist teams that explore and experiment with data. To enable their organizations to work with BI&A at scale, they have defined roles, processes, and responsibilities as part of their governance organization. Since each case company has a

relatively high BI&A maturity and belongs to a different industry, the case selection process followed a literal replication logic allowing the results to be analytically generalized (Benbasat et al., 1987; Yin, 2003).

We gathered information on each case company from multiple sources, i.e. primary sources (interviews) and secondary sources (internal documents), which allowed triangulation and ensured the construct validity (Yin, 2003). As a starting point, we conducted an initial semi-structured interview with the key informants to gain an understanding of their roles, as well as their companies' processes, technologies, and infrastructures. These interviews gave us the opportunity to understand the challenges and approaches in greater depth. In parallel, we collected primary data through the internal documents that the firms provided (e.g. BI&A platform designs, role models, and organizational structures). These documents not only informed us about their approach, but also about the context and related topics, such as the technical infrastructure, as well as the established roles and processes.

Table 12. Case companies

Company	Industry	Revenue / # Employees	Key informants	BI&A context
A	Consumer goods	$50–100 b / ~80 000	Data governance manager, enterprise data architect	Organization: central data and analytics management organization Infrastructure: central big data platform for the innovation and industrialization of analytics use cases
B	Public transportation	$1–50 b / ~35 000	Leader business information management, data governance manager, big data platform architect	Organization: central data management organization and central/decentralized data science team Infrastructure: corporate data lake for data exploration/experimentation and the operation of analytics use case
C	Industry products	$50–100 b / ~110 000	Project manager data lake	Organization: Central data management organization and advanced analytics group Infrastructure: Operation of multiple data lakes and data warehouses
D	Consumer goods	$1–50 b / ~30 000	Head of data and analytics, head of data governance	Organization: Central data and analytics management organization with a high business intelligence maturity Infrastructure: Operation of one_central enterprise data warehouse with extensions to undertake analytics

3.2 Focus Group Meetings

The experts met physically five times between February 2019 and February 2020 (see Table 13). The first meeting was held in February 2019 to discuss the challenges of managing data lakes compared to traditional BI environments. The group realized and agreed that established approaches could not be transferred to data lake environments where data enable data

exploration and experimentation. It became clear that the key challenge lies in managing both environments simultaneously, and that BI&A management should encompass the complete "enterprise analytics platform," meaning all the components that deliver BI&A products, including existing BI and data lake environments. The participants also concluded that they needed a comprehensive approach covering the technological and the organizational aspects. In respect of the technological aspects, the participants called for an understanding and descriptions of the existing and the emerging components of the "*enterprise analytics platform*." In respect of the organizational aspects, the participants called for a clarification of the roles, responsibilities, and processes. Based on the findings of this initial focus group, we conducted five subsequent focus group meetings. At each meeting, we investigated one crucial topic in depth in order to contribute to the larger picture of how enterprises should build their BI&A capabilities.

Table 13. Focus group meetings

Meeting	Date	Participants	Duration	Topic
1	Feb 2019	11 BI&A experts from seven high-profile European companies	2 x 3 hours	BI&A challenges
2	Apr 2019		2 x 3 hours	BI&A products
3	Jun 2019		2 x 3 hours	BI&A technologies and infrastructure
4	Sept 2019		2 x 3 hours	BI&A roles and responsibilities
5	Feb 2020		2 x 3 hours	BI&A processes

3.3 Theoretical Integration

To integrate our findings from the field with those from the literature, we used the theoretical WST lens and analyzed our data according to the work systems framework's and lifecycle model's components. We chose the WST (Alter, 2013) for the following two reasons: First, it *"replaces the prevailing system-as-technical-artifact perspective with a genuine system perspective for focusing on IT-reliant systems in organizations"* (Alter 2013, p. 74). Second, the WST provides a suitable, systematic approach to describe how tangible and intangible resources are orchestrated in an enterprise. In the context of BI&A, it helps provide an understanding of how the required capabilities are built.

The WST comprises three core components: the work system definition, the work system framework, and the work system lifecycle model. A work system is defined as a "*system in which human participants and/or machines perform work (processes and activities) using information, technology, and other resources to produce specific product/services for internal or external*

customers" (Alter, 2013). The work systems' elements and the relationships between them are described by means of the work system framework (see Figure 6). The C*ustomers*, which are displayed at the top, are the receivers of *Products/Services*, which they use "*for purposes other than performing work activities within the work system*" (Alter, 2013). *Products/Services* deliver a certain value to these *Customers* and are a direct outcome of the work system. These products or services are created through a certain set of *Processes/Activities,* which requires *Participants*, *Information,* and *Technologies*. *Participants* are responsible for at least one *Process/Activity*, but can simultaneously be a *Customer*. *Information* represents the "*informational entities that are used, created, captured, transmitted, stored, retrieved, manipulated, updated, displayed, and/or deleted by processes/activities*" (Alter, 2013). *Technologies* are used in *Processes/Activities* to provide customers with *Products/Services*. While the previous elements describe a work system's key elements from an inside perspective, the elements *Infrastructure*, *Strategies,* and *Environment* influence the work system from the outside. *Infrastructure* comprises the resources shared between work systems. *Strategies* influence the work system's the lifecycle and may include the companies' strategy, the business unit strategy, and the work system strategy. The *Environment* encompasses the "*relevant organizational, cultural, competitive, technical, regulatory, and demographic environment within which the work system operates, and that affects the work system's effectiveness and efficiency.*"

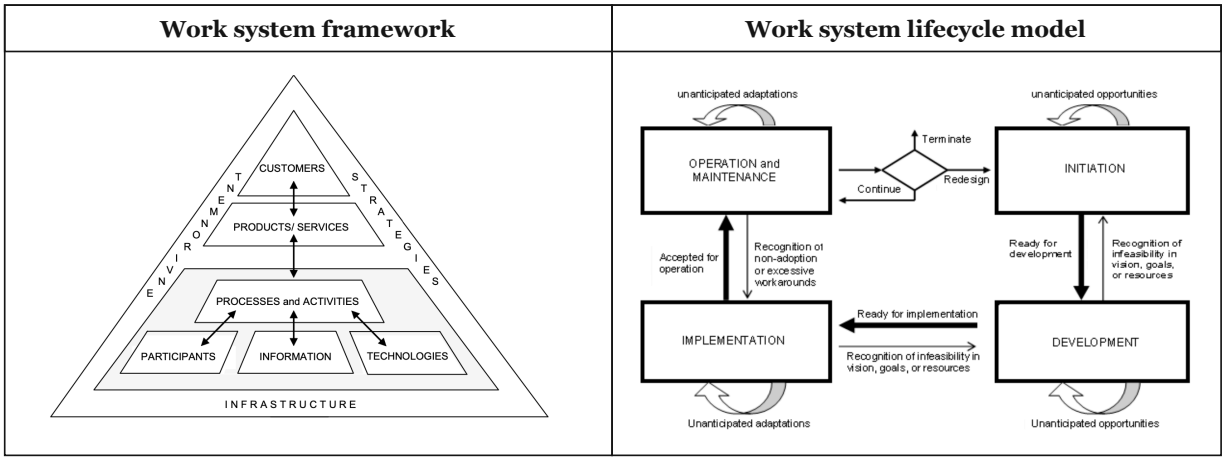


Figure 6. Work system framework and lifecycle (Alter 2013)

While the work system framework is a snapshot of a certain point in time, the work system lifecycle model describes how a work system evolves over time. This system comprises four phases representing planned changes: *Initiation*, *Development*, *Implementation*, and *Operation & Maintenance*. The need for a particular work system is addressed and specified in the *Initiation* phase. After the specification of the requirements, the resources for implementing the work

system are created and allocated in the *Development* phase, which also includes, for instance, the development of software. In the *Implementation* phase, the work system is implemented in the organization through change management and, for instance, through the work system participants' training. After this phase, the work system is in the *Operation & Maintenance* phase.

4 Building Business Intelligence and Analytics Capabilities with Work Systems

By analyzing our data and the literature, we identify four types of BI&A capabilities prevalent in the case companies and extend prior studies (see Table 14): *Reporting*, *Data exploration*, *Analytics experimentation*, and *Analytics production.* Based on the WST, these BI&A capabilities serve specific Customers and Products/Services, which we will present below, together with evidence from the cases and links to the BI&A literature.

4.1 Four Types of Business Intelligence and Analytics Capabilities

The reporting capability comprises periodically providing business users with reports, as well as digital dashboards summarizing the business transactions in the form of key performance indicators and visualizations (Chen et al. 2012; Watson 2009). While companies A and D solely focus on managers as end users, companies B and C also address users on an operational level, who could, for instance, be shop-floor workers at their production facilities. This capability creates value in the form of *Transparency and access* and *Continuous monitoring and proactive adaptation* according to the mechanisms that Grover et al. (2018) suggest.

The *Data exploration* capability allows in-depth and flexible data analysis "without an a priori understanding of what patterns, information, or knowledge it might contain" (Baker et al. 2009, p. 534). In a dedicated environment, business users analyze their domain of interest's data with self-service BI tools to explore and make sense of the data in the investigation context (Alpar and Schulz, 2016). Companies A and D call this BI&A product "interactive visualization," while Company C uses the term "self-service data" and Company B "agile reporting." In keeping with the mechanisms that Grover et al. (2018) suggest, this capability creates value in the form of *Discovery and experimentation and Transparency and access*.

Table 14. BI&A capabilities in case companies

BI&A capability	Reporting	Data exploration	Analytics experimentation	Analytics production
Value creation mechanism (Grover et al., 2018)	Transparency and access/ Continuous monitoring and proactive adaptation	Discovery and experimentation/ Transparency and access	Discovery and experimentation/ Learning and crowd-sourcing	Prediction and optimization/ Customization and targeting
Customers	Business user	Business user	Data scientist	Business user
Products/ Services	Periodically providing reports or real-time updating of dashboards summarizing business transactions in the form of key performance indicators and visualizations (Chen et al., 2012; Watson, 2009).	Providing an environment to explore and make sense of data in a certain domain of interest, without a priori understanding of what information it might contain for the issue being investigated (Alpar & Schulz, 2016; Baker et al., 2009).	Providing a virtual sandbox environment to develop and test analytics use cases and prove their feasibility (H. Watson, 2014).	Providing an up-to-date analytics model in a business application (Watson, 2019).
Company A	Management reporting in each business function	Interactive visualization	Data science labs to develop predictive and prescriptive analytics	Data products industrialize predictive and prescriptive models
Company B	Operational and management reporting	Agile reporting to flexibly analyze different types of data	Data labs/user home for data exploration, experimentation, research, and development	Data apps industrialize the tested algorithms in a data lab
Company C	Reporting for end- and key-users	Self-service data modelling and visualization	Data lab environment	Application programming interfaces to enhance other applications
Company D	Dashboards for corporate performance management	Interactive visualization in data marts	Development of data products	Data mart provides access to deployed, advanced analytical models

The *Analytics experimentation* capability allows enterprises the possibility to develop analytics use cases and prove their feasibility. A data scientist should develop and test analytics algorithms in a virtual sandbox environment (Watson 2014). All the case companies provide access to a dedicated environment in order to allow access to datasets in their raw format for research and development purposes. While companies A-C run their analytics experimentations on a data lake, Company D uses a data mart connected to its enterprise data warehouse. In keeping with the mechanisms that Grover et al. (2018) suggest, this capability creates value in the form of *Discovery and experimentation* and *Learning and crowd-sourcing*.

The analytics models that prove feasible are deployed and made accessible with the *Analytics production* capability, which in turn ensures that the analytics models remain up-to-date throughout their lifecycle (Watson, 2019). A business user accesses an analytics model in business applications. Companies A-C operate their *Analytics production* in a dedicated environment on the data lake. For instance, Company C provides access to analytics models via an application programming interface (APIs). Company D operates advanced analytical models with a data mart connected to the enterprise data warehouse. In keeping with the mechanisms that Grover et al. (2018) suggest, this capability creates value in the form of *Prediction and optimization* and *Customization and targeting*.

5 Business Analytics and Intelligence Work Systems

In the following, we describe how the identified BI&A capabilities are built using the work system framework and lifecycle. We describe each work system's *Customers* and *Products/Services*, the required resources with *Participants*, *Information,* and *Technologies*, and resource orchestration with *Processes/Activities*.

5.1 Reporting Work System

The *Reporting* work system (see Table 15) enables enterprises to create transparency and monitor business operations to improve operational and strategic decision making. Operational data (e.g. business transactions or machine data) are aggregated on a continuous basis to calculate key performance indicators and create visualizations. The operational level's time horizon of analysis is usually shorter than that of for the management level.

The trigger is usually an information need that a business user expresses. In the *Initiation* phase, the business user specifies the analytics product (here, the key performance indicators and the report) by defining the decisions that the report needs to support, including aspects such as the frequency or form of delivery. An analytics expert with business domain knowledge usually supports the report specification process.

In the *Development* phase, an analytics expert identifies the required data for the report and assesses whether they are available in the enterprise data warehouse or in a data catalog. If the data are not available or accessible in the required form, they first need to be onboarded to the data warehouse. In this case, a data architect identifies and models the data according to the reports' data requirements. In addition, a data engineer must extract, transform, and load (ETL) the data into the data warehouse/data mart. If the data are available and accessible in the

required form in the data warehouse, either the analytics expert or a data analyst develops the report directly by means of a BI tool. Once a first version of the report has been created, the Business user validates it. The analytics expert creates material with which to train others to use the report.

In the *Implementation* phase, the data engineer deploys the report, business users are trained to use the report, and the report is documented in a data catalog, along with explanations of its general mechanism and information on access to report and training material.

In the *Operation & Maintenance* phase, a data engineer monitors the ETLs, while a data steward monitors the data quality in general and the report's use in particular.

Table 15. Reporting work system

<table>
<tr><th>Customers</th><th colspan="2">Products/Services</th></tr>
<tr><td>Business user</td><td colspan="2">Periodically providing reports or real-time updating of dashboards summarizing business transactions in the form of key performance indicators and visualizations (Chen et al., 2012; Watson, 2009).</td></tr>
<tr><td colspan="3">Major activities/processes</td></tr>
<tr><td colspan="3">1. Initiation:
• Report specification by BUS/AEX (DCA).
2. Development:
• Data onboarding by DAR/DEN (DWH, DMA, ETL, DCA), and
• report development by AEX/DAR/BUS (BIT).
3. Implementation:
• Deployment of report by DEN (DMA, BIT),
• training of BUSs and report documentation by AEX (DCA).
4. Operation & maintenance:
• Monitoring of ETL by DEN (MOT), and
• data quality and report use by DST (DCA).</td></tr>
<tr><th>Participants</th><th>Information</th><th>Technologies</th></tr>
<tr><td>Business user (BUS)
Analytics expert (AEX)
Data analyst (DAN)
Data architect (DAR)
Data engineer (DEN)</td><td>Operational data
Historic/Real-time
Pre-defined structure
Domain knowledge
Data pull</td><td>BI tools (BIT)
Extract, transform, load (ETL)
Data catalog (incl. business glossaries and data dictionaries) (DCA)
Monitoring tools (MTO)
Data mart (DMA)
Data warehouse (DWH)</td></tr>
</table>

5.2 Data Exploration Work System

The *Data exploration* work system (see Table 16) allows, depending on the issue being investigated, the flexible analyzing a certain domain of interest's data and from different perspectives. This system supports decisions requiring an in-depth data analysis. Depending on the domain, data can stem from various sources and be of different types. Data are pushed to the access tools and give customers the flexibility to select and analyze the required data themselves. This work system relies on data warehouse architecture with online analytical

processing and powerful data visualization tools to allow data to be explored in a self-service way.

In the *Initiation* phase, a business user identifies and specifies data requirements and access modalities with the support of an analytics expert, which support the analysis of the domain of interest and support the task at hand.

In the *Development* phase, the required data are onboarded to the data warehouse. First, a data architect identifies and models the required data. Thereafter, a data engineer implements the extract, transform, and load process according to the data models that the data architect provides. The BI tool is set up according to the specification. Finally, the analytics expert and data analyst create training material.

Table 16. Data exploration work system

<table>
<tr><td>Customers</td><td colspan="2">Products/Services</td></tr>
<tr><td>Business user</td><td colspan="2">Providing an environment to explore and make sense of data in a certain domain of interest, without a priori understanding of what information it might contain for the issue being investigated (Alpar and Schulz, 2016; Baker et al., 2009).</td></tr>
<tr><td colspan="3">Major activities/processes</td></tr>
<tr><td colspan="3">1. Initiation:
• Specification of requirements by BUS/AEX (DCA).
2. Development:
• Data onboarding by DAR/DEN/DOW (DWH, DMA, ETL, DCA),
• BI tool setup by DAN/DEN (DMA, BIT, OLAP), and
• creation of training material by DAN/AEX
3. Implementation:
• Access provisioning by DAN (BIT),
• training of BUSs by DAN/AEX, and
• documentation by AEX (DCA).
4. Operation & maintenance:
• Monitoring of ETL by DEN (MTO) and
• monitoring of data quality by DST (DCA).</td></tr>
<tr><td>Participants</td><td>Information</td><td>Technologies</td></tr>
<tr><td>Business user (BUS)
Analytics expert (AEX)
Data analyst (DAN)
Data architect (DAR)
Data owner (DOW)</td><td>Data of domain of interest (e.g. certain business events)
Historic/real-time
Pre-defined structure
Data push
Domain knowledge</td><td>BI tools (BIT)
Online analytical processing (OLAP)
Extract, transform, load (ETL)
Data catalog (incl. business glossaries and data dictionaries) (DCA)
Monitoring tools (MTO)
Data mart (DMA)
Data warehouse (DWH)</td></tr>
</table>

In the *Implementation* phase, data analysts and analytics experts train business users in conducting descriptive and diagnostic analytics with the BI tool. This capability requires customers to be data literate. The analytics expert documents the data and training material in a data catalog.

In the *Operation & Maintenance* phase, data are continuously pushed to the access tool through the extract, transform, and load process which the data engineer monitors. A data steward takes data quality measures and ensures that the data are fit for purpose.

5.3 Analytics Experimentation Work System

The *Analytics experimentation* (see Table 17) work system provides the possibility to test analytics use cases' feasibility through iterative experiments. A sample dataset is made available in a dedicated environment where experts can access it by, for instance, using interactive programming and development tools. This sample dataset comes with its own requirements and, for instance, requires labels for machine learning tasks.

In the *Initiation* phase, the analytics use case is specified either by means of a top-down (strategic initiation) or bottom-up (business user initiation) approach. Whatever the case, a team comprising an analytics expert (domain knowledge), a data architect (data knowledge) and a data scientist (analytics knowledge) specifies the use case. Besides technical requirements, the specification includes a calculation of the business case and agreements to obtain the data, which might involve further interactions with data stewards and data owners. Thereafter, the use case experiences a funnel process, in which the data and analytics board, which includes business sponsors and senior managers, review and eventually prioritize it. Once the use case has been prioritized, the *Development* phase starts.

In the *Development* phase, the architect models the required data, which the data engineer extracts from the source system(s) and loads it in its raw format to the data lake. After onboarding the data on the data lake, a data engineer creates a dedicated sandbox environment to access the dataset. The relevant data steward and data engineer document the newly onboarded data in the data catalog. In an ideal case, the required data are already onboarded on the data lake and only require the latter steps.

In the *Implementation* phase, the data scientist is given access to the sandbox environment. An analytics expert could help the data scientist understand the business side of the analytics use case.

In the *Operation & Maintenance* phase, the data scientist tests different algorithmic approaches' feasibility regarding addressing the analytics use case. This usually involves multiple iterations of the analytics model's building and evaluation, and might require a data engineer to change or onboard more data.

Table 17. Analytics experimentation work system

<table>
<tr><td>Customers</td><td colspan="3">Products/Services</td></tr>
<tr><td>Data scientist</td><td colspan="3">Providing a virtual sandbox environment to develop and test analytics use cases and prove their feasibility (Watson, 2014).</td></tr>
<tr><td colspan="4">Major activities/processes</td></tr>
<tr><td colspan="4">1. Initiation:
• Analytics use case specification by AEX/DAR/DSC/DOW (DCA) and
• analytics use case prioritization by DAB
2. Development:
• Data onboarding by DAR/DEN/DOW (DLA, EL, DCA),
• creation and configuration of sandbox environment by DEN (DLA, VSO), and
• data documentation by DEN/DST (DCA)
3. Implementation:
• Sandbox provision to DSC by DEN (SEN) and
• support in business understanding of use case by AEX
4. Operation & maintenance:
• Analytics model development by DSC (IDE, ICO, SEN, CRE, DCA)</td></tr>
<tr><td>Participants</td><td>Information</td><td colspan="2">Technologies</td></tr>
<tr><td>Data and analytics board (DAB)
Analytics Expert (AEX)
Data architect (DAR)
Data owner (DOW)
Data scientist (DSC)
Data engineer (DEN)
Data steward (DST)</td><td>Domain knowledge
Sample dataset (incl. labels)
Historic data
Structured/Unstructured
Raw format
Reference data
Pre-trained models</td><td colspan="2">Integrated development environment (IDE)
Interactive computing tools (ICO)
Extract and load (EL)
Programming libraries (PLI)
Sandbox environment (SEN)
Virtualization software (VSO)
Code repositories (CRE)
Data catalog (incl. business glossaries and data dictionaries) (DCA)
Data lake (DLA)</td></tr>
</table>

5.4 Analytics Production Work System

The *Analytics production* work system (see Table 18) deploys analytics models and ensures that they generate business value throughout their lifecycles. While an analytics model is usually developed by using historic data, the deployed model requires access to real-time data and might even use these data to optimize itself over time.

In the *Initiation* phase, a system engineer, a data architect, the responsible data scientist, and an analytics expert review the successfully tested analytics model and specify the requirements for the production. Their tasks include clarifying how often an analytics model needs to be retrained (as part of the analytics model lifecycle) and, for instance, how the quality can be monitored.

In the *Development* phase, the analytics model needs to be optimized for production according to the specification. First, a developer, with the responsible data scientist's support, converts the analytics model to a production-ready form. Second, a system engineer designs the application architecture according to the enterprise architecture. Third, a data architect provides the data models and a data engineer implements the extract, transform, and load process accordingly.

The developer and system engineer then test and deploy the analytics model. In the meantime, the responsible analytics expert and data scientist create a plan and material to train business users in using the analytics model. The system engineer, data scientist, and analytics expert document the analytics model.

Table 18. Analytics production work system

<table>
<tr><th>Customers</th><th colspan="2">Products/Services</th></tr>
<tr><td>Business user</td><td colspan="2">Providing an up-to-date analytics model in a business application (Watson, 2019).</td></tr>
<tr><td colspan="3">Major activities/processes</td></tr>
<tr><td colspan="3">1. Initiation:
• Specification of requirements analytics model production by SEN, DAR, DSC, AEX
2. Development:
• Production version of an analytics model by DEV/DSC (CRE, PLI),
• application architecture design by SEN (SMO),
• data models by DAR (DMO),
• ETL by DEN (ETL),
• deployment by DEV/SEN (DTO),
• creation of training material by AEX/DSC, and
• documentation of analytics model by SEN/DSC/AEX (CRE, DCA)
3. Implementation:
• Training of business users in the use of analytics model by AEX
4. Operation & maintenance:
• Monitoring of analytics model quality by DEN (MTO) and
• maintenance of analytics model by DSC (IDE)</td></tr>
<tr><th>Participants</th><th>Information</th><th>Technologies</th></tr>
<tr><td>Analytics Expert (AEX)
Data architect (DAR)
Data scientist (DSC)
Data engineer (DEN)
Data steward (DST)
Developer (DEV)
System engineer (SEN)</td><td>Analytics model
Required data for analytics model
Historic/Real-time data
Pre-defined structure
Data push</td><td>Integrated development environment (IDE)
Data modelling tools (DMO)
Software modelling tools (SMO)
Deployment tools (DTO)
Extract, transform, load (ETL)
Programming libraries (PLI)
Monitoring tools (MTO)
Code repositories (CRE)
Data catalog (incl. business glossaries and data dictionaries) (DCA)
Data lake (DLA)</td></tr>
</table>

In the *Implementation* phase, Business users are trained to use the analytics model. While the use does not necessarily require any data management or knowledge of statistics, skills in change management are needed for successful implementation of analytics applications.

In the *Operation & Maintenance* phase, business users use the analytics model in business applications. The data engineer continuously monitors the analytics model's quality. In case of changes in the underlying data distribution, which might lead to a drop in the analytics model's accuracy, the data scientist needs to newly optimize the model.

5.5 Integration of the Four BI&A Work Systems

While we outline the four BI&A work systems separately, commonalities can be identified across BI&A work systems:

- **Participants:** The analytics expert, data architect, and data engineer are key roles to build BI&A capabilities and are required in all BI&A work systems. The analytics expert is the business domain expert of the *Participants* and has a two-fold role: to identify and specify business requirements for BI&A products, but also to support their implementation into the organization by training business users and documenting products from a business perspective. While the data architect mainly helps identify and model enterprise data in the *Development* phase, the data engineer is needed in the *Development* phase to implement "data pipelines" (some form of ETL process) and in the *Operation & Maintenance* phase to ensure these data pipelines remain available. Data modelling and data engineering expertise could therefore be bundled in a center of excellence. It could also be argued that the opposite holds true with regard to data analysts and data scientists, who need to collaborate with business users and analytics experts for whom a decentralized model seems to make more sense. While analytics experts reside in their respective business functions, they should be coordinated centrally to democratize BI&A knowledge correctly.
- **Technologies/Infrastructures:** The four work systems obviously share many infrastructure requirements. *Reporting* and *Data exploration* are generally enabled by means of a data warehouse, while the *Analytics experimentation* and *Analytics production* work systems leverage a data lake infrastructure. The former two are complemented by BI tools to visualize and analyze data in an interactive way. All four work systems benefit from a data catalog solution.
- **Processes/Activities:** The work systems *Reporting, Data exploration,* and *Analytics experimentation* all require the data and tool requirements to be specified in the initiation phase. This could potentially be bundled with a request management process (or use case funnel) that prioritizes requests and allocates resources centrally. Moreover, the analytics model lifecycle spans two work systems. An analytics model's feasibility is first tested in the *Analytics experimentation* work system and, if this test is successful, it is productized in the *Analytics production* work system. While the separation of the two work systems seems to be reasonable from a capabilities perspective, both work systems require effective alignment to ensure a seamless transition from an analytics model prototype to an analytics model in production.

Since we can identify commonalities in the BI&A work systems, we argue that a company may lose synergies if they manage their existing BI environments and emerging big data infrastructures separately. From our case analysis, we find that managing the four work systems as an integrated "enterprise analytics platform" creates benefits at organizational and infrastructure level, and helps build superior BI&A capabilities.

6 Conclusion and Implications

Value creation from BI&A is a complex process with multiple stages ranging from the initial investments in resources to obtaining actual value. While existing research mainly focusses on value creation mechanisms, our study addresses resource orchestration and capability building for BI&A. From four case studies and intense exchanges with experts in focus group meetings, we identify four BI&A capabilities prevalent in companies and discussed in the literature: *Reporting*, *Data exploration*, *Analytics experimentation*, and *Analytics production*. For each BI&A capability, we identify patterns in the form of a work system with its specific components. The work system framework provides a structured approach to identify tangible, intangible, and human resources, as well as analyze how these resources are orchestrated to create BI&A capabilities. We thereby do not only explain how enterprises build specific BI&A capabilities, but also suggest potential synergies by identifying commonalities across the suggested BI&A work systems. Our research therefore addresses important questions outlined in the research agenda for BDA related to analytics capabilities' creation, i.e. the ability to integrate, disseminate, explore, and analyze big data, by Grover et al. (2018). On a more general level, we showcase how WST can be used to understand resource orchestration and capability building in IS research.

Our study does have limitations. First of all, the study is of qualitative nature and only allows analytical generalization. Quantitative studies are therefore needed to validate our findings. Furthermore, it should be noted that our sample in both the expert group and the case studies comprises large corporations with high levels of specialization. This implies that the findings might not be transferrable to smaller companies.

Our findings allow practitioners to not only understand the essential resources and their interplay, but also to map them to their organizational context. The documentation in the form of work systems equips enterprises with the possibility to analyze their current situation and define an appropriate organizational and infrastructure setup for their analytics initiatives. While we view the BI&A work systems separately, our findings suggest that companies should

manage their existing BI environments in conjunction with their emerging analytics infrastructures to enable synergies between the different work systems. From an academic perspective, our research contributes to understanding resource orchestration and capability building as a prerequisite to value generation with BI&A. In this field, we see promising research opportunities related to all four BI&A work systems, as well as their integration into an "enterprise analytics platform." For instance, the transition from *Analytics experimentation* to *Analytics production* remains a challenge in practice and requires an in-depth analysis.

7 References

Alpar, P., & Schulz, M. (2016). Self-Service Business Intelligence. *Business & Information Systems Engineering*, *58*(2), 151-155.

Alter, S. (2004). A work system view of DSS in its fourth decade. *Decision Support Systems*, *38*(3), 319-327.

Alter, S. (2013). Work System Theory : Overview of Core Concepts, Extensions, and Challenges for the Future. *Journal of the Association for Information Systems*, *14*(2), 72-121.

Baker, J., Jones, D., & Burkman, J. (2009). Using Visual Representations of Data to Enhance Sensemaking in Data Exploration Tasks. *Journal of the Association for Information Systems*, *10*(7), 533-559.

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly*, *11*(3), 369-386.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics : From Big Data to Big Impact. *MIS quarterly*, *36*(4), 1165-1188.

Davenport, T., & Bean, R. (2019). Big Data and AI Executive Survey 2019. *New Vantage Partners LLC*, 1-16.

Davenport, T. H. (2006). Competing on Analytics. *Harvard Business Review*. https://hbr.org/2006/01/competing-on-analytics

Davenport, T. H. (2018). From analytics to artificial intelligence. *Journal of Business Analytics*, *1*(2), 73-80.

Farid, M., Roatis, A., Ilyas, I. F., Hoffmann, H.-F., & Chu, X. (2016). CLAMS : Bringing Quality to Data Lakes. *Proceedings of the 2016 International Conference on Management of Data (SIGMOID '16)*, 2089-2092.

Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics : A Research Framework. *Journal of Management Information Systems*, *35*(2), 388-423.

Heart, T., Ben-Assuli, O., & Shlomo, N. (2018). Using the Work System Theory to Bring Big Data Analytics to the Inpatient Point of Care. *Proceedings of the 26th International Conference on Information Systems (ICIS)*, 1-9.

Jukić, N., Sharma, A., Nestorov, S., & Jukić, B. (2015). Augmenting Data Warehouses with Big Data. *Information Systems Management*, *32*(3), 200-209.

Kettinger, W. J., Zhang, C., & Li, H. (2019). Information Management Capabilities in the Digital Era : The Senior Manager's Perspective. *Proceedings of the 27th International Conference on Information Systems (ICIS)*, 1-15.

Kohli, R., & Grover, V. (2008). Business Value of IT : An Essay on Expanding Research Directions to Keep up with the Times. *Journal of the Association for Information Systems*, *9*(1), 23-39.

Madera, C., & Laurent, A. (2016). The Next Information Architecture Evolution : The Data Lake Wave. *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, 174-180. http://doi.acm.org/10.1145/3012071.3012077

Marjanovic, O. (2016). Improvement of Knowledge-Intensive Business Processes Through Analytics and Knowledge Sharing. *Proceedings of the 24th International Conference on Information Systems (ICIS)*, 1-19.

Melville, N., Kraemer, K., & Gurbaxani, V. (2004). Review : Information Technology and Organizational Performance: An Integrative Model of It Business Value. *MIS Quarterly*, *28*(2), 283-322.

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. N. (2018). Big data analytics capabilities : A systematic literature review and research agenda. *Information Systems and E-Business Management*, *16*(3), 547-578.

Ransbotham, S., Khodabandeh, S., Fehling, R., LaFountain, B., & Kiron, D. (2019). Winning With AI. *MIT Sloan Management Review*, 1-23.

Ransbotham, S., & Kiron, D. (2017). *Analytics as a Source of Business Innovation* (p. 1-16). MIT Sloan Management Review.

Schryen, G. (2013). Revisiting IS business value research : What we already know, what we still need to know, and how we can get there. *European Journal of Information Systems*, *22*(2), 139-169.

Schüritz, R., Brand, E., Satzger, G., & Bischhoffshausen, J. (2017). How to cultivate analytics capabilities within an organization ? - Design and types of analytics competency centers. *Proceedings of the 25th European Conference on Information Systems (ECIS)*, 389-404.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Éds.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)* (p. 2503-2511).

Seddon, P. B., Constantinidis, D., Tamm, T., & Dod, H. (2017). How does business analytics contribute to business value? : How does analytics contribute to business value? *Information Systems Journal*, *27*(3), 237-269.

Shim, J. P., & Guo, C. (2015). Big Data and Analytics : Issues, Solutions, and ROI. *Communications of the Association for Information Systems*, *37*, 797-810.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263-286.

Soh, C., & Markus, M. L. (1995). How IT creates Business Value : A Process Theory Synthesis. *Proceedings of the 16th International Conference on Information Systems (ICIS)*, 29-41.

Sprague, R. H. (1980). A framework for the development of decision support systems. *MIS quarterly*, *4*(4), 1-26.

Trieu, V.-H. (2017). Getting value from Business Intelligence systems : A review and research agenda. *Decision Support Systems*, *93*, 111-124.

Watson, H. (2014). Tutorial : Big Data Analytics: Concepts, Technologies, and Applications. *Communications of the Association for Information Systems*, *34*(1).

Watson, H. J. (2009). Tutorial : Business Intelligence – Past, Present, and Future. *Communications of the Association for Information Systems*, *25*(487-510), 26.

Watson, H. J. (2017). Preparing for the Cognitive Generation of Decision Support. *MIS Quarterly*, *16*(3).

Watson, H. J. (2019). Update Tutorial : Big Data Analytics: Concepts, Technology, and Applications. *Communications of the Association for Information Systems*, 364-379.

Wixom, B., & Ross, J. (2017). How to Monetize Your Data. *MIT Sloan Management Review*, *58*(3).

Yin, R. (2003). Case Study Research : Design and Methods, Third Edition, Applied Social Research Methods Series, Vol 5. Sage Publications, Inc.

Data Ownership Revisited: Clarifying Data Accountabilities in Times of Big Data and Analytics

Martin Fadler and Christine Legner
Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

First version presented at European Conference of Information Systems 2020
Extended version published in Journal for Business Analytics 2021

Abstract: Today, a myriad of data is generated via connected devices and digital applications. In order to benefit from these data, companies have to develop their capabilities related to big data and analytics (BDA). A critical factor that is often cited concerning the 'soft' aspects of BDA is data ownership, i.e. clarifying the fundamental rights and responsibilities for data. Scholars have investigated data ownership from different disciplinary perspectives. In IS research, this resulted in definitions of data ownership for operational systems and data warehouses, where the purpose of data processing is known. In the BDA context, defining accountabilities for data ownership is more challenging, because data are stored in data lakes and used for new, previously unknown purposes. Based on insights from four case studies with extensive experience in BDA, we identify ownership principles and three data ownership types: *data*, *data platform*, and *data product*. We also discuss implications resulting from repurposing of data. By redefining the concept of data ownership, our research answers fundamental questions about how data management changes with BDA, extending existing concepts on data ownership and lays the foundation for future research on data and analytics governance.

Keywords: Data Ownership, Data Governance, Analytics Governance, Big Data and Analytics, Data Lake, Data Products

Table of Contents

List of Tables

1 Introduction

There is no doubt that data are leading to a rising new economy (The Economist, 2017) and are fundamentally changing how business is conducted (Davenport et al., 2012; Wamba et al., 2015). With decreasing computing costs and the myriad of data generated via connected devices and digital applications, enterprises are seeking opportunities to improve existing processes and products as well as to develop new data-driven business models (Wixom & Ross, 2017). This goes along with improving their capabilities to manage big data and analytics (BDA) (Grover et al., 2018). A cornerstone of BDA is data lakes, which store large volumes of data in various formats and enable innovation through data exploration and experimentation (Farid et al., 2016; Madera & Laurent, 2016; Watson, 2017). Since data are nonrival, the business potential scales with data being used for multiple purposes at the same time without losing their value (Jones & Tonetti, 2019). However, this idiosyncrasy and the increasing number of data consumer-provider relationships leads to complexity in data ownership. While there is consensus that data ownership clarifies fundamental rights and responsibilities for data (Hart, 2002), the related debates in practice and research view the concept from different, often contrasting disciplinary perspectives. The legal perspective is reflected in the increasing number of data privacy regulations that governments issue to give individuals more rights and to control businesses' uses of personal data (Labadie & Legner, 2019). Economists emphasize that data ownership affects and potentially harms social welfare (Jones & Tonetti, 2019). In IS literature, data ownership is often cited as a critical factor concerning the 'soft' aspects in the creation and use of enterprise data, specifically BDA. Data ownership is not only important to gain business value from big data (Alexander & Lyytinen, 2017; Comuzzi & Patel, 2016; Grover et al., 2018); it also clarifies fundamental rights and responsibilities that underpin data governance (Loshin, 2001; Winter & Meyer, 2001). Grover et al. (2018) emphasize:*"[…] governance that delineates responsibility and accountability for data, [is a catalyst] for BDA value creation"* (p. 417).

Data ownership has been discussed since electronic data processing began (Maxwell, 1989; Spirig, 1987; Van Alstyne et al., 1995; Wang et al., 1995). The focus of the subsequent debates has been on data ownership for operational systems and data warehouses (Winter & Meyer, 2001), where the purpose of data processing is known. While we can assume that data ownership is still beneficial in today's corporate environment, practitioners emphasize that data lakes require a different approach to data governance (Chessell et al., 2018). Defining accountabilities for data is more challenging for BDA, because data are stored in data lakes and used for new, previously unknown purposes. When data are repurposed, data flow across organizational units and need

to satisfy different data consumer' requirements in terms of data format, granularity, and quality. Such cross-unit data flows require effective coordination, as emphasized by the concept of enterprise-wide information logistics (Dinter, 2013; Winter, 2008). These developments raise the question how we need to reinterpret and apply data ownership concepts so as to cope with emerging challenges in BDA environments.

To address this gap, our objective is to understand how data ownership concepts change in the context of BDA. Thus, we ask:

RQ: *How do enterprises define and adapt data ownership in the big data and analytics context?*

To integrate academic and practitioner perspectives, we performed an extensive literature review and conducted explorative research based on multiple case studies to explore data ownership in the real-world context (Benbasat et al., 1987; Yin, 2003). From our analysis of the literature and of four companies with significant BDA experience, we identify data ownership principles and three data ownership types: *data*, *data platform*, and *data product*. We also demonstrate the implications of data repurposing on data ownership assignment and data dependencies. Our findings extend the prevailing data ownership concept from IS literature by integrating the *data platform* perspective, which serves as the required mediator between data supply (*data*) and data demand (*data product*) in BDA environments. Our insights into ownership contribute to the data and analytics governance literature generally. They particularly address structural aspects of data governance according to Tallon et al. (2013) and help clarify the decision rights in Tiwana et al. (2013)'s *IT Governance Cube*. Based on Grover et al.'s (2018) research framework, our study lays the foundation for *BDA governance to facilitate the value creation process*. Our findings also complement prior research on enterprise-wide information logistics (Dinter, 2013; Winter, 2008), by adding the perspective of data ownership to cross-unit information flows. The three data ownership types support the effective coordination of enterprise-wide information delivery in order to generate synergies and attain overarching goals.

The remainder of this paper is structured as follows: We start by reviewing the research field on data ownership from different disciplinary perspectives and outline the research gap. We then motivate our qualitative research approach and provide an overview of the research process. Third, we present each case in detail. Based on our cross-case analysis, we synthesize our findings into six propositions. We conclude with a summary and discussions of our contributions as well as an outlook on future research.

2 Background

Data ownership is grounded in the general concept of ownership, which is a fundamental mechanism in our society and can relate to different disciplinary lenses, including legal, economics and management. Accordingly, different paradigms can be applied to determine who could or would be entitled to claim ownership of data. In the IS field, data ownership has been studied since the early days of electronic data processing, resulting in data ownership principles for operational systems and data warehouses. With BDA, an increasing variety of data sources are used for new, previously unknown purposes and are stored in data lakes so as to enable data exploration and experimentation. This requires us to revisit the data ownership concept for the BDA context.

2.1 Relevance of Data Ownership from Different Disciplinary Perspectives

Ownership is a fundamental concept that is grounded in our everyday life and in fundamental mechanisms of society (Shleifer, 1998). It denotes the assignment of rights and responsibilities for a property to an individual or an organization: *"Property rights […] are the rights of ownership. In every case, to have a property right in a thing is to have a bundle of rights that defines a form of ownership (Becker 1980,189–190)"* (cited in (Hummel et al., 2020), p.3). These rights can apply to material and immaterial objects alike (ibid). Independent of the underlying object, the concept of ownership links various research disciplines among them law, economics, or management. In each of these disciplines, data ownership is discussed with varying objectives (see Table 19).

In law, data ownership is mostly associated with the privacy of individuals. With personal identifiable information being collected in an ever-increasing volume by large tech companies, this discipline aims at defining the actual owner of this data collection and the extent of control that remains with the data's subjects. This legal perspective is particularly important as companies must be held accountable when it comes to data leakages or alienation of use that can harm data's subjects as happened in the Cambridge Analytica scandal (Confessore, 2018). Although some governments are increasingly introducing privacy regulations to give individuals more rights and to control businesses' uses of personal data (Labadie & Legner, 2019; Tikkinen-Piri et al., 2018), the dominant legal view remains that data cannot be owned (Hummel et al., 2020). Nonetheless, contractual and intellectual property law have to be respected for governing data in different situations (ibid). They put forward that data property rights can be transferred

through license agreement or that data property rights are obtained through mere creation (ibid).

In economics, property rights for data are defined as the ability to control the amount of data collected and to monetize it (Dosis and Sand-Zantman 2019, pp. 3-4). With the recent explosion of data, economists are seeking answers on *"how different property rights for data determine its use in the economy, and thus affect output, privacy, and consumer welfare"* (Jones and Tonetti 2019, p. 2819). Inherent to the economic perspective are data's unique characteristics as nonrival goods. In contrast to most other goods, data thereby are infinitely usable and are the source of increasing returns for companies (Jones & Tonetti, 2019). This characteristic can have negative economic consequences in cases where property rights for data are wrongly distributed. First, firms may not adequately respect the privacy of consumers (ibid). Second, firms may hoard data and limit potential gains of data being broadly used (ibid). Finding the optimal allocation of property rights for data therefore remains an open quest. Interestingly, a recent study by Dosis and Sand-Zantman (2019, p.32) finds that the optimal allocation of rights crucially depends on the value of the data, or equivalently on the relative weight between the market in which the data are generated and the market in which they are used. Notably, there are already initiatives that drive open access to data (e.g. Open data) (Link et al., 2017) and to machine learning models (e.g. Open AI) (Open AI, 2020) which directly stimulates reuse and thus generates value.

In management, ownership rights are an important element of corporate governance that guarantee the mere survival of organizations. Recent studies argue that property rights of a company should be assigned in a way that increases a companies' overall market value (Schulze & Zellweger, 2020). Here, a company is owner of data that it collects or creates, while the companies' property rights holders are undertaking the inherent risk of this venture. Linked to this perspective are also the separation and delegation of different decision rights to manage an organization's inherent complexity and achieve a desirable outcome (Fama & Jensen, 1983; Winkler & Wessel, 2018). In their seminal study Fama and Jensen (1983) view a company *"as a nexus of contracts (written and unwritten)"* between different agents (p. 321). As implication, an effective system for decision control implies, almost by definition, that the control (ratification and monitoring) of decisions is to some extent separate from the management (initiation and implementation) of decisions (ibid, p. 304). Besides the general differentiation of decision rights and their separation, it remains important to understand for what object (material or immaterial) a certain decision is made. This question is further studied in the corresponding sub-disciplines of management research, for instance in the IS discipline.

In IS, early studies investigate how the allocation of data ownership affects system success (Maxwell, 1989; Spirig, 1987; Van Alstyne et al., 1995; Wang et al., 1995). Although the authors use the term *"data ownership"*, they do not interpret *"ownership"* in the same way as the other disciplines mentioned earlier. Data ownership in the context of IS governance are decision control rights rather than property rights (as in the economic or management perspectives). For instance, in their seminal paper, Van Alstyne et al. (1995) distinguishes between *ownership* as the residual right of control (i.e. the right to determine access privileges for others), and *usage rights* as the ability to access, create, standardize, and modify data as well as all intervening privileges (p. 8). Allocating decision control rights on data has a direct effect on system implementations. Several studies confirm that data ownership should always stay with its origin (i.e. where the data are created) to ensure system success (Maxwell, 1989; Spirig, 1987; Van Alstyne et al., 1995; Wang et al., 1995). While this logic sounds intuitive, its practical implementation remains complex, especially in analytical information systems where data flow across organizational units (Dinter, 2013).

Table 19. Disciplinary perspectives on ownership and data ownership

Discipline	Ownership concept	Objectives of data ownership
Law	Enablement and protection of rights with respect to one's property (external) and identity (internal).	Ensure data privacy, while holding firms accountable for fraudulent data use.
Economics	Allocation of ownership rights for economic goods and their effect on market equilibriums as well as welfare.	Distribute property rights for data in the way that increases output and consumer welfare, while protecting individual privacy.
Management	Allocation of property rights to maintain an organization's survival and increase its value (Firm is owner of the data it collects and creates)	Define a firm's accountability and assess its risk undertaking through data collection and monetization.
Management information systems	Allocation of decision rights for IT artifacts to achieve a desired outcome.	Assign decision control rights for data among different organizational entities to increase value generation through data.

2.2 Data Ownership Paradigms – How to Assign the Data Owner?

In the enterprise context, data ownership provides the underpinning principles for data governance to define roles, responsibilities, and processes (Loshin, 2001; Winter & Meyer, 2001). Grover et al. (2018, p. 417) argued that *"without appropriate organizational structures and governance frameworks in place, it is impossible to collect and analyze data across an enterprise and deliver insights to where they are most needed"*. The assignment of certain ownership rights to roles has proven to be beneficial: most importantly, people feel responsible, act in their self-interest, and take care of data. Thus, data ownership has been found to positively impact on data quality and system success (Loshin, 2001; Van Alstyne et al., 1995). While the assignment of ownership rights and responsibilities has clear advantages, it can also lead to conflict concerning data sharing (Hart, 2002).

Generally, the allocation of data ownership is a *"control issue – control of the flow of [data], the cost of [data], and the value of [data]"* (Loshin 2001, p. 28). Since responsibilities can depend on its context of use, Loshin (2001) explored different data ownership paradigms. Although Loshin (2001) followed a fairly pragmatic approach, the suggested paradigms can be linked to different general philosophical ownership approaches outlined by (Hart, 2002). These approaches can help us to understand the underlying rationale for assigning ownership as well as to structure the research field (see Table 20). We classify the paradigms according to the socio-organizational context into three categories: individual, organizational, and shared ownership (everyone). We will now present each category.

2.2.1 Individuals as data owner (data ownership outside of the organization)

Data ownership is increasingly being claimed by individuals as the subjects of data (*subject as owner*). This paradigm reflects *libertarian theory* by Robert Nozick and John Rawls, where ownership must be allocated in ways that do not limit the freedom of others to act autonomously (Hart, 2002). With the Internet, personal data are being collected, used, and even sold in nontransparent ways. Thus, the private ownership paradigm often emerges as a reaction once the data collection has been unveiled, and *individual data ownership* rights are increasingly enforced with data protection policies such as the European Union's General Data Protection Regulation (GDPR). With the emergence of the Internet of Things (IoT), the debate about individual data ownership has gained a new facet, because it remains unclear who owns personal data produced by machines (Janeček, 2018). For instance, the data collected by smart meters enable electricity providers to optimize their network and service offerings, but also unveil

highly sensitive data about private households, which can easily be misused (McKenna et al., 2012).

2.2.2 Organizations as data owner (data ownership inside of the organization)

In the context of organizations (*enterprise as owner*), the data ownership concept is getting more complex as a result of distributed data creation and processing in organizations (Van Alstyne et al., 1995). Here, three reasons for claiming ownership can be distinguished. First, organizations claim ownership owing to monetary factors of funding (*funding organization as owner*) or purchasing/licensing data (*purchaser/licensor as owner*). These paradigms build on *labor theory* by John Locke and assign ownership according to the extent of value added through labor (Hart, 2002). They always involve two parties: the organization that funds the party who creates data, and the organization that purchases or licenses data owned by another party. While in the first case data ownership is transferred to the funding organization without any restrictions, in the second case, data ownership is transferred to the purchasing/licensing party under certain restrictions. Second, an organization may claim ownership by using data. This approach reflects the view of first *occupancy theory* by Immanuel Kant, which assigns ownership to the first who possesses a property or object (Hart, 2002). This is typically the case for consuming parties (*consumer as owner*) that require high confidence in the data and therefore take over accountability. It may also apply to parties who read data from different sources (*reader as owner*) to create or add these to their knowledge base. Third, organizations create business value through data processing and therefore claim ownership. In line with *personality theory* by Georg Wilhelm Friedrich Hegel, it determines ownership by a person's will to invest in an object, which makes him this object's owner (Hart, 2002). Four paradigms can be distinguished depending on the processing type: creating data (*creator as owner*) or formatting data (*packager as owner*) for a certain purpose, compiling information from various data sources (*compiler as owner*), and decoding data (*decoder as owner*).

Table 20. Data ownership paradigms and discourses

The socio-organizational context	The data ownership paradigm (Loshin, 2001)	Example	The related philosophical perspective on ownership (Hart, 2002)
Individual	Subject as owner	A private person accuses a company of selling his or her personal data to a third party	Libertarian theory: Ownership does not limit the freedom of others
Organization	Consumer as owner	A sales team uses customer phone numbers that are essential for its daily operation	First occupancy theory: Ownership by being the first to possess an object
	Reader as owner	A consultancy collects information on industry trends to extend its knowledge base	
	Enterprise as owner	An enterprise creates, processes (adds value), and distributes data about its products	Labor theory: Ownership through value adding, either by own labor or owning labor
	Funding organization as owner	A company pays a research company to collect panel data	
	Purchaser/Licensor as owner	A company buys an address list of potential customers	
	Creator/Generator as owner	A research firm invests in collecting qualitative data for a market study	Personality theory: Ownership through personal will to invest in an object
	Compiler as owner	A business intelligence department builds a central data warehouse	
	Packager as owner	A web agency designs and formats a web page for a customer	
	Decoder as owner	A company synthesizes information from DNA data	
Everyone	Everyone as owner	A crowdsourced collection of geo-information in a public database	Utility theory: Ownership maximizes the benefits for all involved parties

2.2.3 Everyone as data owner

Data ownership often implies that an individual or organization has sole ownership rights. The opposite is the case in the paradigm *everyone as owner*, which is applied when data are intended to be shared with a broad user group. In this case, data ownership is not assigned to any individual or organizational party; instead, everyone can become an owner of certain data, and with the same access rights. This paradigm builds on *utility theory* by Jeremy Bentham and John Stuart Mill, where ownership maximizes the benefits for all involved parties (Hart, 2002). It is often emphasized in discussions related to open data, which is *"data that anyone can access and use"* (Link et al., 2017). Especially when the data are created in a crowdsourced way – as is the case with OpenStreetMap (OpenStreetMap, 2019), for instance – the community is the data owner and everyone shares the same rights to access and use the data, under certain restrictions. Still, open data repositories require data governance, which is often hard to establish when responsibilities are distributed, and accountabilities cannot be assigned to an individual or organizational entity. This is especially the case with public health data, but also with data collected in smart cities, for instance. Thus, while open data hold the potential for great innovation, issues develop around privacy, confidentiality, and control of data (Kostkova et al., 2016).

2.3 Approaches to Data Ownership for Operational and Analytical Systems

Data ownership has been specifically investigated for operational systems (Maxwell, 1989; Spirig, 1987; Wang et al., 1995) and data warehouses (Winter & Meyer, 2001). Operational systems seek to enable business processes with quality data, defined as data that fit its purpose (Wang & Strong, 1996). Enterprises have sought to centralize operational systems to ease maintenance and control for IT departments. This has resulted in a misconception that IT departments are the data owner and must be responsible for data quality (Van Alstyne et al., 1995). Business users create the data while executing business processes, but also need high confidence (quality) in the data they use. Thus, in operational systems, it is recommended that data ownership holds to its original aim of ensuring high data quality (Maxwell, 1989; Spirig, 1987). This implies that the data ownership paradigms *creator as owner* and *consumer as owner* fall together.

While data ownership in operational systems follows the logic of business processes, data warehouses and particularly data marts (in the means of analytics systems) integrate data from multiple business processes (Watson & Wixom, 2007). Data warehouses bring together data from operational systems (*push*). To fulfill a certain information demand (e.g. management

report), data are integrated for this particular use in data marts (*pull*). Thus, data ownership in data warehouses and data marts must be data-centric and depends on the number of data integration layers. In the case of one data warehouse and one data mart layer, two ownership types can be distinguished (Winter & Meyer, 2001). Since data are typically not changed when it is brought into a data warehouse, data ownership on the data warehouse layer stays the same as in operational systems (*data supply*). On the data mart layer, data are typically changed to fulfill a certain information need. Thus, data ownership on this layer is assigned to the party who requests particular information (*data demand*), which is often also the sponsor of such activities.

In the context of analytical information systems, data are used in different organizational units than from which they originate (Dinter, 2013; Winter, 2008). The resulting data supply issues have been discussed from the perspective of information logistics, i.e. *"[…] the planning, control, and implementation of the entirety of cross-unit data flows as well as the storage and provisioning of such data"* (Winter, 2008, p.41). Hereby, data ownership, in the form of governance structures (Dinter, 2013) enables efficient and effective information delivery.

2.4 The Research Gap

Debates about data ownership have multiple facets and, with increasing privacy concerns, they go well beyond the boundaries in which data are created. In the enterprise context, data ownership remains more complex compared to other assets. Still, data ownership is needed to clarify rights and responsibilities to ensure business value with effective data governance (Grover et al., 2018; Otto, 2011; Tallon et al., 2013). The research distinguishes two approaches to data ownership: In operational systems, data ownership is business process-centric, i.e. the creator and the consumer of operational data are often the same. This perspective stands in contrast to analytical systems (e.g. data warehouses), where data ownership is data-centric: the consumer is not the creator, because a data mart integrates data from multiple business processes. IS research on data ownership has focused mostly on operational systems, although even more managerial challenges emerge in the context of analytical systems (Dinter, 2013; Winter, 2008). To the best of our knowledge, only one early study elaborates specifically on data ownership in data warehouses (Winter & Meyer, 2001). A few studies investigate related topics, such as data governance in the context of data warehousing (e.g. Watson et al., 2004) or governance mechanisms for data analytics (Baijens et al., 2020), data quality management (e.g. Weber et al., 2009), and data lifecycle management (Tallon et al., 2013).

BDA as emerging analytical paradigm differs from traditional business intelligence and data warehouse infrastructures, where the structure is predefined and data are cleaned upfront to

deliver high-quality reports and insights (Watson, 2009). BDA introduces larger volumes and a higher variety of data that are stored in data lakes, without a predefined structure and in raw format, to enable data exploration and innovation (Farid et al., 2016; Madera & Laurent, 2016; Watson, 2017). With this paradigm shift, new challenges emerge for enterprises (Grover et al., 2018; Sivarajah et al., 2017): On the one hand, with data repurposing, they need to manage an increasing number of data provider-consumer relationships. Providing data for multiple purposes (Chen et al., 2012) imposes higher requirements on data quality, data integration, and data security (Grover et al., 2018). In fact, data quality remains one of the key challenges to enable business value from BDA (Abbasi et al., 2016; Grover et al., 2018; Wamba et al., 2015). On the other hand, the development and operation of analytics go beyond the mere aggregation and visualization of data. With artificial intelligence (AI) (Watson, 2017), it is harder to keep track of how data are processed. Further, the high dependency of machine learning applications on data may lead to the risk of high technical debt (Sculley et al., 2015). At the same time, the increasing use of AI is fueling debates about ethical questions. For instance, deep learning techniques operate as *'black box'* algorithms whose working mechanisms are somehow hard to understand (Castelvecchi, 2016). This is why analytics can lead to *"[…] discriminatory effects and privacy infringements"* (Custers 2013, p. 3) and why debates have emerged about accountabilities for algorithmic decision-making (Diakopoulos, 2016).

These developments are resulting in new issues and questions relating to data ownership, while showing the relevance of defining accountabilities for data. Besides the consideration of these contemporary requirements in research on accountabilities, a holistic view on data governance, which comprises operational and analytical systems, is currently missing.

3 Methodology

We seek to understand *how enterprises define and adapt data ownership in the BDA context* – a complex phenomenon that requires that one analyze rich information related to the adoption of BDA and the definition of data-related roles in enterprises. This is why we opted for an explorative case study research design, which is well suited for answering *how* questions (Yin, 2003) and studying such contemporary phenomena in their particular context (Benbasat et al., 1987; Yin, 2003). Specifically, we studied multiple case studies so as to ensure our theory's robustness and to draw generalizable conclusions (Benbasat et al., 1987; Yin, 2003).

3.1 Case Selection

We integrated our research activities into a research program on data management that included close interactions with 11 data management experts from seven high-profile European companies over 12 months. In early 2019, we initiated an expert group to investigate data management challenges in the context of BDA and met 14 times between January and November 2019. The participants were data experts responsible for establishing organizational and technological structures to manage BDA. They represent large corporations from different industries with some maturity in levering BDA.

Table 21. Selected cases

Case name	Industry	Size	Key informants (years in the company)	Big data and analytics context
Company A	Fast-moving consumer goods	Revenue: $50B to $100B Employees: ~80 000	Manager: data governance (>10y), Enterprise data architect (1-5y)	Organization: central data and analytics management organization Infrastructure: central big data platform for innovation and industrialization of analytics use cases
Company B	Public transportation and mobility infrastructure	Revenue: $1B to $50B Employees: ~35 000	Leader: Business information management (>10y), Data governance manager (6-10y), Big data platform architect (1-5y)	Organization: central data management organization and central/decentralized data science team Infrastructure: corporate data lake for data exploration/experimentation and the operation of analytics use case
Company C	Manufacturing	Revenue: $1B to $50B Employees: ~90 000	Director: Data architecture and engineering (6-10y), Project manager: Data platform (3y)	Organization: corporate data management organization and central platform team Infrastructure: central data platform to enable digital innovations and scale the operation of data products
Company D	Healthcare and life science	Revenue: $1B to $50B Employees: ~50 000	Leaders: Head of Data Products and Solutions (>10y), Global Enterprise Data Strategy Lead (1-5y)	Organization: federated organization with data and analytics center of excellence and staff in line of business Infrastructure: Multiple data platforms serving specific analytics needs and an enterprise-wide data platform

The discussions in the expert group allowed us to develop an understanding of the current situation and to select four (out of seven) companies for further investigation (see Table 21). Three companies were discarded because their data lake initiative was only in the pilot phase

and they had limited practical experience with data ownership in BDA environments. The selected four companies had already established an enterprise data lake and had practical experience with introducing data and analytics roles, including the data ownership concept. As each case company has a high BDA maturity and belongs to a different industry, the case selection process followed literal replication logic, leading to similar rather than contrasting results (Benbasat et al., 1987; Yin, 2003).

3.2 Data Collection

Our data collection approach aimed at gathering information from multiple sources, including expert interviews and internal documents, to allow for triangulation and ensure construct validity (Yin, 2003). For the expert interviews, we selected key informants that have strategic and operational responsibility to manage BDA and who are aware of the relevance of and issues relating to data ownership. For identifying the experts, we used snowball sampling approach (Naderifar et al., 2017): We were already in contact with at least one key informant for the data lake initiative in the respective company through the expert group that we formed (see above). We requested them to identify further key informants in case our requirements were not met. Thereby, we interviewed at least two experts per company, which were knowledgeable about BDA platforms, roles and accountabilities. At least one expert was working in the company for more than five years to ensure a solid understanding of the company's strategic initiatives and challenges. As starting point, we conducted one initial semi-structured interview of 1-1.5h with the key informants to understand each's technological and organizational structures to manage BDA. For instance, we asked the open-ended questions "What is the architectural structure of your data lake?", "What are your key accountabilities for managing data on the data lake?" and "How do you assign those accountabilities?". These interviews gave us the opportunity to understand the challenges and approaches concerning assigning accountabilities for data in greater depth. In parallel, we collected primary data through internal documents provided by the firms (e.g. BDA platform designs, role models, and organizational structures). These documents informed us not only about their approach to data ownership, but also about the context and related topics, such as technical infrastructure as well as established roles or processes.

3.3 Within- and Cross-Case Analysis

We performed the case analysis in two steps. First, we conducted a within-case analysis (Yin, 2003) to understand the different data ownership types in each enterprise. Here, we used an analysis framework and documented the company-specific data ownership types, their

descriptions, and the organizational assignment in each type based on the interview transcripts and the additional company documents provided. In a subsequent expert group meeting, we discussed and compared each company's data ownership approach. The discussion helped us to understand the similarities and peculiarities of each case. Second, we performed a cross-case analysis (Yin, 2003), comparing the findings of the within-case analysis with one another so as to identify common data ownership types and their responsibilities. Further, we linked each identified type to the corresponding data ownership paradigms suggested by Loshin (2001), which helped us to understand its mechanism in a simplified way. Based on our analysis, we outlined four propositions for data ownership in the BDA context. We discussed our findings in another expert group meeting, which gave us a better understanding of whether the enterprises agreed with our conclusions or if we had missed aspects we had not reflected on. To verify specific aspects with the case companies and to ensure robust findings, we conducted an additional interview with one key informant from each company. At the end, we held another expert group meeting to discuss common challenges resulting from data repurposing and derived two further propositions.

4 Data Ownership in the Four Case Companies

To provide insights into the case setting, we start by presenting the general context, i.e. BDA's role in each enterprise and each's approach to data ownership.

4.1 Company A

Company A is undergoing a digital transformation and is introducing innovative digital products (in addition to its traditional product portfolio), which shifts its core business model from business-to-business to business-to-consumer. Through this change, the company faces an increasing number of data created via sensors embedded in the digital product and in new customer touchpoints (e.g. points of sale or web applications). This data are enabling company A to improve the way it understands and interacts with its customers; but, to lever this data, the company had to enhance its data and analytics capabilities. In a first step, it formed a central group that is responsible for enterprise data and analytics. It also established a data lake as a central big data platform (commercialized Hadoop stack from Cloudera, on-premise and partially in the cloud), which enables data scientists to conduct analytics across the traditional business functions based on internal and external datasets. This platform is primarily used for exploration and experimentation, but also for industrialization of analytics use cases. It has three major components: the data repository for storing and staging data from internal and external sources, data science labs for exploration and experimentation, and data products for industrialization of analytics use cases.

Company A distinguishes three data ownership types (see Table 22): data source owner, platform owner, and data product owner. The data source owner is "*primary decision maker about the data entities under his responsibility and accountable for the overall integrity, data lifecycle and data quality of data created in his ownership".* This role is typically assigned at a director level or even above, to the head of a business function that creates but also consumes data of this domain. In the data platform context, the data source owner "*provides approval for data usage in data product*". Thus, company A ensures compliant access to sensitive data (e.g. identifiable personal information). When data are then used in a data product, the company arranges a service-level agreement with the corresponding owner of the data sources so as to ensure quality on both sides. Thus, the data source owner must "*fulfill service-level agreements for data products*". The platform owner is accountable for the platform infrastructure (technology stack) and is assigned to the head of the digital analytics team. Concerning data, he "*maintains data sanity and business context while data are going through the technology stack".* This includes that

he "*oversees and controls work in data labs". Further, he "is accountable for the availability of data pipelines".* In this sense, he must ensure that business requirements for data products are being fulfilled. The data product owner, as a head of a business function, represents the data use side and "*addresses business need for data driven by analytics use cases".* This makes him "*accountable for output of the technology stack".* Once a data product is developed and ready to use, he "*ensures the business value of a data product over its lifetime".*

Table 22. Data ownership in case company A

Data owner type	Description	Organizational assignment
Data source owner	*"Primary decision-maker about the data entities under his responsibility and accountable for the overall integrity, data lifecycle and data quality of data created in his ownership."* *"Provides approval for data usage in data product."* *"Fulfils service-level agreements for data products."*	Head of a business function: director level or above
Platform owner	*"Maintains the data sanity and business context while data are going through the technology stack."* *"Oversees and controls work in data labs."* *"He is accountable for the availability of data pipelines."*	Head of the digital analytics team
Data product owner	*"Addresses the business need for data driven by analytics use cases."* *"Accountable for the output of the technology stack."* *"He ensures business value of data product over its lifetime."*	Head of a business function: director level or above

4.2 Company B

Case company B is an infrastructure provider. It is undergoing a digital transformation following a corporation-wide program with three main goals: improve interactions with customers, increase internal efficiency, and enhance capacity management. Thus, the company has invested in new digital applications and sensor technologies to collect data from its assets. Further, it provides noncritical data to third parties through open access so as to stimulate innovation from the outside. Advanced and big data analytics are key drivers of company B's digitalization initiative and are strategically relevant to the company. Thus, it established a central big data platform (commercialized Hadoop stack from Cloudera, on-premise) to provide access to data from diverse sources simultaneously for innovation and production. To ensure the reusability of data on the platform, it was decided that data must be actively managed through corresponding organizational roles and structures. A central data management organization was established to ensure data governance. On the analytics side, a central data science team coordinates the activities, while data scientists form part of each business unit. The platform has four major

components: data lake, data labs, data apps, and user homes. The data lake serves as an underlying data storage and processing entity that operates along a staging, an integration, and a business transformation layer. Data labs operate on the data lake and serve the data scientists' need to explore and experiment with data, for instance, a group of data scientists is accessing machine state data in a data lab to develop a predictive maintenance algorithm. The data app represents an operationalized application that uses data from the data lake, for instance, the predictive maintenance application signals service workers in case of required maintenance activity. A user home comprises specific data from the data lake that is private to the user, for instance, a business analyst conducts ad hoc analyses of daily customers.

Company B distinguishes three of data ownership types on the big data platform, according to its components (see Table 23): data owner, owner of the data lab / data app / user home, and owner of the data lake. The data owner is responsible for a data feed in the context of the big data platform and is typically assigned to a business role. Thus, this role is "responsible for data quality, definition, classification, security, compliance and data lifecycle of a data attribute, set of attributes, or dataset". The data definition (e.g. documentation in data catalog) and classification must be done when data are brought to the big data platform. This implies that the data owner "controls reading access to his data through data feed on big data platform and ensures compliant use through the provision of no-join policies under the respect of interests of existing and future data user". These policies must be revisited as new data are continuously brought to the platform. Since not every data feed has a data owner assigned when it is brought to the big data platform, the data user is required to find the data owner. If the data owner cannot be identified, the user must fill this gap and becomes the owner of the requested data. The owner of the data lake is "accountable for the standardization of the overall big data solution architecture". This includes that he "proves the compliance of analytics solutions". Thus, this role is assigned to the role of the big data solution architect, who is also responsible for platform development and provides "information on planned extensions of the data lake". This role's responsibilities go beyond the architecture of the big data platform, since he "ensures that new and valuable data are onboarded to the data lake according to the business need and potential. For this, he searches proactively new data sources, valuates their business potential, and initiates the onboarding process". In this regard, the owner of the data lake serves as a mediator between the data owner and the owner of the data lab / data app / user home. The latter holds the rights to use data either through a data app that is typically assigned to a business role or through a data lab or user home that is typically assigned to technical roles, for instance, a data scientist. This owner also "manages access to data lab, app, or user home and is accountable for any

activity (operational activity or data privacy) on it over its lifetime". He is also obliged to inform the platform owner about whether the environment still generates value or can be removed. A data scientist, as a user of the owner of the data lab, "needs to comply with a conduct of ethics when working with data in a data lab".

Table 23. Data ownership in case company B

Data owner type	Description	Organizational assignment
Data owner	*"Responsible for data quality, definition, classification, security, compliance, and data lifecycle of data attribute, set of attributes, or dataset."* *"Controls reading access to his data through data feed on big data platform and ensures compliant use through the provision of no-join policies under the respect of interests of existing and future data users."*	Business role
Owner of the data lake	*"Accountable for the standardization of the overall big data solution architecture. Proves compliance of analytics solutions."* *"Gives information on planned extensions of the data lake."* *"Ensures that new and valuable data are onboarded to the data lake according to the business need and potential. For this, he proactively searches for new data sources, valuates their business potential, and initiates the onboarding process."*	Big data solution architect
Owner of the data lab / data app / user home	*"Manages access to the data lab, app, or user home, and is accountable for any activity (operational activity or data privacy) on it over its lifetime."* *"Data scientists must comply with conduct of ethics when working with data in a data lab."*	Business role for the data app Technical role for the data lab/user home

4.3 Company C

Case company C has a long tradition in the automotive industry. It has invested heavily in R&D to embed software in its products to collect and process data. With this data, the company is seeking to monitor its products' conditions and to provide value adding services to its customers. Thus, it strongly relies on data as an essential component of its future business. For traditional data domains, it has established a corporate organization for master data management. Owing to new requirements to manage sensor data and to develop analytics, company A has extended this function's scope and has set up new organizational units. A central platform team has been built up and manages a platform with a virtualized and physical data lake (Microsoft Azure Cloud) to enable digital innovations and to scale the operation of data products. Company C has also flattened its organizational hierarchies so as to become more agile. Its data platform has two major components: a data hub and data solutions. The data hub connects to the data sources and encompasses a physical and a virtual storage for various types and formats of data. The data

solution accesses and processes data to develop/deliver a data application for/to a data consumer.

Table 24. Data ownership in case company C

Data owner type	Description	Organizational assignment
Data domain manager	*"Controls and monitors the data management for his domain."* *"Receives requests for data processing and provides data for data usage."* *"Reports errors and suggests improvements."*	Business role: lower management
Infrastructure owner	*"Develops and operates the data platform."* *"Oversees the implementation and availability of data pipelines to onboard data to the data hub and to provision data to data solutions."*	Corporate IT role: Head of the data platform team
Business logic owner	*"Accountable for a data application over its lifetime, which includes compliant implementation, the maintenance of the data application, and support of users."*	Business or/and IT role: lower management

In the context of the data platform, company C distinguishes between three ownership types (see Table 24): *data domain manager*, *infrastructure owner*, and *business logic ownership*. The *data domain manager "controls and monitors the data management for his domain"*. Each data domain comprises a homogenous set of data attributes describing a business object, for instance, a customer or an asset. This domain approach to structuring data ownership is a typical approach in organizations with mature data management practices. Company C's *data domain manager "receives requests for data processing and provides data for data usage"* and is accountable for data content and responsible for maintaining data according to business requirements. This role is assigned to a business role in lower management to ensure the efficient handling of requests, which corresponds to company C's agile management approach. Company C does not yet distinguish between the input and output data of a data application. Thus, the data domain manager is the owner of input data to the platform and output data of data applications as long as they belong to his domain of responsibility. This includes reporting errors and suggesting improvements. The infrastructure owner is accountable for the data platform's development and operation. Thus, he *"oversees the implementation and availability of data pipelines to onboard data to the data hub and provision data to data solutions"*. At company C, this role is assigned to the head of the data platform team, which is part of the corporate IT function. The *business logic owner* is *"accountable for data applications over its lifetime, which includes compliant implementation, the maintenance of data application, and support of users"*. This role can either be assigned to a business or/and an IT role (central/decentral) depending on a data application's importance and complexity.

4.4 Company D

Case company D is a long-lasting player in the healthcare and life-science industries and exist on the market since more than a century. As science and technology are at the core of this company, data and analytics have become major enablers for the company's ability to develop innovative products. Thus, an enterprise-wide data platform has been established that aims at data democratization by capturing, curating, exposing, and understanding data to answer innovative business questions. This enterprise-wide platform comprises a wide array of capabilities among them are advanced analytics, text analytics, data lake, and a data catalog. Data can be onboarded from internal operational (e.g. CRM) and analytical systems (e.g. data warehouse) as well as external data sources. The data catalog is a central element of this platform that helps in coordinating data onboarding workflows (data supply) and simultaneously in finding relevant data (data demand).

Company D distinguishes four data ownership types in the context of the enterprise-wide data platform (see Table 25): data owner, data product owner, business owner, and platform owner. It makes a clear distinction between so-called left-hand operations and right-hand operations on the enterprise-wide platform. The "left-hand operations are basically how you fill your data catalog and how you curate your data and organize it", the "right-hand operations are actually how you use that data for a certain purpose and that purpose is what we call product".

On the left hand, data owners are assigned to organizational entities that are the primary users of a specific dataset. The data owner "has a strong contributory role in governing the data in the means of their purposeful, compliant, ethical use as well as their quality". This role is accountable for delegating corresponding data responsibilities to data stewards and data custodians. Data owners are nominated for a dataset's context of use which defines the Who, What, Why, Where, How, and When respective the terms and conditions of a dataset's use. This context of use can be adapted when a dataset is used for a different use case, for instance. While small deviations of a dataset's context of use (e.g., a dataset is used as it is for creating a report) have no effect of its responsibilities, greater deviations (e.g., a dataset's attributes must be extended) may lead to defining a new context of use and nominating a dedicated data owner. The central data organization acts as intermediary and is responsible for assigning data owners and negotiating these contractual agreements.

On the right hand, "data product owners look after certain domains like sales, supply chain or marketing and they oversee a portfolio or bundle of use cases which might result or can be bundled to a product". Hence, the data product owner manages a portfolio of use cases and

collaborates with data and analytics experts to bring these use cases on the platform. First, data need to be onboarded to the platform. This data acquisition "is driven by the data product / use case, first data stewards or data detectors find the data, then data engineers acquire the data and lead engineers organize the data". The data product owner is "complemented with a business owner, someone who has skin in the game and makes sure that their staff are really using the data product, e.g. in digital sales". So, while data product owners seek to transform use cases into products in the central data organization, business owners ensure that these products are actually used to generate value in the lines of business.

Besides the accountabilities for managing data supply and demand, case company D is at the moment establishing the data platform owner role who "prioritizes all the requirements coming from all areas and own the platform, and basically give the direction how the platform will develop further". This role is also part of the data organization.

Table 25. Data ownership in case company D

Data owner type	Description	Organizational assignment
Data owner	*"Data owner has a strong contributory role in governing the data in the means of their purposeful, compliant, ethical use as well as their quality."* *"Data owner is accountable for the careful delegation of responsibilities for supporting processes, systems and uses."*	Organizational unit that has primary use of a dataset
Data product owner	*"Data product owners [...] oversee a portfolio or bundle of use cases which might result or can be bundled to a product."*	Data organization
Business owner	*"The data product owner is complemented with a business owner, someone who has skin in the game and makes sure that their staff are really using the data product, e.g. in digital sales."*	Line of business: team leader
Data platform owner	*"Data platform owner prioritizes all the requirements coming from all areas and own the platform, and basically give the direction how the platform will develop further."*	Data organization

5 Data Ownership Types and Principles in the Context Of BDA

Through a cross-case analysis, our study unveils that BDA leads to significant changes and extensions to data ownership. In the following, we formulate propositions related to the three data ownership types and on the specific implications of data repurposing on data ownership.

5.1 Data Ownership Types

Proposition 1: In the context of BDA, companies define data ownership at three levels: data source or dataset (data supply), data product (data demand), and data platform.

Our cross-case analysis reveals that three different data ownership types were present in all four enterprises. These ownership types characterize relevant organizational data accountabilities and responsibilities in the context of BDA. They can be linked to the corresponding data ownership paradigm suggested by Loshin (2001) and the related philosophical assumptions (see Table 26).

Proposition 2: The data owner ensures compliant access to and use of data, not only in the source system, but also on the platform and in data products. This addition extends beyond the traditional responsibility of ensuring data quality and requires one to manage more data dependencies.

The *data owner* is first the creator but can also be user of data (sources) in his or her domain of responsibility. This implies the accountability for the definition, the quality and the lifecycle of data and can be associated with the paradigms of *creator as owner* and *consumer as owner*. The *data owner* is a pure business role in all four case companies, but with varying organizational assignment levels. While in company A, this role is assigned on a director level, in company C, it is assigned to a lower management function so as to ensure efficiency in handling data requests. In company D, this role is assigned to any organizational unit which is primary user of a dataset.

The *data owner* is accountable for making data fit its purpose, as outlined by seminal papers (Wang & Strong, 1996). But data owners also play a key role in advancing the digital transformation by increasing the availability of quality data captured by digital technologies (Vial, 2019). Interestingly, we find that BDA also extends the responsibilities of *data owners* to also provide the input data for new data products. First, the *data owner* is expected to address the particular requirements of data products according to service-level agreements – as in company A and D. For instance, in company D such contractual agreements are handled through

a dataset's context of use and a central organization helps in moderating and defining them. Second, the *data owner* ensures compliant access and use of the data on the platform, i.e. manages data requests, approves usage, and provides access. For instance, the *data owner* in company B must continually revisit the no-join policies so as to ensure compliant use, also when the number of data available on the platform increases. This responsibility requires both additional effort and knowledge of potential implications when data are combined with data from other domains. In this regard, the *data owner* controls the decentralized access, which is one of the key data security issues to be solved in BDA environments (Grover et al., 2018), and may even be needed at an intra-organizational level (Günther et al., 2017).

Proposition 3: The data product owner ensures business value of a data product over its lifetime, including use case portfolio management, development, maintenance, and user support. Depending on the data product's complexity, this role may require technical expertise; thus, this may be a shared role between business and IT.

The *data product owner* is accountable for the data product. Notably, the companies differentiated between data products that are yet in their development (typically, a sandbox environment ("data lab") used by an analytics development team to explore and experiment with a dataset) and data products that are already developed and used downstream in productive systems (e.g., a customer churn prediction model used by sales teams). In case companies A and C, the *data product owner* is accountable for the data product over its lifetime, including development, maintenance, and user support. In company D, the *data product owner* manages data products for a portfolio of use cases of varying maturity. For use cases with low maturity (i.e. hypothesis that yet need to be validated), the *data product owner* collaborates with data and analytics experts to acquire all necessary data and turn these use cases into value generating products. Here, the paradigms *decoder as owner* (e.g. a data scientist who decodes a pattern in the data) or *compiler as owner* (e.g. data analysts who aggregate multiple data sources) are more suitable as the *data product owner* involved in the creation of the data product that is then consumed by a user. In case company A, this role mainly ensures that the data product generates a business value over its lifetime. In case company D, the *data product owner* is complemented with the role of a business owner who makes sure that data products are actually used. In this sense, the *data product owner* can also be associated with the *consumer as owner* paradigm.

Table 26. Data ownership types in the context of big data and analytics

Data owner type	Responsibilities	Support in cases	Exemplary statement
Data owner	Accountable for quality and lifecycle of data in his domain of responsibility.	A, B, C, D	*"[…] accountable for the overall integrity, data lifecycle, and data quality of data created in his ownership."* (A)
	Fulfils quality requirements for data in his domain of responsibility for data products.	A, D	*"Fulfils service-level agreements for data products."* (A)
	Ensures compliant access and use of data in his domain of responsibility by handling requests, providing access, and approving usage.	A, B, C, D	*"Controls reading access […] ensures compliant use through the provision of no-join policies […]."* (B)
Data platform owner	Ensures data quality on the platform by managing data pipelines to onboard and provision data.	A, C	*"Oversees the implementation and availability of data pipelines to onboard data to the data hub and to provision data to data solutions."* (C)
	Accountable for onboarding of valuable data according to a business need and potential.	B	*"Ensures that new and valuable data are onboarded to the data lake according to the business need and potential."* (B)
	Responsible for the development and operation of the data platform. Approves compliance of data products according to data platform standards.	B, C, D	*"Data platform owner prioritizes all the requirements coming from all areas and own the platform, and basically give the direction how the platform will develop further."* (D)
Data product owner	Ensures that a data product addresses a business need and generates business value over its lifetime.	A, D	*"He ensures business value of a data product over its lifetime."* (A)
	Accountable for a data product over its lifetime, including use case portfolio management, development, maintenance, and user support.	A, C, D	*"Accountable for a data application over its lifetime, which includes compliant implementation, maintenance of the data application, and support of users."* (C)
	Ensures compliant access and use of data product.	B	*"Manages access to data lab, app, or user home and is accountable for any activity […] on it over its lifetime."* (B)

Proposition 4: In BDA environments, the data platform owner role facilitates data supply (data owners) and data demand (data product owners). This activity ensures the availability of data on the platform for data exploration and experimentation, but also for the operation of data products.

Companies manage BDA with data platforms, storing data from multiple sources and delivering data products for data exploration/experimentation and for direct use. This observation underpins the disruptive nature of BDA to amalgamate technologies to derive knowledge from big data into platforms (Abbasi et al., 2016). All enterprises have the role of a *data platform owner,* which serves as a mediator and facilitates data supply (*data owner*) and data demand (*data product owner*). While there are many *data owners* and *data product owners*, there is usually only

one *data platform owner* assigned to an IT role in an enterprise. Thus, we can link this ownership type to the paradigms *compiler as owner*, since this role brings data from various sources to the platform, and *packager as owner*, since they reformat data for particular uses in data products. In company B, this role has the important (even strategic) function to *"proactively"* search for and bring valuable data (according to a business potential and need) to the platform. This role is also accountable for the development and operation of the platform – as is also the case in company C and D. This also includes controlling whether data products comply with data platform standards. In sum, the *data platform owner* is responsible for the availability of data on the platform, since she or he manages the data pipelines to bring data to the platform and to provide data to data products. Our findings thereby also support Wamba et al.'s (2015, p. 242) study that *"[…] emphasizes not only the support but also the active involvement of senior management for successful implementation of the shared platform to leverage 'big data' capabilities"*.

5.2 Implications of Data Repurposing

With BDA, the analytical paradigm changes from using data in known ways towards finding innovative ways of using data in unknown ways (data repurposing). From the challenges that enterprises encounter when repurposing data, we derive further propositions related to the assignment of data ownership and changes in responsibilities.

Proposition 5: With data repurposing, data's context of use deviates more often from its origin. Thus, new data owners may be assigned if the data creators are not able to cope with the additional data requirements.

The role of the data owner becomes an elementary role in the context of BDA. As data repurposing results in changes of a dataset's context of use, it often results in new data requirements, e.g. a specific data attribute must be collected at a data source to be used in a data product. Thus, in order to manage these deviations, responsibilities are required at the source level for maintaining data requirements, while ensuring compliant access and use. The identification and assignment of data owners must follow a governed process to align data supply and demand effectively. In case company D, the context of use comprises six dimensions which define a datasets functional bounds a data owner looks after: *Who, What, Why, Where*, *How*, and *When*. *Who* defines the qualifications and skills of dataset user, *What* defines the dataset itself and its sensitivity level, *Why* describes its purpose of use, *Where* the location of use and how data are flowing to and from that location, *How* governs the maintenance and use of data, and *When* specifies a dataset's time of use and retention restrictions. When a dataset'

context of use changes, it is either extended or a new context of use is defined and assigned to a new data owner. The latter case will only happen when one or more dimensions need to be adapted in a way that goes beyond the original data owner's area of expertise, for instance.

Propositions 6: With data repurposing, the number of dependencies between datasets and data products are increasing. The data platform owner assumes additional responsibilities for maintaining transparency and contractual agreements between data owners and data product owners.

Data repurposing immediately results in an increasing number of dependencies between datasets and data products. On the one side, these dependencies need to be managed on the source level where data requirements are maintained. On the other side, these dependencies need also to be managed on the platform level where data products consume data. For instance, engineers at Google warn about data dependencies in machine learning applications that can lead to high technical debt (Sculley et al., 2015). Transparency on these data dependencies is needed to ensure traceability of data quality impacts, for instance. Hence, the *data platform owner* acts as intermediary role with additional responsibilities regarding transparency and contractual agreements between data owners and data product owners. In line with the concept of information logistics, the *data platform owner* plays an important role in coordinating enterprise-wide information flows and managing the increasing number of data consumer-provider relationships.

6 Summary and Outlook

6.1 Contribution

Our study contributes to the emerging field of research on data governance, which is considered a critical success factor for BDA (Grover et al., 2018) and for digital transformation in general (Vial, 2019). More specifically, we link data ownership to the general philosophical assumptions (Hart, 2002) and identify data ownership types that help assigning the decision rights for governing the content of IT artifacts according to Tiwana et al. (2013)'s IT Governance Cube. Our findings confirm that data ownership remains a key concept to clarify rights and responsibilities but should be revisited in the BDA context. While BDA environments come with specific challenges, due to the nature of advanced analytics products and the more frequent repurposing of data, some of the established principles for operational systems and data warehouses still hold

true; most importantly, the clear distinction between the owner on the data supply side (*data owner*) and the owner on the data demand side (*data product owner*). Despite these similarities, BDA environments require also a change in responsibilities and additional role of the *data platform owner* to mediate data supply (*data owner*) and data demand (*data products*). We conclude that building BDA environments leads to even more complex data provider-consumer relationships and requires effective coordination of enterprise-wide information flows. Our propositions and the suggested ownership types represent a first step towards studying *BDA governance to facilitate the value creation process,* which is a key theme of Grover et al.'s (2018) research framework.

6.2 Limitations

This study comes not without limitations. Since the four case companies represent large organizations, the findings may not be transferrable to smaller enterprises. Also, case studies only allow for analytical generalization, and we suggest quantitative empirical studies to further validate our findings.

6.3 Implications for Research

While prior research has mostly looked at either data or analytics governance, our findings illustrate how these two worlds are interconnected and inform future research on these topics. Eventually, the three types of data ownership may guide the definition of governance mechanisms for BDA and should be considered as the basis for more comprehensive data governance roles and frameworks. We show how data governance designs must be extended to include analytics-related accountabilities for data products and data platforms. Moreover, the identified interdependencies between data ownership types underline the need for relational governance mechanisms and illustrate the collaboration between data and analytics teams with business and IT departments. Data and data product ownership are accountabilities ideally assigned to business stakeholders which understand best how to create business value. However, the domain expertise must be complemented with knowledge about data and analytics. This augmentation requires the collaboration with data and analytics experts that facilitate the value creation process and foster data literacy enterprise wide. Platform ownership lies with the data and analytics teams, which onboard the data and deliver data products, and the IT teams, which operate and develop the infrastructure.

From the perspective of enterprise-wide information logistics, the assignment of data ownership can be interpreted as coordination mechanism in analytical information systems. By setting clear data ownership frameworks, organizations foster *"the planning, control and implementation of*

cross-unit data flows in order to realize enterprise-wide (or even inter-organizational) synergies" (Winter, 2008, p.47). In correspondence to Winter (2008), the *data owner* represents the unit in which data are generated, the *data product owner* the unit in which data are analytically processed, and the *data platform owner* manages the platform infrastructure which is essential for information logistics success. We envision that organizations will be highly data-driven in the future. As data demands increase, the organization inevitably evolves into a complex network of data producers and data consumers. The assignment of data ownership plays therefore a significant role to coordinate these raising data provider-consumer relationships and requires further research to understand the involved processes in greater depth.

6.4 Implications for Practice

Practitioners may use our findings to define their approach to ownership as well as the related roles and responsibilities. Our findings can help them to increase consistency in role definitions and establish an understanding of data supply and demand in their data governance initiatives. For instance, the three data ownership types can be used to derive further roles, such as data engineers which typically work alongside data platform owners to implement data pipelines and data scientists which collaborate with data product owners to build advanced analytics models. Moreover, the ownership types and governance structures need to be complemented by new data quality management practices as data repurposing more frequently changes the data use contexts. Ideally, companies establish scalable and agile approaches for onboarding data in the right quality to create immediate business value through data exploration and experimentation.

7 References

Abbasi, A., Sarker, S., & Chiang, R. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, *17*(2).

Alexander, D., & Lyytinen, K. (2017, August 10). Organizing Successfully for Big Data to Transform Organizations. *Proceedings of the 23rd American Conference on Information Systems (AMCIS)*. http://aisel.aisnet.org/amcis2017/DataScience/Presentations/30

Baijens, J., Helms, R. W., & Velstra, T. (2020, June 15). Towards a Framework for Data Analytics Governance Mechanisms. *Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference*.

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly*, *11*(3), 369–386.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20. https://doi.org/10.1038/538020a

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188.

Chessell, M., Scheepers, F., Strelchuk, M., Starre, R. van der, Dobrin, S., & Hernandez, D. (2018). *The Journey Continues: From Data Lake to Data-Driven Organization*. Redbooks.

Comuzzi, M., & Patel, A. (2016). How organisations leverage Big Data: A maturity model. *Industrial Management & Data Systems*, *116*(8), 1468–1492. https://doi.org/10.1108/IMDS-12-2015-0495

Confessore, N. (2018, April 4). Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. *The New York Times*. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

Custers, B. (2013). Data Dilemmas in the Information Society: Introduction and Overview. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases* (pp. 3–26). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30487-3_1

Davenport, T. H., Barth, P., & Bean, R. (2012). How 'Big Data' Is Different. *MIT Sloan Management Review*, *54*(1), 5.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62.

Dinter, B. (2013). Success factors for information logistics strategy—An empirical investigation. *Decision Support Systems*, *54*(3), 1207–1218. https://doi.org/10.1016/j.dss.2012.09.001

Dosis, A., & Sand-Zantman, W. (2019). The Ownership of Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3420680

Fama, E. F., & Jensen, M. C. (1983). Separation of Ownership and Control. *The Journal of Law and Economics*, *26*(2), 301–325.

Farid, M., Roatis, A., Ilyas, I. F., Hoffmann, H.-F., & Chu, X. (2016). CLAMS: Bringing Quality to Data Lakes. *Proceedings of the 2016 International Conference on Management of Data (SIGMOID '16)*, 2089–2092.

Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, *35*(2), 388–423.

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, *26*(3), 191–209. https://doi.org/10.1016/j.jsis.2017.07.003

Hart, D. (2002). Ownership as an Issue in Data and Information Sharing: A philosophically based review. *Australasian Journal of Information Systems*, *10*(1). https://doi.org/10.3127/ajis.v10i1.440

Hummel, P., Braun, M., & Dabrock, P. (2020). Own Data? Ethical Reflections on Data Ownership. *Philosophy & Technology*, 1–28. https://doi.org/10.1007/s13347-020-00404-9

Janeček, V. (2018). Ownership of personal data in the Internet of Things. *Computer Law & Security Review*, *34*(5), 1039–1052. https://doi.org/10.1016/j.clsr.2018.04.007

Jones, C., & Tonetti, C. (2019). *Nonrivalry and the Economics of Data* (No. w26260). National Bureau of Economic Research. https://doi.org/10.3386/w26260

Kostkova, P., Brewer, H., de Lusignan, S., Fottrell, E., Goldacre, B., Hart, G., Koczan, P., Knight, P., Marsolier, C., McKendry, R. A., Ross, E., Sasse, A., Sullivan, R., Chaytor, S., Stevenson, O., Velho, R., & Tooke, J. (2016). Who Owns the Data? Open Data for Healthcare. *Frontiers in Public Health*, *4*. https://doi.org/10.3389/fpubh.2016.00007

Labadie, C., & Legner, C. (2019). Understanding Data Protection Regulations from a Data Management Perspective: A Capability-Based Approach to EU-GDPR. *Proceedings of the 14th International Conference on Wirtschaftsinformatik (WI)*, 3. https://aisel.aisnet.org/wi2019/track11/papers/3

Link, G., Lumbard, K., Germonprez, M., Conboy, K., & Feller, J. (2017). Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations. *Communications of the Association for Information Systems*, *41*, 587–610. https://doi.org/10.17705/1CAIS.04125

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

Madera, C., & Laurent, A. (2016). The Next Information Architecture Evolution: The Data Lake Wave. *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, 174–180. http://doi.acm.org/10.1145/3012071.3012077

Maxwell, B. (1989). Beyond "Data Validity": Improving the Quality of HRIS Data. *Personnel*, *66*(4).

McKenna, E., Richardson, I., & Thomson, M. (2012). *Smart meter data: Balancing consumer privacy concerns with legitimate applications*. https://doi.org/10.1016/j.enpol.2011.11.049

Naderifar, M., Goli, H., & Ghaljaie, F. (2017). Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research. *Strides in Development of Medical Education*, *14*(3). https://doi.org/10.5812/sdme.67670

Open AI. (2020). *Open AI*. Open AI. https://openai.com/

OpenStreetMap. (2019). *OpenStreetMap*. OpenStreetMap. https://www.openstreetmap.org/copyright

Otto, B. (2011). Data Governance. *Business & Information Systems Engineering*, *3*(4), 241–244.

Schulze, W. S., & Zellweger, T. M. (2020). Property Rights, Owner-Management, and Value Creation. *Academy of Management Review*. https://www.alexandria.unisg.ch/259344/

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)* (pp. 2503–2511).

Shleifer, A. (1998). State versus Private Ownership. *Journal of Economic Perspectives*, *12*(4), 133–150.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286.

Spirig, J. (1987). Compensation: The up-front issues of payroll and HRIS interface. *Personnel*, *66*(100), 124–129.

Tallon, P. P., Ramirez, R. V., & Short, J. E. (2013). The Information Artifact in IT Governance: Toward a Theory of Information Governance. *Journal of Management Information Systems*, *30*(3), 141–178.

The Economist. (2017). Data is giving rise to a new economy. *The Economist Group Limited*. https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy

Tikkinen-Piri, C., Rohunen, A., & Markkula, J. (2018). EU General Data Protection Regulation: Changes and implications for personal data collecting companies. *Computer Law & Security Review*, *34*(1), 134–153. https://doi.org/10.1016/j.clsr.2017.05.015

Tiwana, A., Konsynski, B., & Venkatraman, N. (2013). Special Issue: Information Technology and Organizational Governance: The IT Governance Cube. *Journal of Management Information Systems*, *30*(3), 7–12.

Van Alstyne, M., Brynjolfsson, E., & Madnick, S. (1995). Why not one big database? Principles for data ownership. *Decision Support Systems*, *15*(4), 267–284.

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, *28*(2), 118–144. https://doi.org/10.1016/j.jsis.2019.01.003

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, *7*(4), 623–640. https://doi.org/10.1109/69.404034

Watson, H. (2009). Business Intelligence: Past, Present and Future. *AMCIS 2009 Proceedings*.

Watson, H. J. (2017). Preparing for the Cognitive Generation of Decision Support. *MIS Quarterly*, *16*(3).

Watson, H. J., Fuller, C., & Ariyachandra, T. (2004). Data warehouse governance: Best practices at Blue Cross and Blue Shield of North Carolina. *Decision Support Systems*, *38*(3), 435–450.

Watson, H. J., & Wixom, B. H. (2007). The Current State of Business Intelligence. *Computer*, *40*(9), 96–99. https://doi.org/10.1109/MC.2007.331

Weber, K., Otto, B., & Österle, H. (2009). One Size Does Not Fit All—A Contingency Approach to Data Governance. *Journal of Data and Information Quality*, *1*(1), 1–27.

Winkler, T. J., & Wessel, M. (2018, December 13). A Primer on Decision Rights in Information Systems: Review and Recommendations. *Proceedings of the 39th International Conference on Information Systems (ICIS)*.

Winter, R. (2008). Enterprise-wide information logistics: Conceptual foundations, technology enablers, and management challenges. *Proceedings of ITI 2008*, 41–50. https://doi.org/10.1109/ITI.2008.4588382

Winter, R., & Meyer, M. (2001). Organization of data warehousing in large service companies—A matrix approach based on data ownership and competence centers. *Journal of Data Warehousing*, *6*(4), 23–29.

Wixom, B., & Ross, J. (2017). How to Monetize Your Data. *MIT Sloan Management Review*, *58*(3).

Yin, R. (2003). *Case Study Research: Design and Methods, Third Edition, Applied Social Research Methods Series, Vol 5*. Sage Publications, Inc.

Data Governance: From Master Data Quality to Data Monetization

Martin Fadler, Christine Legner, and Hippolyte Lefebvre
Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

First version presented at European Conference of Information Systems 2021

Extended version for submission to an IS journal

Abstract: Even as companies increase their investments in big data and analytics resources, they struggle to achieve the returns they had hoped for. To manage and gain control over value creation from data, companies must effectively implement data governance. While literature on Information Technology (IT) governance is extensive and provides a thorough analysis of different governance mechanisms, research on data governance is still scarce and narrowly focused on specific (mostly structural) mechanisms. Based on a multiple case study involving companies with substantial data governance experience, we conduct a thorough analysis of data governance mechanisms. In accordance with prior IT governance literature, we find that companies implement structural, procedural, and relational data governance mechanisms. We observe that data governance designs evolve with respect to different data strategy directions and data scope, resulting in three archetypes: (1) Improve master data quality, (2) Enable enterprise-wide data management, and (3) Coordinate the network to enable data monetization. Our study fundamentally advances the field of IT governance by providing evidence that data is governed independently from IT. Our findings adds a strategic perspective to data governance research by identifying a complete set of data governance mechanisms and describing three typical data governance designs. For practitioners, our research provides insights into the priorities of data governance initiatives and outlines pathways to manage data as a strategic asset.

Keywords: IT Governance Mechanisms, Data Governance, Data Monetization, Data Quality, Master Data Management.

Table of Contents

List of Tables

1 Introduction

Companies recognize the strategic potential that data have but fail to realize it yet. Although they raise their investments in big data and analytics resources to find new ways to monetize their data (Wixom and Ross, 2017), they struggle to achieve the returns they had hoped for (Grover et al., 2018; Shim & Guo, 2015). Grover et al. (2018) argue that "*without appropriate organizational structures and governance frameworks in place, it is impossible to collect and analyze data across an enterprise and deliver insights to where they are most needed*" (p. 417). Thus, companies need to strengthen their data governance practices in order to manage and gain control of the value creation process from data.

Research on IT governance shows that the implementation of an effective governance design promotes strategic alignment and leads to superior organizational performance (Wu et al., 2015). Based on their organizational purpose, IT governance mechanisms are generally classified into structural (define the hierarchical structure and assign responsibilities), procedural (define and structure decision-making processes), and relational (communicate, share knowledge, align, and collaborate) mechanisms (De Haes and Van Grembergen, 2004; Peterson, 2004).

In contrast to the extensive literature on IT governance that provides a thorough analysis of different governance mechanisms, research on data governance is still scarce. Data has often been seen as integral part of IT governance, and only a few studies have investigated specific data governance designs for data warehouses (Rifaie et al., 2009; Watson et al., 2004), or master data and data quality (DQ) management (Khatri and Brown, 2010; Otto, 2011a, 2011b; Weber et al., 2009). These few studies consider the role of data mainly as an enabler of business processes and reporting, focus on operational aspects such as data lifecycle management (Tallon et al., 2013), and suggest (mostly structural) governance mechanisms such as data-related roles (Korhonen et al., 2013; Weber et al., 2009). However, data has evolved into a strategic asset (Legner et al., 2020) and has a much greater impact on overall business performance than in the past. Research on data governance must consequently reflect this changing role of data (Grover et al., 2018). More specifically, it is necessary to understand governance designs – comprising structural, procedural, and relational governance mechanisms – with the same strategic perspective adopted by studies investigating these mechanisms for IT artifacts. Hence, we ask the following research questions:

RQ1: *How do companies design data governance using structural, procedural, and relational mechanisms?*

RQII: *How do companies implement data governance to address the changing role of data?*

We opt for multiple exploratory case studies (Benbasat et al., 1987; Yin, 2003) because they allow studying governance designs in their real-world context (Paré, 2004) and understand their contingencies on a variety of internal and external factors (Sambamurthy and Zmud, 1999). By investigating a diverse set of nine multinational companies in terms of their industry, strategic contexts, data strategy, and experience with data governance, we can identify specific data governance mechanisms and analyze the variations between data governance designs. Through our analysis, we identify a set of structural, procedural, and relational mechanisms which are commonly used by companies to govern data. Moreover, we identify three archetypes that characterize typical ways of governing data and reflect the changing role of data: (1) *Improve master data quality*, (2) *Enable enterprise-wide data management* and (3) *Coordinate the network to enable data monetization*. From an academic perspective, our study fundamentally advances the field of IT governance by providing evidence that data is governed independently from IT and that data governance should therefore be recognized as such. In addition, we provide a broader perspective on data governance that includes structural, procedural, and relational mechanisms, and considers emerging requirements from big data and analytics. For practitioners, we provide insights into the priorities taken by data governance initiatives and the interplay between different governance mechanisms. Thus, our research outlines the pathways to help manage data as a strategic asset.

The remainder of this paper is structured as follows. Firstly, we synthesize the IT governance mechanisms from the literature to show the gap in research. Secondly, we describe our research approach. Thirdly, we present findings from our cross-case analysis and identify typical archetypes for data governance designs. Lastly, we summarize the contributions and discuss the implications of our research.

2 Background

2.1 Governance Foundations

The rich body of knowledge on IT governance can be structured along three dimensions (Tiwana et al., 2013): (1) *What is governed?* (i.e., IT artifacts in the form of hardware and software, but also the content of these artifacts in the form of data and information), (2) *Who is governed?* (i.e., projects, firms or ecosystems), and (3) *How is it governed?* (i.e., decision rights, control or architecture). Most research has been conducted (Gregory et al., 2018; Tiwana et al., 2013) on governing IT artifacts and stakeholders in a firm's environment using control mechanisms (Gregory et al., 2018; Tiwana et al., 2013). In this context, the prevailing understanding sees governance *"as the decision rights and accountability framework deployed through a mix of structural, processual, and relational mechanisms and used to ensure the alignment of IT-related activities with the organization's strategy and objectives"* (Gregory et al., 2018, p. 1227). IT governance mechanisms refer to human IT resources and complement IT activities in delivering value to organizations (Wu et al., 2015). These mechanisms act as moderating factors, in particular, to generate business value from big data and analytics investments (Grover et al., 2018). Hence, companies must effectively implement a set of IT governance mechanisms to enable strategic alignment and increase organizational performance (Wu et al., 2015, p. 511).

Based on their organizational purpose, IT governance mechanisms are generally classified into structural (define the hierarchical structure and assign responsibilities), procedural (define and structure decision-making processes), and relational (communicate, share knowledge, align, and collaborate) mechanisms (De Haes and Van Grembergen, 2004; Peterson, 2004). Structural and procedural governance mechanisms are often tangible and implemented in a top-down manner, relational governance mechanisms are usually intangible and tacit as they are "voluntary" actions and cannot be programmed (Peterson, 2004, p.15).

To date, structural, procedural, and relational governance mechanisms have mostly been investigated for IT artifacts, and data has been considered an integrative part of these artifacts (Kohli and Grover, 2008). More recent studies have argued for viewing the content of these artifacts – that is, data (information) and their analysis – as dedicated objects of governance (Grover et al., 2018). However, research on data/information or analytics governance is still rather scarce.

In the following, we analyze prior literature on IT governance mechanisms and compare the mechanisms found for IT artifacts with those identified for data/information and analytics as the content of these artifacts (see Table 27). We then describe our research gap.

Table 27. Governance mechanisms in prior literature

Related literature	Governance mechanisms		
	Structural	Procedural	Relational
IT ARTIFACTS			
(Sambamurthy and Zmud, 1999)	**X**		
(Weill and Ross, 2004)	**X**		
(Xue et al., 2008)		**X**	
(Sambamurthy and Zmud, 2000)		**X**	
(Huang et al., 2010)			**X**
(Wu et al., 2015)			**X**
(De Haes and Van Grembergen, 2004)	**X**	**X**	**X**
(Peterson, 2004)	**X**	**X**	**X**
DATA / INFORMATION			
(Weber et al., 2009)	**X**		
(Velu et al., 2013)	**X**		
(Weber et al., 2009)		**(X)**	
(Tallon et al., 2013)	**X**	**X**	**X**
(Abraham et al., 2019)	**X**	**X**	**X**
ANALYTICS			
(Baijens et al., 2020)	**X**	**X**	**X**

2.2 Structural Mechanisms

Structural governance mechanisms take *"the shape of formal positions and (integrator) roles, and/or formal groups and (management) team arrangements."* They specify the organization's hierarchy, positions and roles and define their responsibilities for IT-related decision making. Thus, while these mechanisms mostly focus on the IT organization, they also involve business stakeholders.

As part of governance, a company first defines which decisions have to be made. For IT and information, Weill and Ross (2004) define IT principles, IT architecture, IT infrastructure, business application needs, IT investment, and prioritization as decision domains. Data governance literature comes up with different decision areas that are more operational. For data quality management, Khatri and Brown (2010) divide the decision domains into data principles, data quality, metadata, data access, and data lifecycle. Interestingly, business needs, as well as investment and prioritization, were not adapted from IT governance mechanisms for the data context.

Once the decision domains are identified, an enterprise must define who is responsible for decision making. According to the location of the decision authority, Sambamurthy and Zmud (1999) distinguish central, decentral, and federated decision making. Weill and Ross (2004) go one step further and derive typical archetypes for this assignment, for instance, "business or IT monarchy" for central decision making. Accordingly, researchers investigated the effects and organizational benefits of the centralized and decentralized assignment: While centralized IS decision making allows for company-wide control, efficiency and reliability in the utilization of IT assets, it decreases the local units' flexibility, agility, and innovation potency (Gregory et al., 2018; Huang et al., 2010). A complete decentralization of IT decision making has the opposite effect. Velu, Madnick and Van Alstyne (2013) formulate the assignment decision for data management practices as a function of the uncertainty in and similarity between business units. However, a recent study showed that data must be governed in a different way than IT: Business organizations are data creators and consumers; therefore, accountability for data should never be centralized to ensure value creation (Fadler and Legner, 2020).

While research on these relationships is often quite abstract, several scholars have investigated how decision rights are assigned on a more granular level. For instance, Winkler and Wessel (2018) analyze different decision right classes and distinguish between decision right input, control, and management rights. In the context of big data and analytics, Fadler and Legner (2020) explore how companies adapt their fundamental decision control rights with the concept of data ownership. They distinguish three data ownership types (i.e., data owner, data platform owner, and data product owner) and analyze the fundamental implications on them when companies repurpose their data, e.g., finding novel data usage contexts with data science. While the data owner and data product owner represent established accountabilities in data management and business intelligence, the authors identify the additional role of data platform that coordinates the data flow across organizational units and the increasing number of provider-consumer relationships. Concrete roles and responsibilities have been a focus topic of data governance research for more than a decade. For instance, Weber, Otto and Österle (2009) define the typical data roles needed for managing data quality. Besides the strategic roles, such as the executive sponsor or chief data steward, they also include operational roles, such as the business data steward or technical data steward. As the overarching authority, a data quality board *"defines the data governance framework for the whole enterprise and controls its implementation"* (Weber, Otto and Österle 2009, p. 11). With big data and analytics becoming strategic value drivers, however, companies must incorporate additional roles and responsibilities, especially for the analytical use context (Fadler and Legner, 2021; Grover et al.,

2018). For instance, Lee et al. (2014) argument for the need of a Chief Data Officer that fosters alignment with business and IT stakeholders on a strategic level. In addition, typical governance roles that have been formalized for data management, e.g. technical data steward, must also be defined on the analytical side, e.g. analytics product architect (Fadler and Legner, 2021). Generally speaking, the importance of steering and operational committees has been emphasized in IS research (Huang et al., 2010; Karimi et al., 2000). These committees *"alig[n] IT-related decisions and actions with an organization's strategic and operational* priorities" (Huang, Zmud and Price, 2010, p. 289) and are commonly seen as an effective governance mechanism. In the context of data specifically, Weber et al. (2009) illustrate the necessity of a data quality board which defines the data governance framework and controls its implementation. Also Tallon et al. (2013) emphasize the need of shared oversight for information governance policy setting, monitoring, and revision.

2.3 Procedural Mechanisms

Focusing only on structural mechanisms would ignore the activities and processes taking place inside an organization's established structures (Sambamurthy and Zmud, 2000). Formal processes are also needed to strike a balance between centralization and decentralization (Gregory et al., 2018). Consequently, procedural mechanisms complement organizational structures and roles in defining how decisions are made. Procedural or process governance mechanisms are defined as *"the formalization and institutionalization of strategic IT decision making or IT monitoring procedures"* (Peterson, 2004, p. 15) and ensure that the IT policies meet business requirements (De Haes and Van Grembergen, 2004). Peterson (2004) synthesizes three essential IT governance processes from literature: *"(a) the identification and formulation of the business case and/or business rationale for IT decisions; (b) the prioritization, justification, and authorization of IT investment decisions; and (c) the monitoring and evaluation of IT decision implementation and IT performance"* (p. 15). These processes aim to align strategic IT investment decisions with company goals and the (administrative, sequential, reciprocal, or full) integration of business and IT decisions (Peterson, 2004). In existing data and analytics governance models, the procedural mechanisms relate to operational rather than strategic aspects. One of the CIOs who participated in the study conducted by Tallon, Ramirez and Short (2013) argued that *"procedural practices permit a greater understanding of the changing value of information and how this value needs to be matched with the characteristics of different storage systems that will maximize and protect that value"* (p. 163). Although Weber, Otto and Österle (2009) outline strategic tasks for data quality management, they do not go into detail but concentrate on their

structural mechanisms. For analytics, the procedural mechanisms typically comprise methods that guide analytics experts to successfully execute analytics projects, such as analytics process models like CRISP, or follow an agile development framework (Baijens et al., 2020).

2.4 Relational Mechanisms

While structural and procedural mechanisms define which, by whom, and how IS-related decisions should be made, relational mechanisms facilitate communication, coordination, and a shared understanding between business and IT stakeholders (Gregory et al., 2018). Thus, relational governance mechanisms are *"the active participation of, and collaborative relationships among, corporate executives, IT management, and business management"* (Peterson, 2004, p.15). They also include communication and shared learning (Wu, Straub and Liang, 2015, p. 500). Hence, these mechanisms focus on the specific horizontal link between IT and business departments to distinguish between the different mechanisms. IT units must establish their communication channels to disseminate IT governance policies, roles, guidelines, and procedures (Huang et al., 2010; Wu et al., 2015). Having the appropriate communication channels in place helps companies create shared mental models, facilitate collaboration, and enhance alignment. Collaboration and alignment are achieved through direct stakeholder participation, business–IS partnerships, or colocation (De Haes and Van Grembergen, 2004). Prasad, Green and Heales (2012) emphasize that collaborative structures can also be built by using tools (e.g., Wiki) correctly. Another relational mechanism is knowledge sharing, which aims to enable a shared perception. To put this mechanism in place, an IS organization should provide training to educate professionals and establish a shared language (De Haes and Van Grembergen, 2004). Another way of sharing knowledge is via job rotation, whereby an IT professional (e.g., an analytics expert) works in different business departments of the company (Baijens et al., 2020). Tallon, Ramirez and Short (2013) observed that the purpose of relational practices revolves around leading users to re-orient their perception of storage as a cheap and infinite resource and, instead, regard it as a finite and costly resource (p. 165). Overall, data- and analytics-related research has not investigated relational mechanism in detail, although alignment and collaboration on strategic and operational levels have been emphasized as important drivers of value generated by investing in big data and analytics (Grover et al., 2018).

2.5 Research Gap

In our review, we find that IT governance research has a much more thorough understanding of different mechanisms than studies on data governance, which mostly consider structural

governance designs. Moreover, studies on IT governance have a clear strategic scope and comprise business cases, investment decisions, and the monitoring of their implementation, for instance. By contrast, the few studies on data (including analytics) governance have a rather operational focus. Even the comprehensive study by Tallon et al. (2013) concentrates on data governance mechanisms *"that span all the stages of the information life cycle from the point of data creation through data destruction"* (p. 162) and considers them a responsibility of IT organizations. This view has arguably been correct over the past few decades. However, as data has evolved into a strategic asset (Legner et al., 2020) and is the source of innovation and a competitive advantage (Grover et al., 2018), data today has a much greater impact on overall business performance and strategic value creation than it did in the past compared to IT. As a reaction, companies establish the role of Chief Data Officer, for instance, who is part of the management board (Lee et al., 2014). Consequently, data governance research must broaden its perspective to address the emerging requirements and changing role of data. This includes extending the focus from operational to strategic aspects and integrate structural, procedural, and relational governance mechanisms.

3 Methodology

To answer our research questions (i.e., *RQ I: How do companies design data governance using structural, procedural, and relational mechanisms?, RQ II: How do companies implement data governance to address the changing role of data?)*, we follow an exploratory case study approach that allows us to investigate the particular phenomenon in a natural context (Paré, 2004). Case studies are commonly used and recommended for answering how and why questions (Yin, 2003). Multiple case studies are more likely than single case studies to lead to robust theories and generalizable results (Benbasat et al., 1987; Yin, 2003).

3.1 Case Selection

Our study is integrated with a multi-year research program that aims to analyze and develop practices for data and analytics management at large corporations. Thanks to this program, we have trusted relationships with data experts from more than 20 companies and privileged access to information about their data strategies and data governance initiatives. For this study, we used theoretical sampling (Eisenhardt and Graebner, 2007) to select companies that have diverse characteristics in terms of their industry and strategic contexts, as well as the selected data scope and experience with data governance. The sample (see Table 28) comprises nine enterprises that have different levels of maturity in data governance, as illustrated by the number of roles and data domains that they focus on (see Appendix, Table 32). Through the variation in our sample, we can analyze differences and commonalities in data governance designs and extract generalizable patterns (Dubé and Paré, 2003).

Table 28. Case companies

Company	Industry	Revenue/Employees	Main contact
A	Public transportation	$1B–$50B / ~35 000	Product owner data strategy
B	Manufacturing, chemicals	$1B–$50B / ~5 000	Head of Corporate Data Management
C	Packaging, food processing	$1B–$50B / ~25 000	Head of Data Management and BI
D	Manufacturing, automotive	$1B–$50B / ~90 000	Vice-President: Data and Analytics Governance
E	Consumer goods	$50B–$100B / ~350 000	Master Data Lead
F	Manufacturing, automotive	$1B–$50B / ~150 000	Head of Master Data Management
G	Pharmaceutical	$1B–$50B / ~70 000	Global Data Lead-Enterprise Solution
H	Consumer goods, retail	$1B–$50B / ~30 000	Vice-President: Data and Analytics
I	Consumer goods, retail	$100B–$150B / ~450 000	Head of Data Management

3.2 Data Collection

We collected primary data through semi-structured interviews. At each firm, we selected key informants who are part of the central data organization and have a mandate for enterprise-wide data governance at their organization. In addition, we made sure that these key informants have worked for a longer period in the company and know the history of data governance initiatives and the issues and challenges that come with implementing data governance (e.g., the challenges that come with involving business stakeholders or assigning roles and responsibilities). With each key informant, we conducted a semi-structured interview of 1.5 hours between September and October 2020 and asked them about the company's data strategy and their current and target state for data governance (roles, processes, and alignment). We used Microsoft Teams to conduct and record the interviews. We complemented the interviews with an analysis of additional documents that we had gathered during our research activities (e.g., on the company's data strategy or data roles and responsibilities) and publicly available information (i.e., news articles, financial reports and presentations at conferences). After the interview, a write-up comprising key statements and links to company material were sent to the interviewees to confirm the statements' correctness, clarify misunderstandings, and answer open questions. Through a combination of primary and secondary sources, we could triangulate the gathered information and ensure construct validity (Yin, 2003).

3.3 Within and Cross-Case Analysis

We analyzed the data in two steps respective our two research questions:

In the first step, we coded each interview using an analysis framework that comprises the governance mechanisms presented in our literature review to answer our first research question (i.e., *How do companies design data governance using structural, procedural, and relational mechanisms?*). Two researchers were involved in the analysis process and labeled all the statements relating to structural (e.g., organizational forms), procedural (e.g., a list of processes), and relational governance mechanisms (e.g., alignment and collaboration types). During the coding process, we further extended and refined the initial set of constructs to provide a complete view of the applied data governance mechanisms, the scope and strategic context. A synthesis of the within-case analysis – with details on each company – can be found in the Appendix (see Table 32).

In the second step, we conducted a cross-case analysis and analyzed the differences and commonalities in the implementation of structural, procedural, and relational governance

mechanisms to answer our second research question (i.e., *How do companies implement data governance to address the changing role of data?*). Through the cross-case analysis, we identified patterns among the case companies, which we grouped together to define the data governance archetypes.

We discussed and reviewed the governance mechanisms and archetypes that we identified – firstly, in a focus group meeting with all interviewees, and secondly, in a focus group including data governance experts other than the interviewees. In both focus groups, the participants confirmed the archetypes – in other words, they could position and relate their data governance approach to one of the identified data governance designs. In addition, they found the data governance designs very helpful to articulate their organization's strategy and governance requirements.

4 Data Governance Design

In the following sections, we present the comprehensive set of structural, procedural, and relational data governance mechanisms that are used in the case companies and discuss them with regard to related literature. We start by describing data strategy objectives and scope which determine the data governance design.

4.1 Data strategy objectives and scope

The data strategy objectives and scope (see Table 29) particularly influence data governance design, i.e., the implementation of structural, procedural, and relational governance mechanisms. The *Data strategy objectives* reflects the long-term direction of the companies' data initiatives respective *Data quality*, *Data availability / access*, and *Data monetization*. While all companies consider data quality targets in their strategy, most of them also aim to make data more broadly available, and a few also take into account goals for monetizing data in various ways. Depending on the chosen *Data strategy objectives*, the implemented data governance design varies among the case companies.

The *Data strategy scope* indicates the data types and domains the data organization is responsible for and reflects the overall data maturity of the enterprise. Accordingly, companies have either a *Narrow* or *Broad* scope. In our sample, only the case companies B and G have a narrow scope as they focus primarily on master data and less than five data domains. The other case companies have a much broader data scope as they take into account other data types and extend towards more data domains to manage data more holistically and consider also analytical

use cases. Depending on the set *Data strategy scope*, an enterprise implements different governance mechanisms.

Table 29. Data strategy objectives and data scope

Data strategy	Characteristics	Case companies
Objectives	Data quality	A, B, C, D, E, F, G, H, I
	Data availability / access	A, C, D, F, H, I
	Data monetization	A, C, D, H
Scope	Narrow	B, G
	Broad	A, C, D, E, F, H, I

4.2 Data Governance Mechanisms

Furthermore, we identified structural, procedural, and relational data governance mechanisms, which are commonly used by the case companies (see Table 30). The identified governance mechanisms confirm existing research on data governance, but also extend them through specific mechanisms which have not been considered yet.

Structural data governance mechanisms

In accordance with existing literature, companies make a fundamental data governance design decision by choosing between a centralized, decentralized, or federated data organization. Almost all case companies follow a federated data organization as they have established a central data team that works with decentralized roles or teams in business functions. Only case companies B, G, H have a solely central data organization and case company G has in addition independent decentral data teams.

Another essential element of any data governance design is *Steering and oversight mechanisms*. The case companies implement *Dedicated boards* and assign *Data governance* as well as *Analytics governance roles*. While not all case companies have established a *Dedicated board* and *Analytics governance roles*, all case companies have assigned *Data governance roles* on different organizational levels and, when chosen a federated organizational structure, also to business functions. Previous studies consider mostly *Data governance roles*, but the cases show that companies increasingly manage data and analytics in an integrated fashion. Therefore, case companies A, C, D, H define *Analytics governance roles* to steer analytics initiatives enterprise wide.

Procedural data governance mechanisms

Companies setup specific processes to make strategic, governance, and operational decisions in a structured way. To some extent, these processes can be linked to the decision domains considered in previous literature as structural governance mechanisms. In our context, the case companies define processes and assign them to roles according to the chosen organizational structure. From the case analysis, eight processes could be identified:

Concerning *Strategic planning decision-making*, the case companies establish *Planning and control*, *Investment management*, and *Business case identification* processes. While all enterprises in our case set have a structured means to manage investments in data-related aspects, not all of them implemented processes to plan and control the implementation of the data strategy and proactively identify business cases. In prior literature, strategic decision-making has been reflected primarily through the lens of data quality, but align with the processes identified through our case analysis (Korhonen et al., 2013; Weber et al., 2009).

Concerning *Data governance design and control decision making*, all case companies establish processes to make decisions about *Data standards and guidelines* and *Data models and architecture*. While the former process handles the conceptual and organizational aspects, the latter process comprises activities to technically implement the business requirements into the system landscape. These processes have been identified in prior literature. Weber et al. (2009) outline activities for data quality and master data management on an organizational and information systems level in accordance to the two distinct process mechanisms found in our case set.

Concerning *Operational data management decision-making*, all case companies establish processes for *Data monitoring and support* and *Data lifecycle management*. On the one side, they proactively manage the lifecycle steps from data creation towards deletion to create transparency and control data flow across systems. On the other side, they monitor data quality and provide support to ensure correctness of data across the enterprise.

Relational data governance mechanisms

All case companies have implemented relational governance mechanisms to align with key stakeholders and establish a data culture. They foster alignment and collaboration with business an IT functions through *Collocation*, *Boards*, and *Procedures*. While the case companies are mostly collocated with IT, they use *Boards* and *Collocation* to strengthen the alignment and collaboration with business stakeholders. In addition, the case companies establish *Data knowledge sharing and use* mechanisms. Primarily they manage communities to engage with data users and foster their data understanding. For the latter purpose, they also use training

programs to increase the data literacy of the business professionals. Only one case company performs regular communication on data matters. Although Tallon et al. (2013) identify specific relational mechanisms, they have not been intensively reflected upon in data governance research, especially with a strategic focus. Interesting is that, *Community management* seems to be a common practice among the case companies, but data governance studies have not investigated them any further.

Table 30. Identified data governance mechanisms

Data Governance Mechanisms		Characteristics	Case Companies	Related Data Governance Literature
Structural	**Data organization structure**	Decentral	G	(Velu et al., 2013) (Weber et al., 2009) (Khatri and Brown, 2010)
		Central	B, G, H	
		Federated	A, C, D, E, F, I	
	Steering and oversight	Dedicated boards	A, D, E, F, H, I	DQM (Weber et al., 2009)
		Data governance roles	A, B, C, D, E, F, G, H, I	DQM (Weber et al., 2009)
		Analytics governance roles	A, C, D, H	(Baijens et al., 2020)
Procedural	**Strategic planning**	Planning and control	A, C, D, E, F, H, I	DQM (Weber et al., 2009) DQ (Korhonen et al., 2013)
		Investment management	A, B, C, D, E, F, G, H, I	
		Business case identification	A, C, D, E, F, H, I	
	Data governance design and control	Data standards and guidelines	A, B, C, D, E, F, G, H, I	DQM (Weber et al., 2009)
		Data models and architecture	A, B, C, D, E, F, G, H, I	
	Operational data management	Data monitoring and support	A, B, C, D, E, F, G, H, I	
		Data lifecycle management	A, B, C, D, E, F, G, H, I	(Tallon et al., 2013) (Khatri and Brown, 2010)
Relational	**Alignment and collaboration with business**	Collocation	A, C, D, E, F, G, I	
		Boards	A, B, C, D, E, F, H, I	
		Procedures	B, I, C, E, F	
	Alignment and collaboration with IT	Collocation	A, B, C, D, E, F, I	
		Boards	D, H	
		Procedures	C, B, G	
	Data knowledge sharing and use	Regular communication	E	(Tallon et al., 2013)
		Community management	A, B, C, D, F, G, H, I	
		Training programs	C, D, E, G, I	(Tallon et al., 2013)

5 Data Governance Archetypes

Based on the cross-case analysis, we observed that the data governance initiatives of the nine companies in our sample set had varying data strategy objectives and scopes and chose their data governance mechanisms accordingly. Using pattern matching, we identified three data governance archetypes that characterize typical ways of governing data according to the implemented structural, procedural, and relational governance mechanisms: *Improve master data quality*, *Enable enterprise-wide data management*, and *Coordinate the network to enable data monetization* (see Table 31. Data governance archetypesTable 31). In the following, we start with a brief overview of the data governance archetypes and then illustrate each of them based on our empirical insights from the nine cases (for detailed information on each case company, see Table 32 in the Appendix).

5.1 Overview

Companies (here: B and G) belonging to the first governance archetype have a narrow scope and focus on improving data quality for master data in a few data domains, like customers, products and finance. We characterize this archetype as *Improve master data quality*. Companies use this initial structuring to focus on the most relevant data objects and define distinct areas of responsibility. While this approach remains the same for the other data governance archetypes, Archetype I has distinct characteristics: A central data team is granted operational responsibilities for collecting business requirements, setting up data quality measures, monitoring data quality, and supporting projects that involve data quality issues. Hence, the responsibilities are mainly centralized, although the data content is created in business units.

Companies (here: E, F, H, I) belonging to the second data governance archetype have a broader scope and comprise a diverse set of data domains and more data types than just master data. Hence, we describe this archetype as *Enable enterprise-wide data management*. With this extended scope, the central data team has a wider array of responsibilities and starts defining a data strategy. While data quality remains a key central responsibility to ensure that data stays *fit for purpose* (Wang and Strong, 1996), data strategy and data access/availability are added to the central data team's responsibilities. While Archetype I nominates only a few decentralized roles that support data lifecycle activities, Archetype II decentralizes responsibilities for collecting business requirements and maintaining data according to domain-specific standards and guidelines. Therefore, relational mechanisms are more intensively established than in the first archetype. For instance, roles and responsibilities are communicated, and regular meetings

and steering committees foster collaboration and alignment between data and business professionals.

Companies (here: A, C, D) belonging to the third data governance archetype recognize data as a strategic asset and a major driver of their digital transformation. Therefore, we characterize this archetype as *Coordinate the network to enable data monetization*. Building on their extensive experience in data management, these companies put specific emphasis on finding and enabling new ways to monetize data and establish a coordinated network of data roles that are not centrally organized. As data is considered a major value driver, these companies have an integrated view of data and analytics through which they foster synergies and seamlessly manage data quality and usage. The remaining central data team mostly undertakes strategic responsibility and is closely aligned with C-level executives. Hence, companies establish the role of the Chief Data Officer to foster alignment and steer data monetization activities enterprise wide.

5.2 Archetype I: Improve Master Data Quality

Strategic context and data strategy: Companies B and G are representative of the data governance–oriented Archetype I as both put in place data governance mechanisms for master data quality. Company B has been facing numerous data quality issues in its operational processes, primarily in the financial domain (e.g., incorrect invoices). Hence, achieving high financial data quality for reporting and controlling is the company's major driver in its digital initiative, which debuted in 2020. Company G faces operational challenges regarding its supply chain, which is typical for the pharmaceutical industry (Desai and Peer, 2018). High-quality data is a major pillar of Company G's digital transformation journey, which the company embarked on in 2019 to optimize operations, anticipate business risks and enhance information transparency along the supply chain. The value of the data unfolds *"by bringing more information together, harmoniz[ing] data from different locations and us[ing] analytics to support product development"* (Head of Corporate Data Management, Company B).

Table 31. Data governance archetypes

	DATA GOVERNANCE ARCHETYPES		
	Archetype I	**Archetype II**	**Archetype III**
	Improve master data quality	***Enable enterprise-wide data management***	***Coordinate the network to enable data monetization***
	CASE COMPANIES		
	B and **G**	**E**, **F**, **H**, and **I**	**A**, **C**, and **D**
DATA STRATEGY			
Objectives	Improve data quality to enable business processes/reporting	Improve data quality to enable business processes/reporting, broaden data access/availability	Improve data quality, broaden data access/availability, monetize data
Scope	Narrow scope on master and reference data and few data domains	Broad scope on any data type and increasing number of data domains	Broad scope on any data type including analytical data and stable number of data domains
STRUCTURAL MECHANISMS			
Organizational structure	Central/Decentral	Central/Federated	Federated
Steering and oversight	Small data organization with essential data roles	Dedicated boards Large data organization with data roles, including assigned roles to business stakeholders	Dedicated boards Large data organization of data and analytics roles, manages as an extended network
PROCEDURAL MECHANISMS			
Strategic planning	Some uncoordinated data strategy planning activities, investments in data quality improvements and infrastructure	Emerging data strategy planning process, investments in data quality improvements and infrastructure, business case analysis for new data domains	Data strategy planning and control process, pro-active identification, and management of data monetization opportunities
Data governance design and control	Ad hoc creation of standards and data models for master data	Data governance framework and process for data modeling and architecture design	Data and analytics data governance framework, unified data architecture
Operational data management	Data quality monitoring and support, uncoordinated data lifecycle management	Data quality monitoring and support, coordinated data lifecycle management	Data quality and use monitoring and support, and data lifecycle management in business functions
RELATIONAL MECHANISMS			
Alignment and collaboration with business	Mostly through procedures or extended boards	Boards and collocation	Boards and collocation
Alignment and collaboration with IT	Collocation with 1-2 data roles in IT functions	Collocation with an extended array of responsibilities for data-related aspects in IT function	Collocation or even combined with a focus on delivering data and analytics products
Data knowledge sharing and use	Few communities for master data Training compliant access and use	Regular updates Emerging community management Training in data quality methods	Regular updates Community management for data and analytics Training in data literacy

Structural mechanisms: Both companies have formed a small central data team that comprises fewer than 10 full-time equivalent employees (FTEs) and operates with a narrow scope on a few data domains. This is a typical approach for this particular data governance archetype. This central data team includes data stewards who take over responsibility for the master data quality in a data domain. They work on developing the methods, standards, and guidelines to create and maintain master data and improve data quality in their data domain. Company B manages four (material, product, customer/account, vendor/supplier) and Company G two data domains (customer/account, material). For each data domain, the data teams have defined the company's core business objects (master data). In Company B, the central data team extends the scope to include managing reference data (e.g., product colors) as well. Besides the data stewards, a dedicated data architect has been nominated as part of the IT organization to support data modelling purposes. Company G has a similar role: a data integration expert. While there are no accountabilities for data on a strategic level, the accountability for data's content lies within the business where they are created.

Procedural mechanisms: As yet, neither of the two companies have defined a comprehensive data strategy, but data is either formulated and embedded in their overarching digital strategy (Company B) or *"data and analytics were identified as core pillars of the overall digital transformation initiative"* (Global Data Lead-Enterprise Solution, Company G). In both companies, the central data teams are responsible for most of the data management processes in the organization (data quality monitoring, data standards), while the data lifecycle is mostly decentralized (in regions or business functions). Company G does have independent decentralized data teams that help to monitor data quality, maintain data, and support the central data team on projects. Company G also relies on a shared service center that supports data maintenance activities. Investment flows into data quality management and is driven either by the IT budget or by the budgets of business stakeholders. Hence, procedural mechanisms mainly focus on operational aspects and on deciding about the data's lifecycles.

Relational mechanisms: Data teams in both companies closely collaborate with business and IT. Company G characterizes the relationship with IT as a *"service-provider relationship,"* with IT providing solutions for the central team. In Company B, the data architect is collocated with IT, and the central data team participates in biweekly meetings related to IT enterprise architecture to align with the data requirements. Alignment and collaboration with business stakeholders happen through projects or collocation with process stewards (Company G). In Company G, monthly global and regional communication ensures knowledge sharing regarding common practices in using data. Company B facilitates active communities (e.g., material master data

community) and a governance body for projects, which invites subject matter experts to contribute to data management projects.

5.3 Archetype II: Enable Enterprise-Wide Data Management

Strategic context and data strategy: Companies E, F, H, and I represent the data governance–oriented Archetype II. Company E sees digitization as vital to its evolution in a connected world. It considers customers as business partners and makes customer experience a core dimension of its digital transformation. As a manufacturer in the automotive sector, Company F aims to enhance products and processes by becoming data-driven. Company H is a large retailer and active in an industry that faces serious challenges because of digital competition and highly informed customers (Deloitte, 2013). It heavily relies on data to improve customer satisfaction, conversion rates, and customer reach. Owing to its growth through mergers and acquisitions, Company I relies on data for operational excellence and IT system landscape consolidation. Thus, the data architecture is key to establishing data governance.

Structural mechanisms: All four companies have a larger central data team (more than 15 FTEs) and a much broader scope than those in the first data governance archetype. This central data team is responsible for a wider array of data domains. For instance, Company F has nine data domains (HR, market, purchasing, finance/controlling, supply chain, production, quality, development/engineering, business partners), and Company I has six (customer/consumer, vendor/supplier, product/article, material financial, employee). Company H follows a slightly different approach to define its areas of responsibility and has 26 domains (e.g., accounting/controlling, data assets, sellables/services) defined by *"going through all processes and business objects that we know to create a holistic view"* (VP Head of Data and Analytics, Company H). Besides master data, which is well established for all four companies, other data types are gaining momentum. These new data types include metadata, which is of the utmost importance to document data for different user groups, and transactional data, which is essential for analytics use cases. Besides managing data quality, data availability and access are among the major concerns and responsibilities of the central data team. Accordingly, roles other than data steward are required across the data domains. These include dedicated roles for data quality (e.g., for creating metrics and monitoring), data standards and methods, and metadata management. Data management also contributes to analytics projects with the provided data and support for data architecture. Besides the centrally organized roles, the data team aims to decentralize responsibilities for managing the data lifecycle to business departments. This includes assigning accountabilities to business stakeholders in the core data domains, who

proactively formulate their business requirements for data. Decentralized teams are organized either by region (Company I) or by business function (Company F). They include nominated roles such as data editors and data owners (for content, domain, or data definition), which are accountable for the data lifecycle or assigned governance responsibilities (Company H). Company E relies on a wide network of data standard owners (about 200 non-FTEs) who are spread across domains and nominated by the central team.

Procedural mechanisms: As data has greater strategic importance than Archetype I, the procedural mechanisms focus not only on operational aspects but also on strategic ones. Decisions are continuously made to review and update the data strategy, which is closely aligned with the IT strategy. In 2015, Company E released its master data strategy to integrate common elements across multiple flows and functions. To define the requirements and have an impact on business, Company E will soon put forward an integrated strategy that extends the existing master data strategy to further data types and add analytics. Company F released its data enablement strategy in 2020, which focuses on business process optimization (operational excellence) and explores ways to turn business capabilities into data and analytics capabilities. Company H has had a data governance strategy since 2019 and will unveil its enterprise-wide data (and analytics) strategy in 2021. Since 2019, Company I has followed a "Data and Architecture strategy" synchronized with the "IT Strategy and Digitization Strategy" and aims to address *"how the organization can work on enterprise architecture with a greater leverage"* (Head of DM, Company I). All companies regularly monitor data quality through business stewardship and defined metrics (e.g., data quality KPIs at Company F). The budget for data management activities can be shared or is directly financed by the business. For instance, all master data–related activities are financed by the business at Company F as part of the MDM committee. The central data team ensures that domains have their own procedures to manage the data lifecycle. A roadmap of data management activities and a portfolio of data management projects help these central data teams to manage and monitor investments.

Relational mechanisms: Companies communicate regularly about data-related topics and projects through different channels. Company E uses newsletters and forums. As the data team aims to decentralize responsibilities, communication includes not only standards and compliant use but also roles, responsibilities, and methods that help to achieve the desired behavior. Boards and committees design the roadmap, nominate roles, and ensure the alignment of decision-making on data management activities between different stakeholders. They meet four to six times a year. Hence, the central data team aligns and collaborates more actively with business stakeholders. Collaboration with business can also happen through internal consulting

services (Company I) or a network of support functions (Company H). Companies F and I use online collaboration platforms or chatbots to enable knowledge sharing and develop skillsets.

5.4 Archetype III: Coordinate the Network to Enable Data Monetization

Strategic context and data strategy: This data governance archetype is represented by Companies A, C, and D. Company A is undergoing a digital transformation driven by increasing competitive pressure, changing customer needs, and new legal requirements. It aims to leverage data in order to improve customer satisfaction while reducing costs through automation. As a result, a transformation of the workforce is expected to address new skillset requirements and staff turnover in the coming years. By 2030, Company C aims to grow revenues by augmenting business with data and analytics insights and reducing costs through operational excellence. It has established a roadmap for enterprise data management by connecting data foundation, capabilities, and organization with business value as the outcome. Company D is active in the automotive industry, which is facing numerous challenges such as market changes toward e-mobility and automotive driving, customer requirements, and cost pressure (Koch, 2015). Company D has made major investments in implementing structured data management to support the company's business transformation. It has demonstrated results with regard to data excellence, innovation, and business value.

Structural mechanisms: As companies see data as a vital driver for the whole enterprise, the central data team sets priorities on formulating and rolling out the enterprise-wide data strategy by establishing the right set of data governance mechanisms. Companies in this data governance archetype establish the role of Chief Data Officer (or Head of Data and Analytics) to foster alignment and steer data monetization activities on a strategic level and across the firm. Business units are planning their data strategy and detailing standards for their respective areas of responsibility, having roles established on a strategic and operational level. Company C has a very small central data management team (six FTEs) setting priorities and designing data governance. This team also coordinates a wide, decentralized network of 100 business experts through an extended data leadership team of 22 business leaders. This structure is typical for the other companies as well. The decentralized data leadership team at Company A comprises 15 leading data managers. Company D has implemented data governance across 47 data domains and established enterprise-wide and data domain-specific standards, clarified data ownership, and assigned data management responsibilities. Its remaining central data and analytics governance team (10 FTEs) reports directly to the CEO and coordinates a decentralized network

of 40 data domain managers (FTEs) in business functions, divisions, and regions, as well as 200 data coordinators (non-FTEs). A data council for project oversight and alignment focuses on prioritizing projects and data governance implementation concerns, among others.

Procedural mechanisms: Procedural mechanisms are established for managing investments in data, planning, and strategy. They are conducted in centralized (e.g., investment in data platform) and decentralized (e.g., staff supporting analytics projects) ways. Thus, data monetization opportunities are proactively identified, and business cases are formulated accordingly (e.g., by using analytics to predict machine outages). Company A renews its data strategy every four years – the current version dates to 2017 and is currently being renewed with a focus on having better data quality, developing roles and skillsets, establishing decentralized responsibilities, and increasing business data awareness. A dedicated strategy for analytics - separate from but coordinated with the data strategy – will also be unveiled to support the consolidation and decentralization of the analytics processes. Company C has integrated all data-related strategies (Master Data Management, BI, Marketing, Engineering) under the umbrella of an enterprise data strategy updated in 2019. BI governance is managed centrally while coordination is more spread out, following global processes and regions. At Company D, the data management strategy started with a focus on master data in 2016, and its scope was extended to all data areas in 2018, leading to a large, decentralized data management network. For the four companies, procedural mechanisms are established for data and analytics on strategic and operational levels.

Relational mechanisms: In this archetype, coordinating an increasing number of data communities and experts becomes a key concern. Alignment and collaboration occur on both an operational level (through communities) and a strategic level (through boards). Communication and knowledge sharing happen through data communities, which comprise key data users and are actively coordinated as virtual networks. For strategic alignment and collaboration, companies establish data steering committees in which key business stakeholders regularly assess and review the roll-out of the data strategy. Beyond formulating the company's vision related to data, quantified goals, and required operations, Company C's strategy encompasses topics related to enterprise culture transformation (e.g., training) and organizing (e.g., teams, principles). Establishing data teams in business is highlighted as a key milestone for the development of capabilities such as data literacy and data democratization. Company A also ensures alignment and collaboration through boards (e.g., data management board) and communities (e.g., AI network group, shared learning group for similar jobs). Company D is building its next-generation enterprise architecture, which will include alignment beyond IT

collocation. All three companies ensure alignment and collaboration thanks to regular high-level DM and D&A board meetings.

6 Conclusion and Implications

Our study provides fundamental insights how companies adapt their data governance designs to address the emerging strategic role that data plays in today's organizations. We find that modern data governance designs, in line with the general governance literature, also comprise structural, procedural, and relational mechanisms, which were only partially addressed in prior research. Our findings thereby extend and consolidate the few studies on data governance designs that have focussed on data warehouses (Rifaie et al., 2009; Watson et al., 2004), or master data and data quality (DQ) management (Khatri and Brown, 2010; Otto, 2011a, 2011b; Weber et al., 2009). Data governance extends beyond structural governance mechanisms and the mere definition of data-related roles and responsibilities; especially, relational, and procedural governance mechanisms are found to be important to decentralize data responsibilities and coordinate the increasing network of data professionals and data users in large organizations. Also relational mechanisms are of particular importance to foster collaboration with business and IT stakeholders.

The three archetypes illustrate how data governance design evolve beyond the data quality and operational aspects shown in previous studies (e.g., Otto, 2011; Tallon, Ramirez and Short, 2013). While the first archetype is arguably representative for this initial state, the other two archetypes show how companies use data governance to manage data as a strategic asset (Legner et al., 2020) and leverage data's monetization opportunities (Wixom and Ross, 2017). Thus, data governance designs consider also the analytical use contexts, a requirement other researchers have called for to moderate value creation from big data and analytics (Grover et al., 2018).

Our findings have several implications for research:
Firstly, we find that data monetization has an impact on data governance design and that the narrow scope of existing data governance research has to be broadened. With data monetization, companies proactively search for and monitor their data use cases, which has implications on the strategic decision-making processes and procedural governance mechanisms. They also need to coordinate among a broad network of professionals who monetize data in various ways, which emphasizes relational governance mechanisms.

Secondly, our study provides evidence that data is governed independently from IT and that data governance should therefore be recognized as such. This finding goes somewhat counter to

earlier studies, which advocate the placement of IT and data/information governance under the same structure as the "preferable" option (Tallon et al., 2013, p.169). In contrast to the prevailing view that data is an integral responsibility of IT organizations, our study demonstrates the importance of data governance as separate instrument to sustain a strategic competitive advantage. However, the collaboration between both organizations remains essential in all case companies, albeit the IT organization is more seen as a service-provider. Moreover, the network perspective on managing data throughout the organization aligns with ongoing debates on viewing the IT organization as a pervasive entity rather than a central and separate unit (Peppard, 2018).

For practitioners, our study provides insights into data governance initiatives in multinational corporations and identifies data governance mechanisms that can guide them to manage data as a strategic asset. As an implication, practitioners should not only focus on structural mechanisms, but concretize these roles by establishing data-related processes (procedural governance) and improve collaboration on data-related topics between business, IT and data and analytics groups (relational governance).

Our study comes not without limitations. Firstly, our sample includes only large, multinational corporations that have complex organizational structures and are characterized by a high degree of specialization and division of labor. They also require more resources for alignment and collaboration. Therefore, our findings might not be applicable to smaller organizations. Secondly, we solely focus on understanding data governance mechanisms and comparing them between companies. We did not analyze the interplay between corporate, IT, and data governance, which presents an interesting avenue for future research.

These developments call for further research with a strategic perspective on data governance designs, while structural, relational, and procedural mechanisms supporting data monetization need to be investigated in more depth.

7 References

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance : A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, *49*, 424-438.

Baijens, J., Helms, R. W., & Velstra, T. (2020, juin 15). Towards a Framework for Data Analytics Governance Mechanisms. *Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference*.

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly*, *11*(3), 369-386.

De Haes, S., & Van Grembergen, W. (2004). IT Governance and its Mechanisms. *Information Systems Control Journal*, *1*, 27-33.

Deloitte. (2013). *Analytics in retail : Going to market with a smart approach*. https://www2.deloitte.com/ch/fr/pages/consumer-business/articles/analytics-in-retail.html

Desai, S. S., & Peer, B. (2018). Big Pharma; Big Data—Big Deal ? Yes; Really! Infosys.

Dubé, L., & Paré, G. (2003). Rigor in Information Systems Positivist Case Research : Current Practices, Trends, and Recommendations. *MIS Quarterly*, *27*(4), 597-636.

Eisenhardt, K., & Graebner, M. (2007). Theory Building from Cases : Opportunities and Challenges. *The Academy of Management Journal*, *50*(1), 25-32.

Fadler, M., & Legner, C. (2021). Toward big data and analytics governance : Redefining structural governance mechanisms. *Proceedings of the 54th Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. /paper/Toward-big-data-and-analytics-governance%3A-Fadler-Legner/e44cbe70723251cd3ba5f17e47c0b3a596214e23

Fadler, M., & Legner, C. (2020). Who Owns Data in the Enterprise ? Rethinking Data Ownership in Times of Big Data and Analytics. *Proceedings of the 28th European Conference on Information Systems*, 207.

Gregory, R. W., Kaganer, E., Henfridsson, O., & Ruch, T. J. (2018). IT Consumerization And The Transformation Of It Governance. *MIS Quarterly*, *42*(4), 1225-1253.

Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics : A Research Framework. *Journal of Management Information Systems*, *35*(2), 388-423.

Henderson, J. C., & Venkatraman, N. (1993). Strategic Alignment : Leveraging Information Technology for Transforming Organizations. *IBM Systems Journal*, *32*(1), 4-16.

Huang, R., Zmud, R. W., & Price, R. L. (2010). Influencing the effectiveness of IT governance practices through steering committees and communication policies. *European Journal of Information Systems*, *19*(3), 288-302.

Karimi, J., Bhattacherjee, A., Gupta, Y. P., & Somers, T. M. (2000). The Effects of MIS Steering Committees on Information Technology Management Sophistication. *Journal of Management Information Systems*, *17*(2), 207-230.

Khatri, V., & Brown, C. V. (2010). Designing Data Governance. *Communication of the ACM*, *53*(1), 148-152.

Koch, M. (2015). Big data and analytics in the automotive industry. Automotive analytics thought piece. Deloitte.

Kohli, R., & Grover, V. (2008). Business Value of IT : An Essay on Expanding Research Directions to Keep up with the Times. *Journal of the Association for Information Systems*, *9*(1), 23-39.

Korhonen, J. J., Melleri, I., Hiekkanen, K., & Helenius, M. (2013). Designing Data Governance Structure : An Organizational Perspective. *Journal on Computing*, *2*(4), 11-17.

Lee, Y., Madnick, S., Wang, R., Wang, F., & Zhang, H. (2014). A Cubic Framework for the Chief Data Officer (CDO) : Succeeding in a World of Big Data. *MIS Quarterly Executive*, *13*(1).

Legner, C., Pentek, T., & Otto, B. (2020). Accumulating Design Knowledge with Reference Models : Insights from 12 Years of Research on Data Management. *Journal of the Association for Information Systems*, *21*(3).

Otto, B. (2011a). Organizing Data Governance : Findings from the Telecommunications Industry and Consequences for Large Service Providers. *Communications of the Association for Information Systems, 29*.

Otto, B. (2011b, juin 9). A morphology of the organisation of data governance. *Proceedings of the 19th European Conference on Information Systems (ECIS)*.

Paré, G. (2004). Investigating Information Systems with Positivist Case Research. *Communications of the Association for Information Systems*, *13*(1), Paper 18.

Peppard, J. (2018). Rethinking the concept of the IS organization. *Information Systems Journal*, *28*(1), 76-103.

Peterson, R. (2004). Crafting Information Technology Governance. *Information Systems Management*, *21*(4), 7-22.

Prasad, A., Green, P., & Heales, J. (2012). On IT governance structures and their effectiveness in collaborative organizational structures. *International Journal of Accounting Information Systems*, *13*(3), 199-220.

Rifaie, M., Alhajj, R., & Ridley, M. (2009). Data governance strategy : A key issue in building Enterprise Data Warehouse. *Proceedings of the 11th International Conference on Information Integration and Web-Based Applications & Services (IiWAS)*, 587-591.

Sambamurthy, V., & Zmud, R. W. (1999). Arrangement for Information Technology Governance : A Theory of Multiple Contingencies. *MIS Quarterly*, *23*(2), 261-290.

Sambamurthy, V., & Zmud, R. W. (2000). Research Commentary : The Organizing Logic for an Enterprise's IT Activities in the Digital Era—A Prognosis of Practice and a Call for Research. *Information Systems Research*, *11*(2), 105-114.

Shim, J. P., & Guo, C. (2015). Big Data and Analytics : Issues, Solutions, and ROI. *Communications of the Association for Information Systems*, *37*, 797-810.

Tallon, P. P., Ramirez, R. V., & Short, J. E. (2013). The Information Artifact in IT Governance : Toward a Theory of Information Governance. *Journal of Management Information Systems*, *30*(3), 141-178.

Tiwana, A., Konsynski, B., & Venkatraman, N. (2013). Special Issue : Information Technology and Organizational Governance: The IT Governance Cube. *Journal of Management Information Systems*, *30*(3), 7-12.

Velu, C. K., Madnick, S. E., & Van Alstyne, M. W. (2013). Centralizing Data Management with Considerations of Uncertainty and Information-Based Flexibility. *Journal of Management Information Systems*, *30*(3), 179-212.

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy : What Data Quality Means to Data Consumers. *J. Manage. Inf. Syst.*, *12*(4), 5-33.

Watson, H. J., Fuller, C., & Ariyachandra, T. (2004). Data warehouse governance : Best practices at Blue Cross and Blue Shield of North Carolina. *Decision Support Systems*, *38*(3), 435-450.

Weber, K., Otto, B., & Österle, H. (2009). One Size Does Not Fit All-A Contingency Approach to Data Governance. *J. Data and Information Quality*, *1*(1), 1-27.

Weill, P., & Ross, J. W. (2004). IT Governance : How Top Performers Manage IT Decision Rights for Superior Results. Harvard Business Press.

Winkler, T. J., & Wessel, M. (2018, décembre 13). A Primer on Decision Rights in Information Systems : Review and Recommendations. *Proceedings of the 39th International Conference on Information Systems (ICIS)*.

Wixom, B., & Ross, J. (2017). How to Monetize Your Data. *MIT Sloan Management Review*, *58*(3).

Wu, S. P.-J., Straub, D., Liang, T.-P., Temple University, Korea University Business School, Liang, T.-P., & National Chengchi University. (2015). How Information Technology Governance Mechanisms and Strategic Alignment Influence Organizational Performance : Insights from a Matched Survey of Business and IT Managers. *MIS Quarterly*, *39*(2), 497-518.

Xue, Y., Liang, H., & Boulton, W. R. (2008). Information Technology Governance in Information Technology Investment Decision Processes : The Impact of Investment Characteristics, External Environment, and Internal Context. *MIS Quarterly*, *32*(1), 67-96.

Yin, R. K. (2003). *Case Study Research—Design and Methods* (3e éd.). Sage Publications.

8 Appendix

Table 32. Detailed description of case companies

Company	Data strategy objectives	Data strategy scope	Data governance
A	DM strategy focused on data governance in place since 2017. A new version planned for 2021 will focus on data quality, roles, and decentralized responsibilities. Analytics strategy for 2021 focuses on further decentralizing analytics initiatives. D&A skillset development is also a core part of the company's digital transformation.	Five data domains with varying maturity: Assets, Business Partners, Production Plans, Product, and Material. Complex datasets spanning business processes/divisions assigned to data managers. Focus on master and transactional data. SAP transformation: R/3 to S/4HANA.	Central DM is a support function with four FTEs and three non-FTEs and is responsible for strategy, methods, and governance. Decentralized teams have 17 FTEs and are organized by "clusters" in business units with data owners. The central analytics team is part of IT with 40 FTEs. Analytics is also decentralized in the IT of business units (100+ FTEs).
B	DM strategy drafted in 2020 with a focus on data foundation and aligned with the group digital strategy. To be spread over all areas of the organization. No analytics strategy but currently exploring how to implement it (independent of the data management strategy).	Five data domains: Product, Customer/Account, Material, and Vendor/Supplier. Historically focused on master and reference data, now also external data. SAP MDG-S implemented for 150 users globally. SAP MDG being implemented.	Central DM is a support function with five FTEs (Head of DM, three data stewards, data architect in IT). It defines methods and guidelines, data models, oversees DQ initiatives, and supports business/IT projects with data know-how. Eight decentralized data experts in business functions (non-FTEs).
C	The company's 2030 strategy will drive D&A initiatives with the goal of monetizing data. The firm's strategic program integrates all data-related strategies since 2019: MDM, BI, Marketing, and Engineering. The first MDM strategy dates to 2005 and the BI strategy to 2009. Developing a corporate data culture is core to the data strategy.	Six data domains: People, Customer, Supplier, Finance, Products/Material, Brand/Category. Self-service exists in BI and AI, with SAP BW and Alteryx. Currently engaging SAP transformation from R/3 to S/4HANA. Strategic program data scope: master, transactional, purchase, machine.	Central data governance team (six FTEs) with decentralized leadership (22 non-FTEs) and business experts (100+ non-FTEs). Two central services for MDM and material data maintenance (total of 32 FTEs). Central BI team operates in IT. BI coordinators and the network of BI experts are decentralized in regions and by process. Domains are assigned ownership, standards, and a model.
D	DM strategy since 2018 and MDM since 2016. DM and analytics will be integrated into the IT and digitalization strategy in 2021. The current analytics strategy is focused more on IT.	Forty-seven data domains with all data types, either established or emerging. Data domains are structured by data objects. D&A is spread across business functions, divisions, and regions.	Central D&A governance agile team (10 FTEs), no role model. Decentralized D&A in domains (FTEs: 40 data domain managers, eight KPI managers, 15 advanced analytics managers; non-FTEs: 200 data coordinators)
E	MDM strategy revised in 2015 to expand the scope of the data. A separate analytics strategy is currently being drafted and will be coordinated with MDM strategy to make a bigger impact.	Six data domains: Customer, Vendor, Product, Material, Financial, and Employee. Master data is well established, and other data types (e.g., internal) are emerging.	Central data governance and methods (15 FTEs). Central analytics in IT without a role model. A network of data standard owners in business functions (200 non-FTEs). Seven shared services for master data operations (100 FTEs).

F	The first draft of "Data enablement strategy" presented to management in March 2020. It will focus on operational excellence and digital transformation: creating data capabilities and analytics capabilities from business capabilities/use cases.	Nine data domains: HR, Market, Finance, Quality, Purchasing, Supply Chain, Development, Production, Business Partners. Group data classes in domains with high governance. SAP family tree for MD domains.	The central team in IT called "data and insights analytics" has four pillars: MDM (17 FTEs), classical BI, finance reporting and advanced analytics. Decentralized data stewardship in domains for DQ and demand. Decentralized reporting in other IT departments.
G	Data strategy is not defined, but data and analytics are separate pillars of the overall digital transformation initiative to be launched in 2021 and are addressed as two separate enablers.	Two data domains: Material and Account. Governance is established only over Material master data. Secured sponsorship from a VP to extend the scope.	Central data team (10 FTEs) with data support and maintenance decentralized in the regions (55 FTEs, including a special team for DQ). Central analytics team (six FTEs) attached to supply chain.
H	Data governance strategy since 2019. BI strategy since 2015. Integrated enterprise-wide D&A strategy in progress, with a release planned for 2021. A data governance framework is currently being rolled out.	Twenty-six data domains and 100 sub-data domains defined by business objects (and functions). All the data-related terms have a glossary. Master, transactional and reference data are established.	Central D&A team of more than 20 FTEs (three for governance) reporting to controlling, while data science reports to strategy. The decentralized data organization in business functions has 15 data stewards (equivalent three FTEs).
I	Data scope extended from MDM to DM through the “Enterprise Architecture and data strategy” (released in 2020) is synchronized with IT strategy and is an enabler of the enterprise-wide digitization strategy.	Six data domains: Article, Vendor Customer, Material, Financial, and Employee. Master data are well established. Transactional, behavioral, and classic analytical data are not fully covered by DM.	Central data management organization (60+ FTEs) working mainly on master data. Decentralized data organization at the branches and also by retail countries with country managers (30 FTEs). Shared services for article master data.

Machine Learning Techniques for Enterprise Data Management: A Taxonomic Approach

Martin Fadler, Christine Legner, and Valérianne Walter

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

First version presented at Pre-ICIS SIGDSA Conference 2019

Extended version for submission to an IS journal

Abstract: Data quality is considered a significant factor for organizational success and a major challenge when adopting Big Data and analytics. Actually, data are often erroneous and comprise outliers, missing values, and duplicates, or violate integrity constraints. Traditional enterprise data management (EDM) practices rely on rule-based approaches, which require the involvement of domain experts, even though these practices do not scale well with increasing data demands. Recent studies confirmed the significant potential of machine learning (ML) techniques to learn from data, despite the fact that these techniques have a strong technical focus and only address isolated problems. Against this backdrop, our study sheds light on how ML techniques can advance EDM. Based on an analysis of 60 ML cases, it contributes in two ways: it proposes a taxonomy that links ML applications to concepts drawn from EDM and data curation; and it identifies nine archetypes that provide an overview of typical application areas of ML in EDM. We find that ML techniques induce a shift from manual data maintenance in a reactive mode to data creation in a proactive mode. Our analysis also reveals that some archetypes build on the rich body of research emerging from the database community.

Keywords: Enterprise Data Management, Machine Learning, Data Curation, Data Quality, Taxonomy

Table of Contents

List of Figures

List of Tables

1 Introduction

Data quality (DQ) poses a major challenge to the creation of value from Big Data and analytics (Baesens et al., 2016; Grover et al., 2018). In fact, data are often erroneous and comprise outliers, missing values, and duplicates, or violate integrity constraints. These DQ issues have a significant impact on organizational success (Otto & Österle, 2015) and are estimated to cost an enterprise between 15–25% of its annual revenue (Redman, 2017). Hence, it is not surprising that DQ is a major concern when implementing artificial intelligence (Pyle and José, 2015), business analytics, or self-service business intelligence (BARC, 2021). Enterprise data management (EDM) aims at handling DQ issues in a structured way and making data fit for use by data consumers (Legner et al., 2020; Otto, 2011; Wang, 1998). Common EDM practices rely on rule-based approaches, requiring domain expertise (Stonebraker and Ilyas, 2018) and manual efforts. A study by Kim et al. (2003) identifies 33 "dirty data" types of which at least 24 require intervention by a domain expert, while only nine are handled automatically. To cope with increasing data volumes and to overcome bottlenecks, EDM must find ways to innovate practices.

Machine learning (ML) is seen as a promising solution because it can automatically learn rules from data and thereby drastically reduce the number of domain-expert interventions (Stonebraker and Ilyas, 2018). Even though researchers have been developing ML techniques to improve DQ since the early 2000s (e.g., Li et al. (2000) or Sarawagi and Bhamidipaty (2002)), the declining costs of data-related technologies have led to significant advances (e.g., deep learning architectures) and have also fostered the democratization of ML methods and tools. As a result, learning-based approaches are receiving wide recognition as a viable solution to facilitate EDM, with interesting examples emerging from research (e.g. Zhu et al., 2014) and practice (e.g. Bean, 2017). The most prominent example is *Data Tamer* (commercialized under the name *Tamr*), a data curation system developed at the Massachusetts Institute of Technology, which involves ML and has been able to reduce EDM costs in three real-world examples by about 90% (Stonebraker et al 2013). More recent studies show that ML outperforms other approaches in predicting missing data values (called data imputation) (R. Wu, Zhang, et al., 2020) or creating "golden" records from duplicated data (called record fusion) (Heidari et al., 2020).

While some evidence point to the significant potential of ML to support EDM, the existing studies are scattered and develop solutions to very specific DQ issues. To date, there has neither been a meta-analysis of ML application areas in this domain, nor have the ML techniques been linked to foundational EDM concepts and to the data production processes of enterprises. Through this study, we aim to add a more systematic understanding of this dynamic and rapidly

evolving field by asking the overarching question: *How do machine learning techniques support enterprise data management?* Accordingly, we address two sub-questions:

RQI: *Which elements describe machine learning techniques for enterprise data management?*

RQII: *Which archetypes of machine learning for enterprise data management can be distinguished?*

The main contributions of our study are twofold. First, we provide a taxonomy that assists the identification and classification of ML techniques for EDM. Following Nickerson et al.'s (2013) taxonomy development guidelines, we applied a combined deductive-inductive approach to develop our classification framework. This taxonomy is an important prerequisite to understand the ML application areas in EDM. It represents *"systems of groupings which are derived conceptually or empirically"* (Nickerson et al 2013, p.3) and is useful to structure a dynamic field, as it orders disorderly concepts while describing their natures and the relationships between them. According to Gregor's classification of theory types, a taxonomy represents a *theory of analyzing* and is one of the most basic forms of theory that provides the foundation for more advanced explanatory and predictive theories. Second, based on this taxonomy, we analyze 60 cases of ML scenarios in EDM. To obtain a comprehensive empirical basis, we rely on three main sources, namely academic literature, focus groups and expert interviews, and an analysis of emerging ML-based DQ tools. We classified these cases and identified nine archetypes, as homogeneous groups of typical application scenarios, based on the empirically observed ML techniques.

Our findings contribute to EDM and DQ research. In accordance with the framework for DQ research suggested by Zhu et al. (2014), our study is positioned as a contribution to classify database-related technical solutions for DQ by using the methods of AI and data mining. Furthermore, by structuring the field, our findings lay the groundwork to advance EDM practices with ML and to integrate learning-based approaches into existing data production processes. Additionally, the suggested taxonomy and archetypes guide practitioners in assessing and selecting potential application areas of ML to improve their EDM.

The remainder of this paper is structured as follows. In order to position our study and to identify the research gap, we begin by summarizing relevant literature on managing data and improving DQ. In the next section, we provide a detailed overview of our research process and explain how we developed the taxonomy and derived the archetypes. Thereafter, we present the main research outcomes, namely the taxonomy and the archetypes. We conclude with a discussion of our findings and an outlook for further research.

2 Background

Over the past decades, IS and computer science communities have investigated DQ from two different but complementary perspectives. First, IS research (Ballou and Pazer, 1985; Legner et al., 2020; Wang et al., 1995; Zhu et al., 2014) emphasizes that DQ must not only be addressed as a technical problem, thereby introducing the organizational perspective on data production processes with EDM. Second, database research develops advanced techniques to curate data and improve DQ in databases, involving data mining and machine learning.

2.1 IS Research: DQ in the Context of EDM

IS researchers define DQ as a context-dependent, multidimensional concept, namely *"as data that are fit for use by data consumers"* (Wang and Strong 1996, p. 6). They argue that DQ problems do not arise for technical reasons alone, but that they are primarily organizational in nature. Earlier studies compared information systems to manufacturing systems for physical goods (e.g., Arnold 1992) and coined the term "data production process," defining it *"as the process that transforms a set of data units into [predefined] information products"* (Ballou et al. 1998, p. 463). This perspective construes data as value transmitters from data providers to either internal or external customers. Consequently, the provider-consumer view is the predominant lens to analyze DQ issues from an organizational perspective. Along the data production process, Strong, Lee, and Wang (1997) distinguish three distinct groups of professionals with varying interests and responsibilities (see Table 33). At the beginning of the data production process, the initial data input is provided by data collectors or producers. The input is either done by manual data entry into systems or is created automatically, e.g., by machines or software. However, the data input, especially in the manual case, is often incorrect or has missing values. These issues directly impact data consumers, who use the provided data for their task at hand, i.e., integration, aggregation, presentation, or interpretation. To fit DQ to the needs of the data consumers, data custodians play an intermediary role by cleaning and maintaining data.

To resolve DQ problems, Wang (1998) introduced the Total Data Quality Management Methodology (TDQM), which conceptualized the delivery of quality data products to data consumers. Ever since, several approaches have been proposed to solve DQ problems. Master data management is an approach that has gained broad acceptance and is a core EDM activity in many enterprises (Otto and Österle, 2015). Smith and McKeen (2008) define it as:

> "*An application-independent process which describes, owns and manages core business data entities. It ensures the consistency and accuracy of these data by*

> *providing a single set of guidelines for their management and thereby creates a common view of key company data, which may or may not be held in a common data source.*" (pp. 65-66)

Hence, major EDM tasks are the definition of adequate business rules to prevent DQ issues and the proactive unification of data across systems. These rule-based approaches require the involvement of domain experts, but they do not scale well with increasing data demands. Therefore, EDM often requires significant efforts to resolve DQ issues in a reactive manner. A typical example is the situation when a company acquires another corporation. When systems containing interrelated data are merged, EDM is confronted with duplicated data (e.g., customer relationship management (CRM)). Although rule-based approaches help to a certain extent (e.g., customer records can be considered duplicates when they have the same company name and address), domain experts (e.g., account manager) are eventually required to confirm the resolved entities.

Table 33. Roles in the data production process

Role	Description	Responsibilities
Data collectors or producers	People or other sources who create/produce data as initial input to the organization.	Enter, create, and collect data
Data custodians	People who maintain data and coordinate data storage and distribution.	Design, develop, maintain, store, protect, and distribute data
Data consumers	People who consume data in the way that they integrate, aggregate, present, and interpret it.	Integrate, aggregate, present, and interpret data

2.2 Database Research: Data Issues and Data Curation

Data issues and the techniques that address them are among the key topics of database research. For instance, Kim et al. (2003) provide a taxonomy of "dirty data" that results either from *missing* or from *non-missing* data. *Non-missing* data can be either *wrong data* due to the *non-enforceability* of *integrity constraints*, or *wrong data that is unusable*, e.g., because of duplication or ambiguity. The same study points to a low level of automation and high manual efforts, since most "dirty data" types require intervention by a domain expert. In order to address these data issues, data curation refers to activities *"to maintain and add value to data over its lifetime, and more specifically the tools and algorithms that attempt to reduce human curation effort by automating some of these important activities"* (Arocena et al. 2011, p. 47). Arocena et al. (2011) highlight eight data curation tasks that have received significant attention in database literature,

i.e., data cleaning, entity resolution, data transformation, data integration, data provenance, metadata or schema discovery, data and metadata profiling, and data archiving. Different data curation techniques have been developed (see Table 34), partially using data mining and ML, for different data types. For instance, these techniques focus mainly on relational data (structured data) and operate on data and/or data's semantics (constraint/schema) to detect and correct errors or resolve entities. Other techniques stem from the semantic web research community. These techniques include methods to extract specific data attributes from HTML documents (semi-structured data) or to unify data by mapping it to a given ontology. While most of the suggested techniques follow a rule-based approach (e.g., manually adjusting weights or defining thresholds), only a few use a learning-based mode.

Table 34. Taxonomies of data curation techniques

Source	Data curation tasks	Dimensions addressed
(Ilyas and Chu, 2015)	Data cleaning (Anomaly detection)	Error types, automation, BI layer
	Data cleaning (Data repairing)	Repair target, automation, repair model
(Rahm and Do, 2000)	Data cleaning	Single-source and multi-source problems Schema level and instance level
(Köpcke and Rahm, 2010)	Entity resolution	Entity type, blocking methods, matchers, combination of matchers, training selection
(Rahm and Bernstein, 2001)	Schema matching	Instance vs. schema, element vs. structure matching, language vs. constraint, matching cardinality, auxiliary information
(Shvaiko and Euzenat, 2005)	Schema matching	Input, process, output
(Mukkala et al., 2015)	Ontology and schema matching	Element level, structure level
(Shvaiko and Euzenat, 2013)	Ontology matching	Input, output, GUI, operation, terminological, structural, extensional, semantic
(Simmhan et al., 2005)	Data provenance (in E-science)	Use, subject, representation, storing, dissemination

While the use of rule-based approaches is effective when data are static, these approaches are less useful when data are dynamic and change over time (Volkovs et al., 2014). In the latter case, approaches that learn from user interactions and improve over time (Volkovs et al., 2014; Yakout et al., 2011) are more suitable. In particular, because of the advent of Big Data and data lakes, learning-based approaches are required that enable the discovery of and experimentation with different data types. Hence, data lakes call for new methods to manage data's heterogeneity,

while keeping the data consistent and clean (Nargesian et al., 2019). With the recent developments in ML, in particular deep learning, researchers emphasize its potential to curate Big Data (Thirumuruganathan et al., 2018). Accordingly, the first research prototypes accentuate the potential of ML in EDM (Stonebraker et al. 2013, Thirumuruganathan et al. 2018).

2.3 Research Gaps

The IS research stream emphasizes the organizational, holistic perspective on EDM, specifically the design of data production processes to improve DQ. Database research has produced techniques to solve typical data issues, e.g., for entity resolution (Rahm and Bernstein, 2001) or data repairing (Ilyas and Chu, 2015). However, these approaches focus mainly on data maintenance, while – as Chen et al. (2010) notice – approaches are missing to improve DQ at the point of data entry. A link between the IS and database research streams seems to be an essential step to advance EDM, which still requires considerable manual efforts and relies on processes that are to a large extent not automated (Abedjan et al. 2016; Stonebraker and Ilyas 2018). With data lakes, data becomes even more dynamic and new approaches are required to manage semi-structured and unstructured data. Despite first research endeavors, a thorough understanding of how ML can support EDM is still lacking. This understanding would provide guidance to practitioners when selecting suitable ML approaches in their enterprise contexts and would present researchers with an opportunity to contribute to the advancement of EDM practices.

3 Methodology

To address these research gaps, we followed the taxonomy development process suggested by Nickerson, Varshney, and Muntermann (2013), as it provides a concise method and is frequently applied to structure emerging fields in the IS domain (Beinke et al., 2018; Püschel et al., 2016).

3.1 Data Collection

To obtain a comprehensive empirical basis, we used three different sources to identify cases that use ML techniques for EDM (see Table 35). First, we reviewed scientific literature to identify specific ML techniques that have been suggested by the database community for data curation. Second, we conducted two focus groups with EDM experts and five expert interviews, which provided us with a better understanding of their vision of using ML and with concrete examples. This source resulted in twelve cases. Third, we screened the market for innovative tools that

offer ML techniques for EDM. Following the review our set consisted of 60 cases, which we used for taxonomy development.

Table 35. Sources considered for taxonomy development

Sources	Applied method	ML techniques
Academic literature	Literature review to identify suggested ML techniques for EDM in research.	29 academic cases
Focus groups and expert interviews	Focus groups and expert interviews to identify ML techniques that companies have started to explore and use for EDM.	12 practitioner cases
Market analysis	Screening of tools and suites that offer ML techniques for EDM.	19 applications

3.2 Taxonomy Development

As a first step, Nickerson et al. (2013) suggest determining the purpose of the taxonomy. This, in our case, is to describe and classify ML techniques for EDM in a systematic way. Moreover, the taxonomy will assist researchers and EDM practitioners to easily understand the potential use of an ML technique and to obtain an overview of the general application scenarios for ML in EDM. This purpose materializes in meta-characteristics that guide the development of the taxonomy's dimensions and characteristics. Based on the literature review, we define two meta-characteristics that reflect the IS and database research perspectives, i.e., the *EDM context*, which represents the organizational situation in which ML is applied; and the *ML application*, which describes the implementation details.

The second step requires the determination of ending conditions, which are validated after each iteration in order to either continue or terminate the development process. We rely on the suggested ending conditions proposed by Nickerson et al (2013), comprising – with one exception – objective and subjective criteria. We did not use the condition that every object must be classified under at least one characteristic, as non-classified characteristics allow the identification of potential areas of future research. The third step involves the selection of either a conceptual-to-empirical (CTE) or an empirical-to-conceptual (ETC) development process. In the CTE iteration, dimensions and characteristics are conceptualized first, using relevant literature (step A4c). Thereafter, these dimensions and characteristics are used to classify empirical observed objects. Based on this classification, the taxonomy is then created or revised (steps A5c and A6c). In the ETC iteration, the empirical, observed objects are examined first (step A4e). Based on an analysis of differences and commonalities, a distinct set of dimensions and characteristics is derived (step A5e), which leads to a revised or new taxonomy (step A6e).

We developed our taxonomy in four iterations (see Figure 7), as described in the next sections.

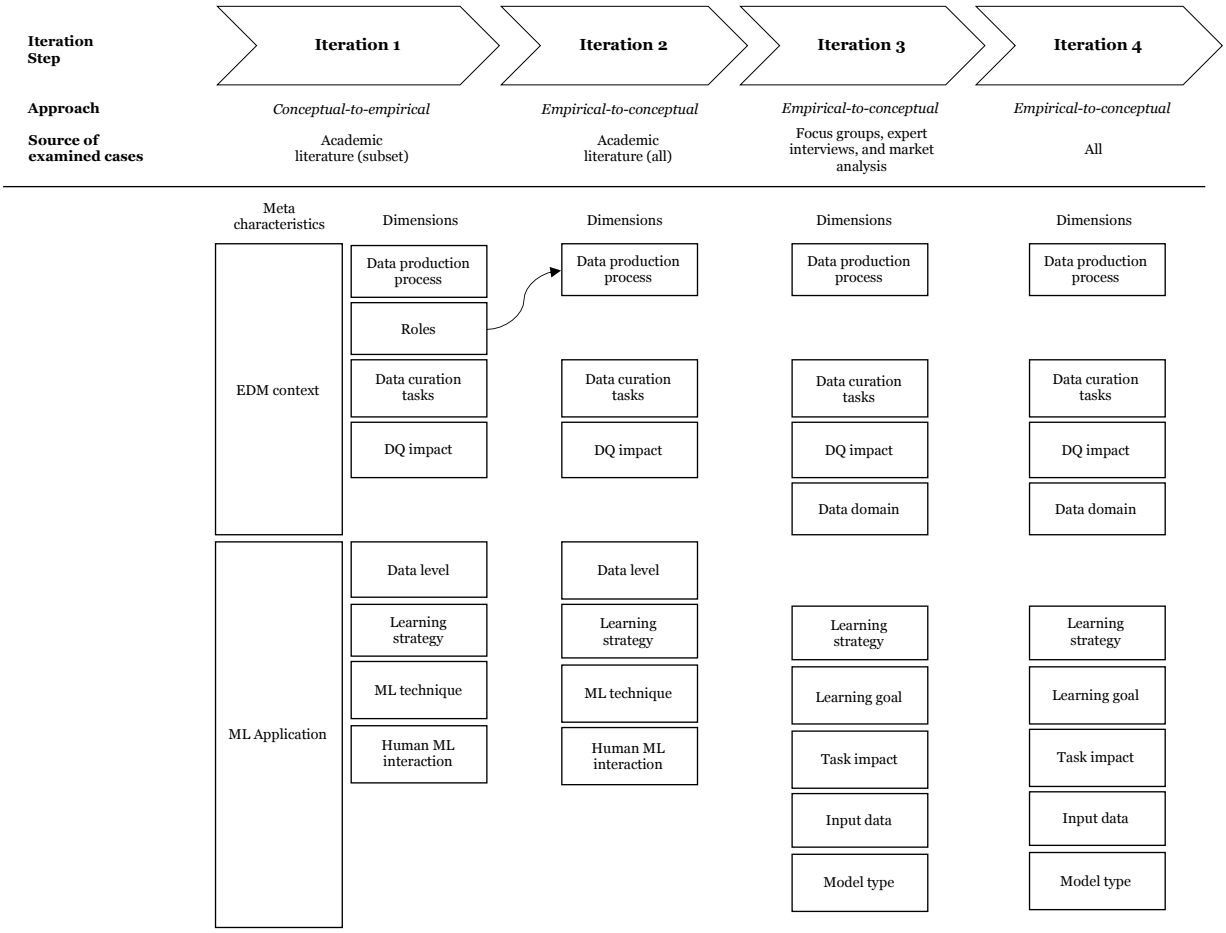


Figure 7. Taxonomy development iterations

Iteration I: Conceptual-to-Empirical (Subset of academic cases)

In the first iteration, we used a CTE approach to base our taxonomy on categories found in EDM, data curation and in ML literature, in order to demonstrate scientific rigor and to ensure that all dimensions and characteristics are sufficiently expressive to answer our first research sub-question. To come up with a comprehensive description of the *EDM context*, we derived dimensions from the DQ and database literature. We selected the dimension *Roles* to describe the target user and the dimension *Data curation tasks* to define the scope of the ML technique. To assess the benefits of each ML technique, we added the dimension *Data quality impact*. For the meta-characteristic *ML application*, we reviewed literature in the domains of data curation, data mining and ML. In order to gain a general understanding of the applied ML techniques, we considered three dimensions. We drew the dimension *Data level* from an existing taxonomy of

ontology matching methods, which seemed to suit our context (Shvaiko and Euzenat, 2013). We added the classical distinction of ML techniques in the dimension *Learning strategy*, which were derived from the ML literature (James et al., 2013). The characteristics of the dimension *ML method* were taken from data mining literature, which proposes general categories applicable to ML (Fayyad et al. 1996, James et al. 2013) and which describes the method at an adequate level of detail. By using this set of dimensions and characteristics, we examined a subset of the academic cases and realized that the dimensions *Data production processes* and *Roles* express near-similar aspects. Therefore, we merged *Roles* with the dimension *Data production process*.

Iteration II: Empirical-to-Conceptual (All academic cases)

In the second iteration, we used an ETC for all cases found in the academic literature. Furthermore, we presented and discussed the resulting version of the taxonomy in focus groups and expert interviews to assess the subjective ending conditions from an external point of view.

Iteration III: Empirical-to-Conceptual (All practitioner cases and applications)

In the third iteration, we again chose an ETC approach. This time, we examined only practitioner cases and applications, while also considering the feedback received from the focus group. We added the dimension *Data domain* on the EDM side as suggested by the EDM experts, removed the dimension *Data level*, and instead added the dimension *Data input* that explains the type of data used for the ML technique in a more comprehensive way. Furthermore, we renamed the dimension *ML technique* as *Learning goal* and added the dimension *Model type*. While *Model type* describes the used ML model from a technical viewpoint, the dimension *Learning goal* depicts outcome when using a particular ML model. In addition to our previous changes, we renamed the dimension *Human-ML interaction* as *Task impact*; a label that, to an extent, is congruent with the dimension *Data quality impact* on the ML application side of the taxonomy.

Iteration IV: Empirical-to-Conceptual (All cases)

In the fourth iteration, we examined all cases. As we were unable to apply any further modifications and since we complied with all ending conditions, we concluded the development process with the fourth iteration.

3.3 Evaluation

While we evaluated the taxonomy ex ante by assessing the ending conditions after each iteration, we also applied an ex post evaluation step to test the viability of the final taxonomy. The entire case base was classified independently by two researchers, according to the

taxonomy's dimensions and characteristics. Thereafter, we calculated the intercoder reliability using both classifications to assess the taxonomy's robustness (see Table 36). To avoid "by chance agreement," we used Cohen's Kappa as a measure for intercoder reliability (Cohen, 1960; Lavrakas, 2008). As expected, considering the number of previous development iterations and extensive evaluation steps, the intercoder reliability was high for both meta-characteristics, yielding coefficients close to 0.8 or above for all dimensions.

Table 36. Intercoder reliability (Cohen's Kappa)

	Dimension	Cohen's Kappa
Data Management	Processes (E-to-C)	0,97
	Data domain	0,82
	Data quality impact	0,90
	Task (C-to-E)	0,89
ML Application	Data input type (C-to-E)	0,89
	Learning strategy (E-to-C)	0,79
	Model type	0,78
	Learning goal (C-to-E)	0,85
	Task impact	0,83

3.4 Application

Given the coding's high reliability scores, it was assumed to be stable enough to answer our second research question: *Which archetypes of machine learning for enterprise data management can be distinguished?* In a first step, we conducted a frequency analysis of the characteristics found in each classified ML technique. In a second step, we analyzed the patterns using a qualitative clustering of the cases based on the classification agreed upon by the researchers and used in the frequency analysis. We identified the archetypes in the means of typical application scenarios of ML in EDM through a qualitative analysis of the classified case base. Accordingly, an archetype represents a group of ML techniques that follow a certain pattern in at least one of the taxonomy's dimensions; a group that can be viewed as a homogenous.

As we were interested in understanding how ML techniques support EDM, we started with the dimensions of the meta-characteristic *EDM context*, more specifically the dimension *Data production process*, before integrating the dimensions associated with the *ML application*. The first authors grouped use cases with a similar coding as the other dimensions of the meta-characteristics *EDM context* and *ML Application* into clusters of similar cases. The authors thereby identified nine clusters within specific data-production processes that showed high coding overlaps, especially for the dimensions *Data curation task*, *Data quality impact*, and

Learning goal. To validate the comprehensibility of the clusters, the second author classified the cases into the identified clusters that yielded similar results as in the first run.

4 A Taxonomy of ML Techniques for EDM

In the following sections, we present the final version of the taxonomy and answer our first research question (RQI): *Which elements describe applications of machine learning techniques for enterprise data management*? The final taxonomy comprises nine dimensions that are structured – in accordance with the meta-characteristics – into the *EDM context* and *ML application*. Table 37 presents the different dimensions and characteristics of the taxonomy with a footnoted indication if a characteristic was added to the empirical-to-conceptual iteration.

4.1 EDM Context

The *EDM context* characterizes the specific situation in which the ML technique is applied to solve a data-related issue and reflects the IS research perspective. We propose four dimensions: the *Data production process* (the high-level process), the *Data domain* (the data's use context), the *Data curation task* (the specific activity performed to curate data), and the *Data quality impact* (the benefit derived from data curation).

Table 37. Taxonomy of ML techniques in EDM

	Dimension	Characteristics				References
EDM context	**Data domain**	Party	Thing	Location	Any	(Cleven and Wortmann, 2010)
	Data production processes	Acquire and create[1]	Unify and maintain[1]	Protect and retire[1]	Discover and use[1]	(Strong, Lee, and Wang 1997)
	Data curation tasks	Data cleaning	Entity resolution	Data trans-formation	Data integration	(Ilyas and Chu, 2015) (Rahm and Do, 2000) (Rahm and Bernstein, 2001) (Pavel Shvaiko and Euzenat, 2005) (Mukkala et al., 2015) (Köpcke and Rahm, 2010) (Elmagarmid et al., 2007)
		Metadata or schema discovery	Data archiving	Data enrichment[1]	Data monitoring[2]	
	Data quality impact	Intrinsic	Contextual	Representa-tional	Accessibility	(Wang and Strong 1996)
ML application	**Input data**	Structured[1]	Semi-structured[1]	Unstructured[1]	Any	(G. Li et al., 2008)
	Learning strategy	Supervised	Semi-supervised	Unsupervised	Reinforcement learning[2]	(James et al., 2013)
	Learning goal	Classification	Regression	Clustering		(Fayyad et al., 1996) (James et al., 2013)
		Summarization	Dependency modeling	Change and deviation detection		
	Model type	Shallow[1]		Deep[1]		(LeCun et al., 2015)
	Task impact	Substitution	Augmentation	Assemblage		(Rai et al., 2019)
Legend: [1] characteristic added in the empirical-to-conceptual iteration; [2] no object could be classified under this characteristic						

Data domain

Enterprise data can be organized by data domains that characterize the data's context of use or origin. Three general data domain categories, common to master data but also applicable to other data types, are broadly distinguished (Cleven and Wortmann, 2010): *Party* (data domains: customers, suppliers, distributors, employees, or citizens), *Thing* (data domains: products, services, or assets), and *Location* (data domains: places, sites, or regions).

Data production processes

EDM coordinates data production processes. First, *Acquire and create* means that data are acquired from an external source or created either manually by a data collector or automatically by a machine. The subsequent process, *Unify and maintain*, is typically performed by a data custodian, and aims at cleaning and integrating data across multiple systems to build a unified view. *Protect and retire* ensures that sensitive data are safeguarded, but also that data are removed due to regulations or when outdated. Through *Discover and use*, data consumers find relevant data and use it in their daily work-related practices. Because these processes overlap considerably with the roles of the data collector, data custodian, and data consumer identified in Wang et al.'s seminal paper, they are therefore merged with the ETC iteration.

Data curation tasks

This dimension emphasizes the different tasks that ML techniques support to curate data. Based on prior literature (Arocena et al., 2011), we distinguish between six data curation tasks. *Data cleaning* aims to detect and repair quantitative and qualitative data errors (Abedjan et al., 2016). *Entity matching* identifies and resolves data belonging to the same real-world entity, e.g., duplicated records in a relational database (Köpcke and Rahm, 2010). *Data transformation* aims to transform data from a source format into a target format. *Data integration* is a common EDM task when data from heterogenous sources need to be combined to create a unified view. In this context, *Ontology and schema matching* is applied (Mukkala et al., 2015). In *Metadata discovery*, metadata are extracted from various data types. *Data archiving* moves outdated data that are no longer required to long-term storage. Through empirical examination in the ETC iteration, we add *Data enrichment* to extend the initial attribute set of a data record for later usage in additional contexts.

Data quality impact

To capture the benefits of ML techniques for EDM, we rely on Wang and Strong's (1996) hierarchical framework that classifies DQ in terms of four categories: *Intrinsic*, *Contextual*, *Representational*, and *Accessibility*. *Intrinsic* DQ comprises attributes that assess whether a data value conforms with the actual or true value of an object and includes believability, accuracy, objectivity, and reputation. *Contextual* DQ defines attributes that reflect whether data are applicable to a certain context or purpose as defined by the user. These attributes are value addition, relevancy, timeliness, completeness, and the appropriate amount of data. The *Representational* and *Accessibility* categories consider DQ from a system perspective; data needs to be presented in a way so that it is not only interpretable, but also available to or obtainable

by a user. *Representational* DQ comprises the attributes of interpretability, ease of understanding, representational consistency, and concise representation. Finally, *Accessibility* DQ includes accessibility and access security.

4.2 ML Application

The *ML application* meta-characteristic comprises five dimensions, namely the *Input data* (the data type to train the machine learning model), the *Learning strategy* (the way in which the machine learning model is trained), the *Learning goal* (the output of a machine learning model), the *Model type* (the machine learning model architecture), and the *Task impact* (from a task perspective, the benefit of using the machine learning model).

Input data

ML techniques learn and operate on a certain data input to fulfill their learning goals. We differentiate between the following categories of data types for the *Input data* dimension (G. Li et al., 2008), namely structured data (e.g., records, schemas, or transactions), semi-structured data (e.g., HTML documents, log-files, or rules), and unstructured data (e.g., documents, pictures, or videos).

Learning strategy

ML techniques are generally classified according to their learning strategies (James et al., 2013). First, learning can be done in a *Supervised* way where the model is trained with a labeled data sample and learns a function that maps an input to an output. Second, through an *Unsupervised* learning strategy, the model is trained with an unlabeled training sample and learns structures in the input data on its own. Third, *Semi-supervised* learning combines supervised and unsupervised learning strategies, typically using a small amount of labeled data and large amount of unlabeled data for training. *Reinforcement learning* is a learning strategy that requires no training data since it learns through trial and error. No ML technique for EDM was found that uses *Reinforcement learning*. Therefore, we excluded this *Learning strategy* from the taxonomy.

Model type

Typically, an ML technique can be categorized into two types: *Shallow* and *Deep*. An ML technique is *Shallow* when it uses only one processing layer (e.g., support vector machines, linear regression, or k-nearest neighbor), whereas it is *Deep* ("Deep learning") when it uses more than one processing layer. Recently, major breakthroughs have been made with *Deep* learning architectures (LeCun et al., 2015).

Learning goal

By using a certain ML technique, we aim to achieve a specific learning goal. To identify the different learning goals, we rely on data mining and statistical learning literature and distinguish between five distinct method categories (Fayyad et al., 1996; James et al., 2013). *Classification* refers to the methods where a function is learned that maps data into pre-defined categories. *Regression* methods are used to learn a function that maps data to a continuous, distributed variable. In addition, the relationship between variables forms part of this analysis. *Clustering* uses methods to identify groups of data that share the same characteristics or are closely related based on some measure of distance. *Summarization* describes a subset of data in a compact form. *Dependency modeling* models significant relations between variables. Through *Change and deviation detection*, a model learns to distinguish between normal and abnormal behavior in the data.

Task impact

ML redistributes tasks between machines and humans (Rai et al., 2019). In the case of *Substitution*, machines replace humans in doing the task at hand. With *Task augmentation*, machines and humans augment each other to increase their performance when executing the task. Finally, in the *Assemblage* scenario, humans and machines work as an integrated unit to complete the task. Thereby, this scenario extends beyond the first two, where the work is divided between humans and machines.

5 Applying the Taxonomy

We used the taxonomy to classify the collected cases that apply ML for EDM purposes, in order to answer our overarching research question: *How do machine learning techniques support enterprise data management?* In this section, we present a frequency analysis (see Table 38) and the derived archetypes as typical application scenarios of ML in EDM (see Table 39). Through these patterns, we also answer our second research questions (RQ II): *Which archetypes of machine learning for enterprise data management can be distinguished?* The identified archetypes represent a group of ML techniques that support a distinct data production process and that can be viewed as a homogenous group.

5.1 Frequency Analysis

Concerning the *EDM context*, the frequency analysis reveals that most ML techniques support *Acquire and create* and *Unify and maintain*. This pattern is expected as these are traditional EDM processes. *Protect and retire* and *Discover and use* are both emerging production processes, due to new data privacy regulations and a greater emphasis on the accessibility of data to a broader range of employees. While some ML techniques exist that are data domain specific, most of them can be applied in any domain. This classification in particular relates to the techniques that stem from academic literature. They often describe a general method that is data domain agnostic. The distribution in the *Data curation tasks* dimension is also as expected. A large set of ML applications are classified under *Data cleaning* and *Entity resolution*, both of which are core EDM activities to improve the DQ. *Data enrichment* is another frequent characteristic among the cases, specifically in the *Acquire and create* process, but it also gains in importance when detecting sensitive data in the *Protect and retire* process. Apart from the *Data quality impact* characteristic, *Accessibility*, the techniques equally often improve the *Intrinsic*, *Contextual*, and *Representational* DQ dimensions.

Regarding the *ML application*, half of the techniques operate on structured data, while a fourth use unstructured data as *Input data*. The latter, in particular, co-occurs with *Data transformation* as the *Data curation task*; a task in which, for instance, certain data attributes are extracted from text, e.g., an address. Most of the techniques use a *Supervised learning strategy*, which requires labeled training data. This learning strategy is also most common in practice and promises the best results. Conversely, most of the techniques follow the *Learning goal* of *Classification.* However, somewhat more than a fourth of the techniques also leverage *Dependency modelling*, using the *Unsupervised learning strategy*. The distribution in the dimension *Model type* is of

particular interest. While most of the techniques use *Shallow* models, the use of more complex *Deep* architectures is increasingly evident. This trend confirms the significance of recent advances in the domain of natural language processing, which is fundamental to EDM. In the *Task impact* dimension, most of the techniques are classified under *Augmentation*. In this respect, it is concluded that ML does not necessarily substitute EDM workers; rather, ML facilitates their jobs.

Table 38. Frequency analysis characteristics

	Dimension	Characteristics			
EDM context	**Data production processes**	Acquire and create (39.2%)	Unify and maintain (44.2%)	Protect and retire (8.3%)	Discover and use (8.3%)
	Data domain	Party (5.0%)	Thing (30.9%)	Location (0.0%)	Any (65.8%)
	Data curation tasks	Data cleaning (27.5%)	Entity resolution (16.7%)	Data transformation (15.8%)	Data integration (10.0%)
		Metadata or schema discovery (6.7%)	Data archiving (0.8%)	Data enrichment[1] (21.7%)	Data monitoring (0.0%)
	Data quality impact	Intrinsic (38.3%)	Contextual (26.7%)	Representational (26.7%)	Accessibility (8.3%)
ML application	**Input data**	Structured (51.7%)	Semi-structured (12.5%)	Unstructured (22.5%)	Any (13.3%)
	Learning strategy	Supervised (65.0%)	Semi-supervised (10.8%)	Unsupervised (24.2%)	Reinforcement learning (0.0%)
	Learning goal	Classification (56.7%)	Regression (1.7%)	Clustering (3.3%)	
		Summarization (3.3%)	Dependency modelling (29.2%)	Change and deviation detection (5.8%)	
	Model type	Shallow (63.3%)	Deep (36.7%)		
	Task impact	Substitution (15.8%)	Augmentation (75.8%)	Assemblage (8.3%)	

Color coding:

>= 80%	>= 60%	>= 40%	>= 20%	>= 10%	< 10%

5.2 Archetypes: ML Techniques for EDM

Archetype 1: Support manual data entry

This archetype comprises ML techniques for EDM that support manual data entry and ensure that data are entered into systems in the expected way. For instance, Chen et al. (2010) (ID 5) propose an ML application that uses a Bayesian network trained on a labeled dataset to learn in which sequence questions must be presented in forms, in order to improve DQ. When a user starts to enter data, the form adapts dynamically – based on the provided user input – and re-asks questions to ensure that it is correctly completed. In this case, the ML application works dynamically together with the user. Toda et al. (2010) (ID 1) demonstrate a scenario where an ML model is trained in a similar way, but data is provided as full text and the algorithm automatically extracts the values for the form from the text. Ali and Meek (2009) (ID 4) describe an application where even the form values are predicted, based on a learned probabilistic network, once the user enters a value into the form. In the enterprise context, this scenario could be applied to structured data, which must be collected and entered manually into the system. Here, a probabilistic network is learned from previous data entries. ML can also go beyond predicting the structure and values of a form and correct syntactic errors with the help of natural language processing. In this case, the ML technique learns to detect errors in text based on pairs of erroneous and correct text fragments (Grammarly 2018, ID 2).

Archetype 2: Automated transformation of data

In this archetype, we summarize ML techniques for EDM that automate the manual transformation of data from a source format into a target format. In the enterprise context, data are often provided as text in documents, where certain data attributes need to be extracted. ML has the potential to automate these tasks. For instance, Sarawagi et al. (2001) (ID 13) trained a Hidden Markov Model with a set of labeled training data to learn the automatic extraction of address data from free text fields; a problem that is common in corporate databases. Although not limited to address data, Hu et al. (2017) (ID 9) trained a convolutional neural network to extract text segments from free text without the need of labeled data. By using three real-world datasets, they demonstrate the possibility of extracting person, house, and car-related attributes from free text and outperforming state-of-the-art approaches by 10%. Beside the extraction of data attributes from text, ML can also be used to detect data attributes from pictures. In one of the cases a convolutional neural network was trained on a labeled dataset to identify material of the company's infrastructure in pictures (ID 12). This classifier is used to allow the end-customer to submit defects via a picture over a mobile app, the moment they detect it. A similar approach

is under development in another enterprise where pictures are used to search for product information (ID 10). In addition, data can also be generated automatically. AX semantics (2018) uses natural language generation methods, which apply deep learning and generative adversarial networks to produce product descriptions (ID 15). The user needs to provide an initial input and the application then automatically transforms this input. Finally, enterprises – as multinational corporations – often need to have their text translated into different languages. Through deep learning, human language can be represented and text can be transformed in a numerical format that also captures the semantics. Based on a training dataset of pairs of text segments in two different languages, a model can be trained to translate between both languages (DeepL 2018, ID 8; Wu et al. 2016, ID 14).

Archetype 3: Support data enrichment

This archetype bundles several ML techniques that support data enrichment. This often requires the classification of existing data records and documents into predefined categories. For instance, assigning product data records to the correct product category is a prerequisite to make them findable and accessible in E-commerce applications. In a similar way, customs and trade regulations require enterprises to classify their product and customer data according to international standards. ML can support the assignment of codes and categories based on training data. In the case of material master data, a random forest classifier was trained on a labeled training dataset to predict tax tariffs, achieving an accuracy of 90% (ID 23). This implies that only assignments with low confidence need to be reviewed by a human expert, resulting in a highly scalable process that frees the capacity of experts for other tasks. In order to support the publication of products in multiple E-commerce shops (e.g., Amazon), commercetools (2018) (ID 21) provides an ML application that automatically classifies product names in product categories, depending on the specific E-commerce shop. In a first step, this application represents the product names in a numerical form and then trains a logistic regression algorithm using a labeled dataset that correctly assigns product names to their corresponding categories. Reltio (2018) (ID 22) has developed further uses of data classification, e.g., by adding customer segmentation attributes like purchasing power or churn propensity, based on address, purchasing and interactions data.

Archetype 4: Support data cleaning

This archetype comprises ML techniques that help to clean data and improve DQ in a reactive manner. ML supports reactive approaches to correct data errors by detecting quantitative errors (as outliers) or qualitative errors (in the form of duplicates, rule or pattern violation) (Abedjan

et al., 2016). Errors can be either wrong values or constraint violations. ML not only helps to detect these errors, but also to repair them. In this archetype, we find use cases that leverage an active learning strategy where the human and ML application dynamically interact with each other (assemblage). For instance, to predict types of data repairs, Volkovs et al. (2014) (ID 26) propose a system of continuous data cleaning where a logistic classifier learns from past user repair preferences. As the user selects the types of data repairs needed to resolve an inconsistency, the ML technique incorporates this feedback to improve its future accuracy. This approach is of great advantage in environments where data change continuously and challenge prevailing practices that are applied in static environments (e.g., Chiang and Miller 2011). In the previous case, the correction of errors still remains a task that is done by a human, but approaches exist that automatically repair data errors. For instance, Wu et al. (2020) (ID 32) suggest deep learning-based architecture to predict the missing values of continuous and discrete data in datasets. As complete automatic repairs remain risky, Yakout et al. (2011) (ID 27) propose an application that leverages active learning. This application only suggests data repairs to the user, who then denies or confirms the updates, iteratively. The system continuously improves its precision by increasing user feedback.

Archetype 5: Support data matching

With this archetype, ML techniques support the detection of duplicated data. Generally, the finding and matching of entities that share identical or similar characteristics is a common problem in enterprises. For instance, duplicated records frequently occur in the case of mergers and acquisitions, since companies tend to have an overlapping customer or vendor base. As companies collect more and more data from internal and external sources, consolidating this data and creating a "golden record" for a given entity poses a challenge. This archetype can be linked to the research on entity matching, which has accumulated a rich body of knowledge over the past decade. ML has not only demonstrated its ability to improve the performance, but also the usability of entity matching systems by becoming more human-centric and interactive (Doan et al., 2017). Sarawagi and Bhamidipaty (2002) (ID 37) propose a deduplication function that is learned from a labeled dataset with less than ten records, through an interactive approach where the algorithm asks the user to label the most challenging pairs of duplicates. They show that the amount of required training data can be reduced by two orders of magnitude to achieve a certain level of precision. Other approaches for deduplication leverage both supervised and unsupervised techniques (Elmagarmid et al., 2007). A recently published approach by Mudgal et al. (2018) (ID 38) uses deep learning and outperforms existing approaches when matching textual and "dirty structured" instances. To resolve entities, Wu et al. (2020) (ID 39) proposes an

unsupervised approach that requires no training data and that attains a comparable level of performance as supervised learning methods. Heidari et al. (2020) (ID 40) apply ML to create a unique record, known as a "golden" record, from a set of duplicate candidate pairs. A particular company intends using ML to classify matching pairs in order to harmonize its product master data with a commonly used industry-related product standard. Reltio (2018) (ID 35) provides an ML solution to match data from various sources for the same entities, e.g., customers or employees, using an active learning approach.

Archetype 6: Support data integration

This archetype contains ML techniques that help to integrate data from different systems. In order to remain agile on the market, large corporations are typically divided into several business units that have some measure of freedom to manage their data. As these business units define their own data entities and schemas, data cannot easily be shared across business units and remain siloed. This is also the case when companies merge. In research, multiple techniques have been proposed for semi-automated and automated data integration, but most of the traditional approaches are not yet scalable for large amounts of data as they are rule-based and rather static approaches (Stonebraker and Ilyas, 2018). ML provides great potential to these scenarios. Stonebraker et al. (2013) (ID 43), using *Data Tamer* (commercialized under the name *Tamr*), developed an end-to-end integration system that leverages ML and dynamically interacts with the user. It has proved to reduce curation costs in three real-world settings by 90%. In the Tamer system, data are integrated in two phases. First, an individual schema attribute is compared to other attributes in a pairwise manner with different similarity measures. The user is only prompted to intervene when the combined measure of similarity falls under a certain threshold. Second, ML is used to deduplicate the integrated records. The Learning Source Descriptions (LSD) system presented by Doan et al. (2001) (ID 44) follows a supervised learning strategy to find semantic mappings between schemas. To enable this, they formulate schema matching as a classification problem and train a set of learners (ensemble) on schema and data-related features, and also incorporate user feedback to further improve accuracy. In their experiments, LSD achieved a 71–90% precision level. Experts recognize the promising potential of deep learning to improve schema-matching approaches (Mukkala et al., 2015). Through recent advances, word vector representations trained on a large corpora of text can capture syntactic and semantic regularities in text (Pennington et al., 2014). This has always been an issue for traditional schema approaches, which perform well on syntactic similarity comparisons but not on the comparison of semantics.

Archetype 7: Creation of data quality rules

This archetype comprises ML techniques that support the creation of DQ rules. The techniques learn the dependencies of items as rules from transaction data (Fan et al. 2009; Liu et al. 2012). Hipp et al. (2001) (ID 48) demonstrate the potential of EDM using association rules learning to extract significant association rules between items in business transactions. In the database literature, association rules are better known as conditional, functional dependencies, and multiple approaches have been suggested to mine these dependencies from transaction data (Fan et al. 2009; Chiang and Miller 2008; Liu et al. 2012). To extract rules of significance and interest with association rules learning, the user needs to define the parameters' support and confidence before running the algorithm. Support indicates how often an itemset occurs in the data, and confidence how often this itemset is true (Agrawal et al., 1993). After a few iterations, rules are discovered that can be implemented in systems to improve DQ (Hipp et al. 2001). In a case (ID 49), this approach was used to extend the existing set of 13 DQ rules for product data to approximately 400 rules. Of these rules, 25% could be applied to critical fields, which were the major focus of this endeavor. Another company applies clustering after the association rules mining step to identify relevant DQ rules in a faster and more scalable way (ID 50).

Table 39. Archetypes of ML techniques for EDM

Processes	Archetype	Description	Classification
Acquire and create	1. Support manual data entry Case IDs: 1 – 7	Learn data entry patterns to prefill values and to adapt the sequence of form elements for faster data entry and higher DQ by minimizing the risks of invalid/wrong data entries, blanks, or typos.	CURATION TASK: data cleaning DQ IMPACT: intrinsic DATA INPUT: structured STRATEGY: supervised, unsupervised LEARNING GOAL: dependency modelling TASK IMPACT: augmentation
	2. Automated transformation of data Case IDs: 8 - 19	Learn how to transform data from a source to a target format, i.e., extract structured data from texts, photographs, or videos, and how to translate text or automatically generate text from structured data.	CURATION TASK: data transformation DQ IMPACT: representational DATA INPUT: structured/unstructured STRATEGY: supervised, unsupervised GOAL: clustering TASK IMPACT: substitution
	3. Support data enrichment Case IDs: 20 - 24	Learn to classify records and documents to enhance further processing and analysis.	CURATION TASK: data enrichment DQ IMPACT: contextual DATA INPUT: structured/unstructured STRATEGY: supervised GOAL: classification TASK IMPACT: augmentation
Unify and maintain	4. Support data cleaning Case IDs: 25 - 33	Learn to detect and correct data errors from existing datasets and user feedback, in order to accelerate reactive data cleaning.	CURATION TASK: data cleaning DQ IMPACT: intrinsic DATA INPUT: structured STRATEGY: supervised GOAL: classification TASK IMPACT: assemblage
	5. Support data matching Case IDs: 34 – 42	Learn to identify similar data entities, in order to reduce the number of duplicates and to enhance data unification.	CURATION TASK: entity resolution DQ IMPACT: intrinsic, representational DATA INPUT: structured STRATEGY: supervised GOAL: classification TASK IMPACT: augment., assemblage
	6. Support data integration Case IDs: 43 - 47	Learn to link data and tables from heterogenous sources based on semantic and syntactic similarities, in order to accelerate data integration and discovery.	CURATION TASK: data integration DQ IMPACT: representational DATA INPUT: structured/semi-structured STRATEGY: supervised GOAL: classification TASK IMPACT: augmentation
	7. Creation of data quality rules Case IDs: 48 – 51	Learn the dependencies between data attributes to extract and discover new DQ rules, in order to facilitate proactive data management.	CURATION TASK: metadata discovery DQ IMPACT: intrinsic DATA INPUT: structured STRATEGY: unsupervised GOAL: summarization TASK IMPACT: augmentation
Protect and retire	8. Detection of sensitive and out-of-date data across systems Case IDs: 52 - 56	Learn to identify sensitive data and detect life-cycle events, e.g., when data needs to be retired to reduce the risk of non-compliance with data protection regulations.	CURATION TASK: metadata discovery DQ IMPACT: accessibility DATA INPUT: structured/unstructured STRATEGY: supervised GOAL: classification TASK IMPACT: augmentation
Discover and use	9. Support the discovery of relevant data Case IDs: 57 - 60	Learn data usage patterns and deep representations of data and tables to make dataset recommendations, in order to enhance discovery and use.	CURATION TASK: data integrat., enrich. DQ IMPACT: contextual DATA INPUT: structured/semi-structured STRATEGY: super-, unsupervised GOAL: classif., regress., clustering TASK IMPACT: augmentation

Archetype 8: Detection of sensitive and out-of-date data across systems

This archetype describes ML techniques that detect sensitive and out-of-date data. Regarding data protection regulations, e.g., General Data Protection Regulation (GDPR), enterprises are required to ensure complete transparency – across their systems – over identifiable personal data (Voigt and von dem Bussche, 2017). There are different tools and services that leverage ML to identify sensitive data across systems. Dathena (2019) (ID 54) offers a service where data is classified into different levels of confidentiality. Both Amazon Macie (2018) (ID 55) and Pingar (2018) (ID 56) offer a service that leverages ML to detect sensitive data across systems. In one of the cases, in order to be regulatory compliant, an ML classifier was trained on a labeled dataset to predict the timepoint when personal data must be retired (ID 53).

Archetype 9: Support the discovery of relevant data

This archetype encompasses ML techniques that help to discover relevant data for a specific purpose. With the availability of increasing amounts of data to companies and data scientists who are eager to utilize it, data discovery becomes a key capability to extract value from Big Data (Fernandez et al. 2018). Fernandez et al. (2018) (ID 57) use deep learning in order to link datasets from heterogenous sources, which are semantically related. Here, a semantic matcher unit uses deeply learned, word vector presentations (word embeddings) to link objects that share syntactic and semantic characteristics. In this way, data analysts can find relevant data more quickly. Enterprises have begun to establish data catalogs to provide enterprise-wide data access. In this case, ML becomes a core capability of these tools to support the discovery of relevant data. For instance, Alation (2018) (ID 59) leverage ML to recommend data that might be relevant, while the user is typing a query.

6 Conclusion and Outlook

In view of the increasing volume and variety of data, EDM is confronted by certain challenges. On the one hand, existing data management practices rely on human intervention and do not scale. On the other hand, the number of data consumers and their expectations increase with the proliferation of data science. Against this backdrop, our study sheds light on how ML techniques can advance EDM.

Based on the analysis of 60 ML cases from research and practice, our study makes two contributions. First, the suggested taxonomy provides a classification scheme that links ML techniques to EDM and data curation concepts. It thereby connects different research streams – previously unconnected – that address complementary organizational and technical aspects. We argue that this connection is important since ML techniques will significantly change the mainly manual EDM practices and lead to a redistribution of tasks between humans and machines. Therefore, a thorough socio-technical analysis is needed to determine how ML can be used in EDM and how it affects existing work practices. Second, the archetypes provide an overview of typical application areas of ML in EDM. Our analysis reveals that some archetypes build on the rich body of research that has developed over the past decades, for instance in the case of archetype 5 (entity matching or data integration). However, the existing research is scattered, and our study is among the first to provide a comprehensive overview of how ML can provide support in these scenarios. In addition, we also observe archetypes that open interesting, new fields of research, such as archetype 9 that applies ML in order to support the discovery of data by users. This observation shows that EDM goes beyond the traditional focus on DQ and that it places more emphasis on actively improving data discovery and use.

Our findings are relevant for both practice and research. The taxonomy and archetypes support practitioners in selecting and assessing suitable ML techniques to resolve their data problems. For purposes of research, the taxonomy helps to assess the impact of ML and to inform future EDM practices. Therefore, our findings provides the groundwork for future research on advancing EDM practices with ML. Based on Zhu et al. (2014)'s framework, we structure the field of database-related technical solutions for DQ, using the methods of ML and data mining.

Several interesting insights and implications emerge from our research. We find that ML supports both reactive and proactive EDM. Actually, ML techniques shift manual data maintenance efforts that were previously executed by data custodians (in a reactive mode) to the data collector (in a proactive mode). This implies that, with ML, EDM becomes an integrated part of each business function, rather than being delegated to specialized data management

units. While acknowledging the significant potential of ML, we find that although it provides support in most cases by proposing a solution, it does not completely substitute human activities. Although some part of the data-related tasks can be automated, human effort is still required in all of the cases. This observation creates manifold research opportunities related to the socio-technical design of data production processes that integrate augmentation or assemblage with ML techniques.

7 References

Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M., & Tang, N. (2016). Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*, *9*(12), 993–1004. https://doi.org/10.14778/2994509.2994518

Agrawal, R., Imielinski, T., Swami, A., Road, H., & Jose, S. (1993). Mining Association Rules between Sets of Items in Large Databases. *Proceedings of ACM Sigmod*, 10.

Alation. (2018). The First Data Catalog Designed for Collaboration. *Alation*. https://alation.com/product/

Ali, A., & Meek, C. (2009). *Predictive Models of Form Filling* (Technical Report MSR-TR-2009-1; p. 8). Microsoft Research.

Amazon Macie. (2018). Amazon Macie | Discover, classify, and protect sensitive data | Amazon Web Services (AWS). Amazon Web Services, Inc. https://aws.amazon.com/macie/

Arnold, S. E. (1992). Information Manufacturing: The Road to Database Quality. *Database*, *15*(5), 32–39.

Arocena, P., Glavic, B., Mecca, G., Miller, R. J., Papotti, P., & Santoro, D. (2011). Benchmarking Data Curation Systems. In J. Newman, *Green Education: An A-to-Z Guide*. SAGE Publications, Inc. http://sk.sagepub.com/reference/greeneducation/n8.xml

AX semantics. (2018). *AX semantics*. https://www.ax-semantics.com/de/home.html

Baesens, B., Bapna, R., Marsden, J., & Vanthienen, J. (2016). Transformational Issues of Big Data and Analytics in Networked Business. *Management Information Systems Quarterly*, *40*(4), 807–818.

Ballou, D. P., & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*, *31*(2), 150–162.

Ballou, D., Wang, R., Pazer, H., & Tayi, G. K. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, *44*(4), 462–484.

BARC. (2021). *Data, BI & Analytics Trend Monitor 2021*. BARC. http://barc-research.com/research/bi-trend-monitor/

Bean, R. (2017). How Big Data Is Empowering AI and Machine Learning at Scale. *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale/

Beinke, J. H., Nguyen, D., & Teuteberg, F. (2018). Towards a Business Model Taxonomy of Startups in the Finance Sector using Blockchain. *Proceedings of International Conference of Information Systems*, 9.

Chen, K., Chen, H., Conway, N., Hellerstein, J. M., & Parikh, T. S. (2010). USHER: Improving Data Quality with Dynamic Forms. *Proceedings of IEEE 26th*, 12.

Chiang, F., & Miller, R. J. (2008). Discovering data quality rules. *Proceedings of the VLDB Endowment*, *1*(1), 1166–1177.

Chiang, F., & Miller, R. J. (2011). A unified model for data and constraint repair. *2011 IEEE 27th International Conference on Data Engineering*, 446–457. http://ieeexplore.ieee.org/document/5767833/

Cleven, A., & Wortmann, F. (2010). Uncovering Four Strategies to Approach Master Data Management. *2010 43rd Hawaii International Conference on System Sciences*, 1–10. https://doi.org/10.1109/HICSS.2010.488

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

commercetools. (2018). *Commercetools | The eCommerce solution for innovators and visionaries*. Commercetools. https://techblog.commercetools.com/boosting-product-categorization-with-machine-learning-ad4dbd30b0e8

Dathena. (2019). *Dathena*. https://www.dathena.io/products/dathena-classify

DeepL. (2018). *DeepL*. https://www.deepl.com/home

Doan, A., Domingos, P., & Halevy, A. Y. (2001). Reconciling Schemas of Disparate Data Sources: A Machine-learning Approach. *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, 509–520. https://doi.org/10.1145/375663.375731

Doan, A., Suganthan, G. C. P., Zhang, H., Ardalan, A., Ballard, J., Das, S., Govind, Y., Konda, P., Li, H., Mudgal, S., & Paulson, E. (2017). Human-in-the-Loop Challenges for Entity Matching: A Midterm Report. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics - HILDA'17*, 1–6. https://doi.org/10.1145/3077257.3077268

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1–16. https://doi.org/10.1109/TKDE.2007.250581

Fan, W., Geerts, F., Lakshmanan, L. V. S., & Xiong, M. (2009). Discovering Conditional Functional Dependencies. *2009 IEEE 25th International Conference on Data Engineering*, 1231–1234. https://doi.org/10.1109/ICDE.2009.208

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Advances in Knowledge Discovery and Data Mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, Eds.; pp. 1–34). American Association for Artificial Intelligence. http://dl.acm.org/citation.cfm?id=257938.257942

Fernandez, R. C., Abedjan, Z., Koko, F., Yuan, G., Madden, S., & Stonebraker, M. (2018). Aurum: A Data Discovery System. *Proceedings of International Conference of Data Engineering*, 1001–1012.

Fernandez, R. C., Ilyas, I. F., Madden, S., Stonebraker, M. O. M., & Tang, N. (2018). Seeping Semantics: Linking Datasets using Word Embeddings for Data Discovery. *ICDE*, 12.

Grammarly. (2018). *Write your best with Grammarly.* https://www.grammarly.com/

Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, *35*(2), 388–423.

Heidari, A., Michalopoulos, G., Kushagra, S., Ilyas, I. F., & Rekatsinas, T. (2020). Record fusion: A learning approach. *ArXiv:2006.10208 [Cs, Stat]*. http://arxiv.org/abs/2006.10208

Hipp, J., Guntzer, U., & Grimmer, U. (2001). DATA QUALITY MINING. *DMKD*, 6.

Hu, M., Li, Z., Shen, Y., Liu, A., Liu, G., Zheng, K., & Zhao, L. (2017). *CNN-IETS: A CNN-based Probabilistic Approach for Information Extraction by Text Segmentation*. 1159–1168. https://doi.org/10.1145/3132847.3132962

Ilyas, I. F., & Chu, X. (2015). Trends in Cleaning Relational Data: Consistency and Deduplication. *Foundations and Trends® in Databases*, *5*(4), 281–393. https://doi.org/10.1561/1900000045

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., & Lee, D. (2003). A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery*, *7*, 81–99.

Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, *69*(2), 197–210. https://doi.org/10.1016/j.datak.2009.10.003

Lavrakas, P. (2008). *Encyclopedia of Survey Research Methods*. Sage Publications, Inc. https://doi.org/10.4135/9781412963947

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Legner, C., Pentek, T., & Otto, B. (2020). Accumulating Design Knowledge with Reference Models: Insights from 12 Years of Research on Data Management. *Journal of the Association for Information Systems*, *21*(3).

Li, G., Ooi, B. C., Feng, J., Wang, J., & Zhou, L. (2008). EASE: An effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. *Proceedings of the 2008*

ACM SIGMOD International Conference on Management of Data - SIGMOD '08, 903. https://doi.org/10.1145/1376616.1376706
Li, W.-S., Clifton, C., & Liu, S.-Y. (2000). Database Integration Using Neural Networks: Implementation and Experiences. *Knowledge and Information Systems*, *2*(1), 73–96. https://doi.org/10.1007/s101150050004
Liu, J., Li, J., Liu, C., & Chen, Y. (2012). Discover Dependencies from Data—A Review. *IEEE Transactions on Knowledge and Data Engineering*, *24*(2), 251–264. https://doi.org/10.1109/TKDE.2010.197
Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., & Raghavendra, V. (2018). Deep Learning for Entity Matching: A Design Space Exploration. *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18*, 19–34. https://doi.org/10.1145/3183713.3196926
Mukkala, L., Arvo, J., Lehtonen, T., & Knuutila, T. (2015). *Current state of ontology matching* (No. 4; University of Turku Technical Reports). University of Turku.
Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: Challenges and opportunities. *Proceedings of the VLDB Endowment*, *12*(12), 1986–1989. https://doi.org/10.14778/3352063.3352116
Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*(3), 336–359. https://doi.org/10.1057/ejis.2012.26
Otto, B. (2011). Data Governance. *Business & Information Systems Engineering*, *3*(4), 241–244.
Otto, B., & Österle, H. (2015). *Corporate Data Quality Prerequisite for Successful Business Models*. http://nbn-resolving.de/urn:nbn:de:101:1-2015112720186
Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162
Pingar. (2018). Automatic Classification and Categorization. *Pingar*. http://pingar.com/automatic-classification-categorization/
Püschel, L., Röglinger, M., & Schlott, H. (2016). What's in a Smart Thing? Development of a Multi-layer Taxonomy. *Proceedings of International Conference of Information Systems*, 19.
Pyle, D., & José, C. S. (2015). *An executive's guide to machine learning | McKinsey*. https://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning
Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, *10*(4), 334–350. https://doi.org/10.1007/s007780100057
Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3–13.
Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's Comments—Next-Generation Digital Platforms: Toward Human–AI Hybrids. *MIS Quarterly*, *43*(1), iii–ix.
Redman, T. C. (2017). Seizing Opportunity in Data Quality. *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/
Reltio. (2018). *Machine Learning Delivers Quality Data at the Speed of the Business*. Reltio Cloud - Data-Driven Applications - Master Data Management - Big Data. https://www.reltio.com/about/news/2017/1/machine-learning-delivers-quality-data-at-the-speed-of-the-business
Sarawagi, S., & Bhamidipaty, A. (2002). Interactive Deduplication using Active Learning. *Proceedings of SIGKDD*, 10.
Sarawagi, S., Deshmukh, K., & Borkar, V. (2001). Automatic segmentation of text into structured records. *ACM SIGMOD*, 12.
Shvaiko, P., & Euzenat, J. (2005). A Survey of Schema-Based Matching Approaches. In S. Spaccapietra (Ed.), *Journal on Data Semantics IV* (Vol. 3730, pp. 146–171). Springer Berlin Heidelberg. https://doi.org/10.1007/11603412_5

Shvaiko, P., & Euzenat, J. (2013). Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering*, *25*(1), 158–176. https://doi.org/10.1109/TKDE.2011.253

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A Survey of Data Provenance in e-Science. *SIGMOD Rec.*, *34*(3), 31–36. https://doi.org/10.1145/1084805.1084812

Smith, H. A., & McKeen, J. D. (2008). Developments in Practice XXX: Master Data Management: Salvation Or Snake Oil? *CAIS*, *23*(4).

Stonebraker, M., Bruckner, D., & Ilyas, I. F. (2013). Data Curation at Scale: The Data Tamer System. *CIDR Proceedings 2013*. Conference on Innovative Data Systems Research.

Stonebraker, M., & Ilyas, I. F. (2018). Data Integration: The Current Status and the Way Forward. *IEEE Technical Committee on Data Engineering*, 7.

Strong, D., Lee, Y., & Wang, R. (1997). Data Quality in Context. *Communications of the ACM*, *40*(5).

Thirumuruganathan, S., Tang, N., Ouzzani, M., & Doan, A. (2018). Data Curation with Deep Learning [Vision]. *ArXiv:1803.01384 [Cs]*. http://arxiv.org/abs/1803.01384

Toda, G. A., Cortez, E., da Silva, A. S., & de Moura, E. (2010). A probabilistic approach for automatically filling form-based web interfaces. *Proceedings of the VLDB Endowment*, *4*(3), 151–160. https://doi.org/10.14778/1929861.1929862

Voigt, P., & von dem Bussche, A. (2017). Organisational Requirements. In P. Voigt & A. von dem Bussche (Eds.), *The EU General Data Protection Regulation (GDPR): A Practical Guide* (pp. 31–86). Springer International Publishing. https://doi.org/10.1007/978-3-319-57959-7_3

Volkovs, M., Fei Chiang, Szlichta, J., & Miller, R. J. (2014). Continuous data cleaning. *2014 IEEE 30th International Conference on Data Engineering*, 244–255. https://doi.org/10.1109/ICDE.2014.6816655

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), 58–65.

Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, *7*(4), 623–640. https://doi.org/10.1109/69.404034

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Wu, R., Chaba, S., Sawlani, S., Chu, X., & Thirumuruganathan, S. (2020). ZeroER: Entity Resolution using Zero Labeled Examples. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 1149–1164. https://doi.org/10.1145/3318464.3389743

Wu, R., Zhang, A., Ilyas, I. F., & Rekatsinas, T. (2020). Attention-based Learning for Missing Data Imputation in HoloClean. *Proceedings of the 3rd MLSys Conference*, 19.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv:1609.08144 [Cs]*. http://arxiv.org/abs/1609.08144

Yakout, M., Elmagarmid, A. K., Neville, J., Ouzzani, M., & Ilyas, I. F. (2011). Guided data repair. *Proceedings of the VLDB Endowment*, *4*(5), 279–289. https://doi.org/10.14778/1952376.1952378

Zhu, H., Madnick, S., Lee, Y., & Wang, R. (2014). Data and Information Quality Research: Its Evolution and Future. In H. Topi & A. Tucker (Eds.), *Computing Handbook, Third Edition* (pp. 16-1-16–20). Chapman and Hall/CRC. https://doi.org/10.1201/b16768-20

8 Appendix

ID	Archetype	Source	Short description	ML application	Reference
1	Support manual data entry	Research	Prefill form from text	Bayesian network	Toda et al. (2010)
2		Tool	Grammarly: text correction	Natural language processing	www.grammarly.com
3		Expert	Prefill forms for master data entry	Probabilistic network	Expert 1
4		Research	Prefill form for faster entry	Bayesian network	Ali and Meek (2009)
5		Research	Adapt form to improve DQ	Bayesian network	Chen et al. (2010)
6		Tool/ Service	Association rules for form filling	A priori and FPgrowth	Camelot
7		Research	Embeddings for table population and retrieval	Word embedding and classification	Deng at al (2019)
8	Automated transformation of data	Tool	DeepL: text translation	Natural language processing	www.deepl.com
9		Research	Information extraction by text segmentation	Convolutional neural network	Hu et al. (2017)
10		Expert	Search by product picture	Convolutional neural network	Expert 1
11		Tool	Flixstock: generate catalog pictures	Convolutional neural network	www.flixstock.com
12		Expert	Detect material data from pictures	Convolutional neural network, classification	Expert 2
13		Research	Transform text in records	Hidden Markov model	Sarawagi et al. (2001)
14		Research	Google: text translation	Long short-term memory	Wu et al. (2016)
15		Tool	AX semantics: generate product descriptions	Natural language generation	www.ax-semantics.com
16		Research	SVM for metadata extraction	Support vector machines	Fox et al (2003)
17		Tool	Structuring: extract master from pictures or flat text files (Structuring)	Hidden Markov model	Camelot
18		Research	Structuring: extract master from pictures or flat text files (Structuring)	Active learning	Muslea et al (1999)
19		Research	Image tagging	Convolutional neural network	Zalando Research
20	Support data enrichment	Expert	Assignment of commodity code	Logistic regression	Expert 1
21		Tool	Commerce tools: categorize product descriptions	Convolutional neural network	www.techblog.commercetools.com
22		Tool	Reltio: machine learning assisted data enrichment	Classification, regression	www.reltio.com
23		Expert	Assignment of tariff code	Random forest	Expert 3
24		Research	Predict accurate size for customers	Hierarchical Bayesian	Zalando Research

ID	Arche-type	Source	Short description	ML application	Reference
25	Support data cleaning	Tool	FirstEigen: Big Data validation	Outlier detection	www.firsteigen.com
26		Research	Continuous data cleaning	Active learning/logistic classifier	Volkovs et al. (2014)
27		Research	Guided data repair	Active learning	Yakout et al. (2011)
28		Research	Outlier detection with statistical inference	Multivariate Gaussian mixture	Madden et al (2016)
29		Expert	Automated DQ assurance with autoencoders	Autoencoder	InCube
30		Tool/ Library	Holistic data repairs with probabilistic inference	Various	HoloClean
31		Research	Few-shot learning for error detection	Neural network	HoloDetect
32		Research	Anomaly detection using deep autoencoders predicting missing data values with deep learning	Attention-based autoencoder network	Wu et al. 2020
33		Research	Anomaly detection using deep autoencoders	Autoencoder	CERN
34	Support data matching	Tool	Talend: data matching	Classification	www.talend.com
35		Tool	Reltio: data matching	Active learning	www.reltio.com
36		Expert	Product master data harmonization	Classification	Expert 4
37		Research	Interactive deduplication	Active learning	Sarawagi and Bhamidipaty (2002)
38		Research	Entity matching	Recurrent neural network	Mudgal et al. (2018)
39		Research	Unsupervised learning for entity resolution	Gaussian mixture model	Wu et al. (2020)
40		Research	Supervised learning for “golden record” creation	Stagewise additive model	Heidari et al. (2020)
41		Tool/ Library	Python package for performing entity and text matching	Recurrent neural network	Deep Matcher
42		Tool/ Library	Deduplicate and find matches	Hierarchical clustering	Dedupe.io
43	Support data integration	Research	Tamr: data integration	Clustering/classification	Stonebraker et al (2013)
44		Research	LSD: schema matching	Classification/ensemble	Doan et al. (2001)
45		Research	Foreign key discovery with ML	Classification	Leser et al (2009)
46		Research	Ontology mapping	Classification/ensemble	Halevy et al (2002)
47		Research	Database integration	Self-organizing map	Liu et al (2000)
48	Automatic derivation data quality rules	Research	DQ rules mining	Association rules mining	Hipp et al. (2001)
49		Expert	DQ rules mining	Association rules mining	Expert 5
50		Expert	DQ rules mining and clustering	Association rules mining/clustering	Expert 1
51		Expert	DQ rules mining	Association rules mining	Expert 6

ID	Arche-type	Source	Short description	ML application	Reference
52	Automatic detection of sensitive and out-of-date data across systems	Tool	Abby: document classification	Natural language processing	www.abbyy.com
53		Expert	Predict retirement of data	Classification	Expert 5
54		Tool	Dathena: classify on level of confidentiality	Classification	www.dathena.io
55		Tool	Amazon Macie: discover and classify sensitive data	Classification	www.aws.amazon.com/macie/
56		Tool	Pingar: discover and classify sensitive data	Natural language processing	www.pingar.com
57	Support the discovery of relevant data	Research	Semantic linking of datasets	Pre-trained word embeddings	Fernandez et al. (2018)
58		Expert	Forecast processing time of a master data change	Regression	Expert 1
59		Tool	Alation: recommend data to join	Recommender engine	www.alation.com
60		Research	Find articles based on image	Convolutional neural network	Zalando Research

All Hands on Data:
A Reference Model for Enterprise Data Catalogs

Martin Fadler, Clément Labadie, Markus Eurich, and Christine Legner
Faculty of Business and Economics (HEC), University of Lausanne

Version for submission to an IS journal

Abstract: As data evolves into an important asset, companies are looking to meet the increasing demand for data inside the organization. In this context, data democratization can play a critical role in making data more broadly available to employees. However, research has not yet addressed the means and, specifically, the platforms that support data democratization. Our study addresses this gap by focusing on enterprise data catalogs (EDCs) as an emerging platform that serves as a data inventory and helps technical and business professionals find, access, and use data. Although the idea is intuitive and intriguing, EDCs lack a sound academic conceptualization, and their scope and role in future IT landscapes have yet to be fully understood. Following a design science research approach, this study develops an EDC reference model that outlines the key components of three architecture views: organization, data documentation, and function. We find that EDCs extend beyond metadata management concepts (e.g., data dictionaries and business glossaries) and provide rich functional capabilities (e.g., data discovery, data governance) to facilitate data democratization. From an academic perspective, our study provides a grounded definition of EDCs and outlines their key constituents as a cornerstone of the emerging enterprise data and analytics platforms. Practitioners can use the reference model to scope, assess, and select suitable EDC solutions and guide their implementation.

Keywords: Enterprise Data Catalog, Metadata Management, Data Curation, Data Management, Data Discovery, Reference Model

Table of Contents

List of Figures

List of Tables

1 Introduction

Data is at the core of emerging business models and has become one of the cornerstones of decision-making and business processes (Dallemulle and Davenport 2017; George et al. 2014; Wixom and Ross 2017). However, the more companies invest in building up data lakes and rolling out analytics infrastructures, the more the availability of and access to enterprise data is becoming an obstacle. It has been widely discussed that data scientists often spend more than 80% of their time searching and preparing data (Bowne-Anderson 2018); and many challenges arise because interrelated enterprise data is distributed over multiple databases and remains in operational silos (Hai et al. 2016; Halevy et al. 2016; Roszkiewicz 2010). To overcome these issues, companies need to efficiently allocate data supply activities and align them with the increasing demand for data.

Data democratization is referred to as a concept of making data more broadly available to employees (Awasthi and George 2020, p.1) and, thereby, addressing the data demand from extended user communities (Díaz et al. 2018; Hyun et al. 2020; Upadhyay and Kumar 2020). Prior research has mainly emphasized data democratization as a prerequisite to leverage data's business potential (Zeng and Glaister 2018) but has not yet addressed the means and, specifically, the platforms that support data democratization. One of these emerging platforms is enterprise data catalogs (EDCs), which serve as a unified data inventory and support technical as well as business professionals in finding, accessing, and using data. EDCs are an integral component of future enterprise IT landscapes (Belissent et al. 2019), and companies are increasingly turning to data catalogs to make their data FAIR (findable, accessible, interoperable, and reusable – Labadie et al. 2020). Yet, while the idea of having a central catalog for enterprise data seems intuitive, its conceptualization and implementation are not. From an academic perspective, the term "enterprise data catalog" is not well defined and to date has been neither conceptualized nor related to prior concepts and enterprise applications. From a practical perspective, companies have varying scopes and goals ranging from pure metadata management to business glossaries and full-fledged data integration and collaboration platforms. This is also reflected by the dynamics of the EDC market, where the scope of EDC functionalities varies among solutions from different vendors (Goetz et al. 2020; Sallam et al. 2020; Zaidi et al. 2017). Hence, making sense of the EDC concept can open up new interesting research opportunities while providing insights into the means for democratizing data in enterprises.

To address this research gap, we ask the following research question:

RQ: *What are the main components of an enterprise data catalog as an emerging platform for data democratization?*

The goal of our study is to propose a reference model (Frank 2014) that synthesizes the key constituents of an EDC and lays the foundation for understanding an EDC's role as a platform for data democratization in enterprises. As a specific type of conceptual model (Frank et al. 2014; Vom Brocke 2007), reference models are commonly used in research and industry to design and plan complex systems while fostering communication with prospective users and providing a sound basis for system implementation (Frank 1999, p. 695). Following the guidelines of design science research (Peffers et al. 2007), we built a reference model for EDCs while developing close industry–research collaboration over 18 months. The resulting reference model is grounded in prior academic research on platforms supporting data democratization, such as digital libraries (Borgman 2003) and dataspaces (Franklin et al. 2005), and integrates insights from focus groups and ongoing analysis of current EDC solutions and implementations. As a multilayered reference model, it synthesizes EDC's key components and organizes them into three views: organization, function, and data. The organization view consists of eight data-related roles that reflect the increasing number of business users and technical experts working with data inside an organization. The function view defines nine function groups with corresponding sub-functions that support data demand and supply. The data view identifies 22 metadata objects that guide the documentation of data for technical and non-technical user roles.

Using this reference model, we characterize the EDC as an evolutionary metadata management concept (Roszkiewicz 2010; Sen 2004) that integrates existing approaches (e.g., business glossaries or data dictionaries) and provides rich functional capabilities to facilitate data democratization (e.g., data governance or data discovery). The EDC reference model contributes to both research and practice. From an academic perspective, we conceptualize EDCs through their key components organized into three architectural views. Thus, our findings inform research in the field of data management (Legner et al. 2020) and complement studies on enterprise analytics platforms and big data infrastructures (Hyun et al. 2020; Fadler and Legner et al. 2020). Practitioners can use the EDC reference model to understand the scope and characteristics of EDCs, assess and select a suitable solution, and guide the implementation.

The remainder of this article is structured as follows. In the background section, we elaborate on prior concepts that address similar but complementary ideas to EDCs. We then present our research design and process in detail. Next, we elaborate on the considerations underlying the reference model's development and its main components: an organization view, a function view,

and a data view. To demonstrate its applicability, we use the reference model to classify 15 vendor solutions and derive two archetypes based on this assessment. We conclude with a discussion and the future outlook of our research.

2 Background: Platforms for Data Democratization

An EDC supports companies looking to democratize their data. Using prior research, we identify two concepts that facilitate data democratization and pursue goals similar to those of an EDC: First, the digital library (DL) focuses on making digital scholarly material, such as textual content and research data, accessible to the research communities (Wilcox 2018). Second, the dataspace (DS) describes the technical infrastructure for making interrelated data findable and accessible across distributed databases (Franklin et al. 2005). Both approaches establish a fundamental understanding of platforms that support data democratization and develop architecture considerations that can be applied to EDCs.

2.1 Digital Library

Libraries have always played an important role in democratizing information for a large audience (Wallace and Van Fleet 2005). Today, the digital library (DL) has become a central component of knowledge infrastructure (Borgman et al. 2015) and is considered one of the most complex information systems (Fox and Sornil 2003). The concept was first formulated with Licklider's (1965) vision of the library, where he raised concerns about the limitations of preserving printed material in physical libraries. With the advent of the Internet at the beginning of the 1990s and the surge in scholarly material, the number of DLs surged. Borgman's (2003) influential definition of a DL comprises two parts: *"1. Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching and using information. […] The content of digital libraries includes data, metadata that describe various aspects of the data (e.g., representation, creator, owner, reproduction rights), and metadata that consist of links or relationships to other data or metadata, whether internal or external to the digital library. 2. Digital libraries are constructed, collected and organized – by [and for] a community of users, and their functional capabilities support the information needs and uses of that community"* (Borgman 2003, p. 42). Early DL architecture blueprints like the Fedora architecture, which was originally developed by the Digital Library Research Group at Cornell University, is still maintained today (Staples et al. 2003). Another example is the Kahn–Wilensky architecture (Kahn and Wilensky 1995), which gained a significant amount of attention and encompasses four different types of components (Calhoun 2014): First, repositories, file systems, and distributed storage systems;

second, search functionalities enabled through indexing or metadata; third, an identifier system for digital objects; fourth, user interfaces for user services for browsing, visualizing, or delivering the content. Further components and parts of other DL architectures include user authentication and collaboration support (Calhoun 2014). Because of the steadily growing amount of digital content, the World Wide Web has also been considered a DL. This has led to ambitious initiatives like the Stanford Integrated Digital Library project to *"develop the enabling technologies for a single, integrated and 'universal' library, providing uniform access to the large number of emerging networked information sources and collections. These include both online versions of pre-existing works and new works and media of all kinds that will be available on the globally interlinked computer networks of the future"* (Stanford 1999). Members of this project included Sergey Brin and Larry Page, who in 1998 presented their work on the Pagerank algorithm to efficiently crawl and index the web, which ultimately became the starting point for their company, Google. While DLs initially had a major focus on managing textual content, their scope has expanded to manage multimedia resources and research data as well. In research communities, DLs are important *"for purposes of reuse, verification, or reproducibility"* of publications and data (Borgman et al. 2015, p.5). They play a key role in making data FAIR (i.e., findable, accessible, interoperable, and reusable) for humans and machines (Wilkinson et al. 2016) and help democratize data within research communities (Wallace and Van Fleet 2005; Wilcox 2018).

2.2 Dataspace

In database research, Franklin et al. (2005) suggest the dataspace (DS) concept as a reference architecture for finding interrelated data distributed over multiple databases. DSs "*provide base functionality over all data sources, regardless of how integrated they are*" (Franklin et al. 2005, p.2). The DataSpace Support Platform (DSSP) comprises five components: *catalog and browse*, *search and query*, *local store and index*, *discovery,* and *source extension*. The *catalog* serves as "*an inventory of data resources, with the most basic information about each, such as source, name, location in source, size, creation date and owner, and so forth. The catalog is infrastructure for most of the other dataspace services, but can also support a basic browse interface across the dataspace for users*" (Franklin et al. 2005, p. 29). With *search and query*, a DSSP provides different services to find the relevant data. Here, either data or metadata can be queried. Additionally, a service to monitor data could be implemented. With a *local store and index* structure, data can be efficiently found and retrieved. *Discovery* ensures that data objects can be located in the DS and relationships can be tightened either by the user or semi-automatically. With *source*

extension, a DS should be capable of extending data sources with value-added information that is not held directly by the data source but within the DS. Examples of value-added information could be classifications, ratings, or annotations. Based on this reference architecture, the database community has developed various DSSPs. For instance, Google proposes a catalog (named GOODS) that manages the metadata of datasets distributed over heterogenous systems and provides services to users to find relevant datasets more quickly (Halevy et al. 2016). Hellerstein et al. (2017) argue that the changing requirements for data management with regard to data exploration and innovation call for new approaches to metadata management. They present Ground, a data context service, as *"a system to manage all the information that informs the use of data"* (Hellerstein et al. 2017, p.1). While these systems can support data democratization in companies, they focus on technical architectures and services, but neither of them explores their integration into enterprise IT landscapes nor elaborates on potential use-case scenarios in an enterprise setting.

2.3 Research Gap

To the best of our knowledge, the EDC concept is mainly discussed among practitioners (Russom 2017; Zaidi et al. 2017), and a rigorous definition and conceptualization are lacking.

Drawing on our literature review of the concepts of DS and DL, we isolate three essential components that can be translated into EDCs (see Table 40). First, both DS and DL contain metadata in their inventory of data resources. According to Borgman (2003, p. 42), metadata should describe various aspects of the data (e.g., representation, creator, owner, reproduction rights), as well as links or relationships to other data or metadata. Halevy et al. (2016) specify metadata groups and metadata for Google's DS system, such as the *Content-based* (schema, number of records, similar datasets) or *User-supplied* (description, annotations) metadata groups (Halevy et al. 2016). Second, DLs "are constructed, collected and organized – by [and for] a community of users, and their functional capabilities support the information needs and uses of that community" (Borgman 2003, p. 42). Similarly, EDCs support the "needs and uses" of different enterprise roles and comprise both data experts and non-experts. A clarification of these roles is also needed in the context of EDCs to understand their requirements in terms of data access and use. The third component is functions. Both DL and DS comprise, on the one hand, functions to store, index, and catalog data and, on the other hand, user functions to create, search, browse, discover, and use data (Borgman 2003; Franklin et al. 2005).

Table 40. Platforms for data democratization

	Digital library (DL)	**Dataspace (DS)**	**Enterprise data catalog (EDC)**
Authors	(Borgman 2003; Calhoun 2014; Fox and Sornil 2003)	(Franklin et al. 2005; Halevy et al. 2016; Hellerstein et al. 2017)	No academic definition yet; here: translation of DL and DS concepts, informed by practitioner literature on EDCs (Russom 2017; Zaidi et al. 2017)
Purpose	To provide access to large numbers of academic information sources	To find interrelated data across distributed databases	To facilitate data democratization in companies
Content	Textual content, multimedia content, research data, Metadata (structural, administrative, terminological)	Datasets Metadata (structural, administrative, terminological, use)	Enterprise data Metadata (structural, administrative, terminological, governance, context, use)
Functions	Storage, object identification, search	Catalog, object identification, search, discover	Not clearly defined but represent an evolution of data dictionaries, business glossaries, and metadata repositories
Users	Communities of users, mainly from education/research	Not clearly defined: organizations on various levels (e.g., enterprises, government agencies, libraries, "smart" homes)	Communities of users in the enterprise (data experts and data non-experts)
Examples	Stanford Integrated Digital Library (Stanford 1999) Fedora (Staples et al. 2003)	Google Dataset Search (GOODS) (Halevy et al. 2016) International DataSpace (IDS) (Otto et al. 2019)	Enterprise data catalog solutions

In the enterprise context, these topics have been addressed, in part, by various metadata concepts, such as data dictionaries, business glossaries, and metadata repositories, albeit with a narrower scope. Data dictionaries provide data documentation at the database level, i.e., basic documentation of tables and fields (Uhrowczik 1973), specifically catering to the needs of technical users. At the other end of the spectrum, business glossaries document key terms in a way that business users can understand. Metadata repositories enable data documentation on an abstraction layer, linking multiple storage instances of data (Chaki 2015), as direct relationships between technical and business terms are impractical and non-scalable in complex IT landscapes (Kumpati 1988). Yet, these concepts are not integrated and only address restricted user and functional scopes if compared with DLs and DSs. Extensions in both areas are essential to data democratization and are addressed by EDCs.

3 Research Design

3.1 Research Objectives and Approach

Our goal with the research is to provide an understanding of the EDC concept as an emerging platform for data democratization by developing a reference model. Reference models are important artifacts that help accumulate design knowledge from academic and practitioner communities and have become very popular to guide data-related topics (Legner et al. 2020). A reference model is defined as "a normative construction (or artifact) created by a modeler who describes a system's universal elements and relationships as a recommendation, thus creating a center of reference" (Ahlemann and Riempp 2008, p.89). As a specific type of conceptual model (Frank et al. 2014; Vom Brocke 2007), reference models are commonly used in research and industry to design and plan complex systems while fostering communication with prospective users and providing a sound basis for system implementation (Frank 1999, p. 695). They are one approach to accelerate the development of enterprise-specific models (Fettke and Loos 2003, p. 35) and are, therefore, ideal to fulfill our research goals.

Reference models are usually developed in iterations of design and evaluation following design science principles (Winter and Schelp 2006). Since the emerging solutions (EDC) address a contemporary problem (data democratization) but have not been well defined in research and practice, we chose the design science research method outlined by Peffers et al. (2007) and the *Objective-Centered Solution* initiation (see Figure 8). Based on our review of prior literature, as well as existing EDC solutions, we designed the EDC reference model iteratively with frequent and early iterations with practitioners to reach an effective solution design and evaluation (Sonnenberg and vom Brocke 2012). As generic and abstract design knowledge, the EDC reference model thereby explicates the (implicit) design knowledge that we derived from situational inquiry (i.e., insights from company-specific EDC initiatives) and materialized instantiations (i.e., EDC solutions and pilot implementations).

Throughout the research process, we gained insights into EDC evaluation and implementation projects by conducting focus groups and interviews with data management experts from 13 large international companies (see Table 41). The experts who joined the group were overseeing EDC initiatives or closely involved in key implementation aspects. Although they all shared the key objectives of democratizing data, they were looking at the issue from various angles and with different priorities: Some of the participants' main interests were data supply, with metadata management and data governance, while others aimed to lower the barriers for data

consumption and, specifically, for analytics purposes. In addition to our insights from focus groups and interviews, we observed or participated in EDC implementation projects in five companies and continuously monitored and analyzed the market for EDC solutions. To complement our practical insights, we continuously reviewed the academic and practitioner literature on data democratization and EDCs.

Table 41. EDC projects of participating companies

Company	Industry	Revenue range	Purpose	Status
A	Adhesives	€1B to €50B	Metadata management	Rollout and onboarding
B	Pharmaceuticals	€1B to €50B	Support for data analytics	Implementation in progress
D	Chemistry	€50B to €100B	Support for data governance and data analytics	Rollout and onboarding
C	Sportswear	€1B to €50B	Support for data analytics	Rollout and onboarding
E	Manufacturing	€1B to €50B	Metadata management	Rollout and onboarding
F	Pharmaceuticals	€1B to €50B	Support for data governance and data analytics	Rollout and onboarding
G	Manufacturing	€50B to €100B	Metadata management (register, search & retrieve data)	Pilot
H	Automation	€1B to €50B	Support for data governance	Tool selection
I	Retail	>€100B	Support for data governance	Continuous usage and maintenance
J	Tobacco	€50B to €100B	Support for data governance	Continuous usage and maintenance
K	Information technology	€1B to €50B	Support for data governance and data analytics, metadata management	Continuous usage and maintenance
L	Fashion and jewelry	€1B to €50B	Data glossary	Rollout and onboarding
M	Packaging	€1B to €50B	Support for data governance, analytics, inventory, and automation	Scoping and tool selection

3.2 Iterations

Following the steps outlined by Peffers et al. (2007), we developed the EDC reference model in three major iterations over 18 months, each comprising a design and evaluation step. As the model reached a stable state with version 1.0, we also included demonstration steps.

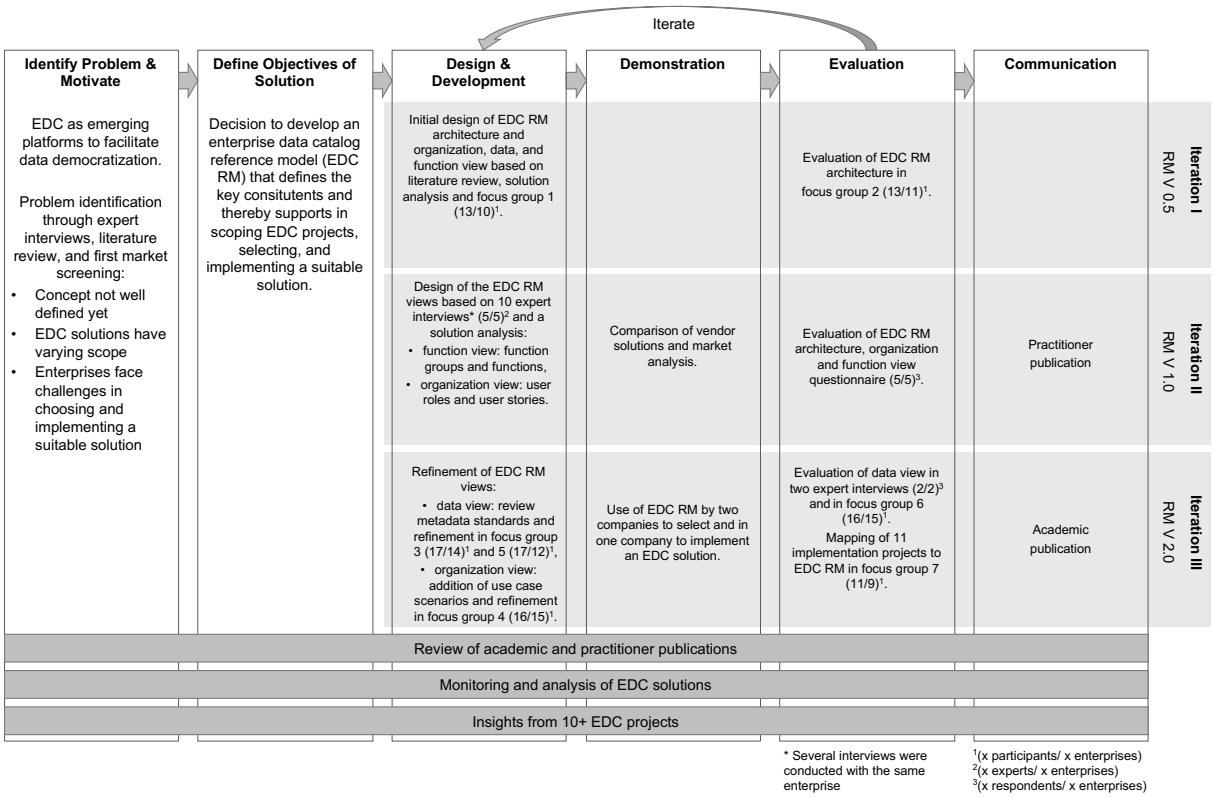


Figure 8. Research process

Iteration I – Reference model Version 0.5 (January 2018–June 2018): We designed the initial version of the EDC reference model (Version 0.5) based on three inputs: The literature review on related concepts (DL and DS) informed us about essential architecture components; from the first analysis of selected EDC solutions, we gained insights into the functional scope of EDCs; and focus group 1 helped us identify typical users. We translated these insights into a multilayered reference model (Frank 2014) with three views: an organization view, which outlines eight user roles and user stories; a function view, which specifies three function groups and functions; and a data view, which defines metadata objects and attributes. This version of the reference model was evaluated by a focus group of 13 data management experts from 11 companies. The participants assessed the general structure and confirmed its usability for their own EDC projects. Major points of improvement were emphasized in the function view, which was found to be too coarse. While the eight user roles in the organization view received general agreement, the user stories that were used as examples were not yet representative enough to satisfy the companies' own requirements.

Iteration II – Reference model Version 1.0 (July 2018–September 2018): In the design and development step of the second iteration, we enhanced the EDC RM (Version 1.0) – primarily the function view – based on feedback from the previous iteration. To this end, we conducted a series of expert interviews (one or more interviews per expert; 10 interviews in total), as well as

a detailed analysis of EDC solutions on the market and gained insights on EDC requirements for the selection and implementation of five EDC projects. As part of the market analysis, we first scanned analyst reports and considered a broader range of solutions, including tools for metadata management, data governance, and data lake management (De Simoni, Dayley, et al. 2018; De Simoni, White, et al. 2018; Duncan et al. 2016; Goetz et al. 2018; Peyret et al. 2017; Zaidi et al. 2017). The initial list was expanded by online searches for further tools and by insights from interviews with practitioners. From about 100 identified solutions, we filtered out 15 that are in line with companies' priorities and understanding of an EDC (see Table 43). Based on a detailed analysis of openly available information material, analyst reports, and documents from companies considering implementing an EDC solution, we specified eight function groups with their own distinct functions. As a demonstration step, we mapped the 15 selected solutions onto the correct function groups and assessed the extent to which the related functionalities were covered. These developments were communicated through a practitioner publication.

In the evaluation step, the reference model version was assessed and specified by means of the individual interviews with five EDC project managers and through a semi-structured questionnaire based on the evaluation criteria proposed by Prat et al. (2015). Respondents were asked to rate the relevance of user roles and their example user stories, as well as the function groups and their functions. We captured the answers by using a five-point Likert scale (*strongly disagree*, *disagree*, *uncertain*, *agree*, *strongly agree*). For the user roles and user stories, all respondents answered *agree* or *strongly agree* concerning the relevancy for their company. For most of the function groups and functions, the respondents answered *agree* or *strongly agree* concerning the relevancy for their company. Only for the function groups data assessment and data analytics did the respondents mostly respond with *uncertain* or *agree*. We also asked the respondents to rate whether the reference model is complete, easy to understand, and useful for their company. Overall, the respondents agreed that the reference model is easy to understand and useful. However, a few were uncertain whether the reference model was already complete, and we included their feedback in the next design iteration.

Iteration III – Reference model Version 2.0 (September 2018–November 2019): Based on the expert feedback collected in the second iteration, a minor change was made to the function view, where we separated one function group into two. After we integrated this change, we further refined the organization view by deriving EDC use-case scenarios that establish links between the function and organization views. The use-case scenarios were outlined through a template with instructions and subsequently discussed and completed in focus group 4. The user roles within the use-case scenarios as proposed by the participants could all be mapped with our

organization view. The practitioners were asked to add function groups, but they could not find any missing. This insight means that all use-case scenarios could be accurately described using the organization and function views. At this point, the focus group reached a consensus that the organizational and function views had together reached a stable state.

In parallel, we resumed the development of the data view. To anchor it in existing knowledge, we started by reviewing domain-agnostic metadata standards, of which we identified 14. After excluding those solely specifying data formats or technical interchange and encoding schemes, we retained four standards as relevant for EDCs: the Dublin Core Schema (DC) (Dublin Core Metadata Initiative n.d.), the Data Catalog Vocabulary (DCAT) (World Wide Web Consortium (W3C) n.d.), the Common Warehouse Metamodel (CWM) (Poole et al. 2002), and the ISO 11179-3 Metadata Registry Metamodel and Basic Attributes (MDR) (International Organization for Standards / International Electrotechnical Commission (ISO/IEC) 2013). Based on these insights, we designed an EDC metadata model, which we iterated internally and in focus groups 3 and 5 to attain a stable version. This version was further refined through expert interviews with representatives from two external organizations, who had experience developing similar models in the context of EDC implementation projects. We integrated the experts' feedback and subsequently evaluated the metadata model with our broader participant sample in focus group 6.

As part of our evaluation activities for the overall EDC reference model, we analyzed and compared 11 EDC implementation projects by asking representatives from organizations to map them onto the reference model. Thanks to focus group 7, we found that the EDC reference model was extensive enough to categorize and support EDC implementation projects. This was confirmed in demonstration steps, where two organizations (company I from our participant sample (see Table 41), as well as an external organization active in the energy industry) used the reference model (particularly the organization and function views) during the request for proposal (RfP) meetings with EDC vendors to compare offerings and select an EDC solution. Furthermore, company B (Table 41) relied on the EDC reference model to guide its overall implementation initiative. Finally, this publication is part of the communication step.

4 EDC Reference Model

In line with Frank (2014), the EDC reference model comprises multiple levels: the reference model architecture as the first level "*to decompose the overall problem domain into smaller manageable units and provide a high-level overview of the reference model*" (Ahlemann and Riempp 2008, p. 92) and three views as the second level to deconstruct in multiple domain-specific layers. We constructed the EDC reference model architecture based on a synthesis of related DL and DS components (see Section 2.3) and the prevailing IS architecture conceptualizations (Chang et al. 2007; Scheer 2001; Scheer and Schneider 2006). The reference model architecture distinguishes three views (organization, function, and data) and how they relate to each other (see Figure 9). In the second level, we deconstruct each view into its key constituent parts.

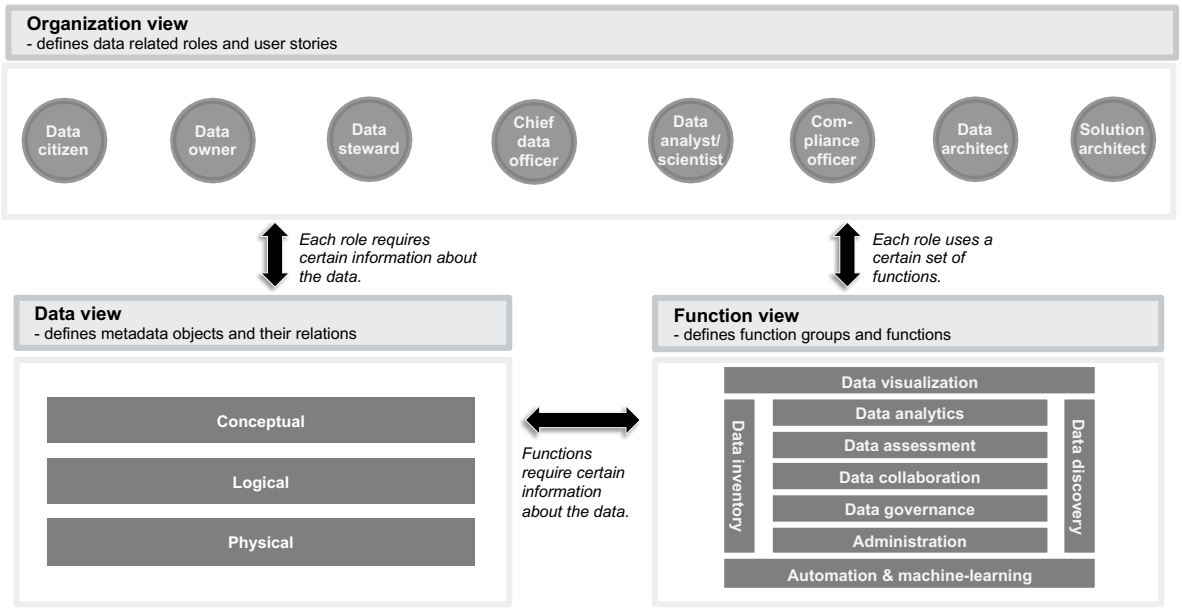


Figure 9. EDC reference model architecture

4.1 Organization View

Following prior literature (section 2.3), platforms for data democratization should address the needs of a certain user community. In the enterprise context, this community consists of employees with varying levels of data-related expertise and expectations. Based on the literature, expert input, and insights from implementation projects, we identified eight user roles for EDCs together with example user stories (see Table 42).

Table 42. Organization view: data roles and user stories

User roles	User stories	Related function groups and functions	Related metadata objects
Data citizen	Understand how to correctly enter data into a system Understand how to interpret data in a report Find the right data for a specific task (e.g., report creation) and identify trusted sources Provide feedback on data (e.g., leave a comment regarding a data error) Identify the right person(s) to contact for data-related questions	Data analytics: *documentation/data stories* Data collaboration: *following/updates, user communication rating, commenting* Data inventory: *business glossary* Data discovery: *search, recommendation, data subscription* Data governance: *rules and policies* Data visualization: *drill-down (process/report on data)*	Business term Business object Business object attribute Application Transformation Report
Data owner	Register data under ownership Maintain definitions and value domains (lists), incl. validation and approval processes Provide metadata on data (e.g., about data quality) Grant access to data under ownership and share guidelines & definitions Compare default and real-life values in systems Access usage data regarding data under ownership	Data inventory: *data registration, business glossary, data dictionary, data access* Data collaboration: *sharing* Data governance: *workflows, roles & responsibilities* Data assessment: *data quality*	Business term Business object Business object attribute Data object Data object attribute Value domain
Data steward	Assess data in the area of responsibility (e.g., quality, maturity, usage) Analyze dependencies between data elements (e.g., business objects, attributes) Investigate data issues and identify faulty data element(s) in process failures (e.g., data quality root cause) Document data (metadata, e.g., quality, maturity)	Data inventory: *metadata management* Data assessment: *data usage, data profiling, data quality* Data collaboration: *tagging, user communication* Data governance: *workflows, roles & responsibilities* Data visualization: *drill-down (process/report on data)*	Data domain Business object attribute Data object attribute Value domain Role Actor Board/council

User roles	User stories	Related function groups and functions	Related metadata objects
Chief data officer	Gain overview on data assets Classify assets according to specific criteria (e.g., quality, costs, usage, risk) Assign roles and tasks to data assets Create workflows for data governance	Data assessment: *data usage, data risk, data quality, data valuation, benchmarking* Data governance: *workflows, rules and policies, roles and responsibilities*	Business domain Data domain Business terms Role
Data analyst/scientist	Understand problem domain Explore and obtain relevant data for a given problem (starting from business meaning or technical field) Provide or retrieve documentation on analytics work with data Publish datasets, possibly with a data story of a successfully implemented analytics application Provide feedback on datasets (e.g., usability, quality)	Data assessment: *profiling* Data discovery: *search, recommendation, subscription, data delivery* Data analytics: *documentation/data stories, data application repository* Data collaboration: *tagging, rating, commenting, sharing, following/updates*	Business term Business object Business object attribute Application Transformation Report
Compliance officer (e.g., data protection officer)	Discover compliance-sensitive data and locate systems/ attributes Understand compliance issues in a specific dataset Label data (attributes) that need(s) to be protected Check who uses and has access to which data Prove the compliance of data usage	Data governance: *rules and policies, data authorizations, handling sensitive data* Data assessment: *data risk* Automation & machine learning (ML): *automated classification/tagging* Data inventory: *metadata management* Data discovery: *search*	Regulations & guidelines Data domain Business term Business process Business object Business object attribute Data object attribute System
Data architect	Manage data models (e.g., create, change, delete) Assess how data is used across systems Link business definitions to the physical layer (e.g., reports)	Data inventory: data lineage, metadata management, data dictionary, business glossary Automation & ML: *automated scanning/ ingestion* Data analytics: *data application repository*	Business object Business object attribute Data object Data object attribute Data structure

User roles	User stories	Related function groups and functions	Related metadata objects
Solution architect	Retrieve and update documentation on data Discover the data schema of a specific system Map data schemas between systems Understand compliance issues in a specific dataset Understand cross-system data lifecycle	Data inventory: *data lineage, data dictionary, metadata management, upload/link content* Data assessment: *data profiling* Data visualization: *data flow/network visualization* Automation & ML: *normalization/data similarity*	Data object Data object attribute System Data structure Interface Application

User roles revolve around the general purposes of an EDC to support data supply, demand, and curation (Borgmann 2003, Lord et al. 2004, p.1) and the three data-related role categories (Lee and Strong 2004): data collectors, data consumers, and data custodians. On the supply side, data collectors are responsible for collecting and inventorying data resources into an EDC. For instance, data architects and the solution architects who model, maintain, and create data to be referenced and documented within the EDC. From a curation perspective, data custodians work with data that has been integrated with the system thanks to data collectors and ensure that it is fit for use. For instance, data owners oversee a specific data domain and manage their creation and access, while data stewards use the EDC to assess and document various aspects of datasets (e.g., quality, maturity, usability), supporting data demand by maintaining relevant definitions. On the demand side, data consumers use data to support their specific business purposes. For instance, data citizens need to find data and understand data practices, and data analysts require precise data documentation to analyze them. As for chief data officers and chief compliance officers, they benefit from gaining an overview of data assets, as well as information on where (e.g., systems, business units), when (e.g., processes), and by whom data is used inside the company.

Each user role has specific data requirements and views data differently; for example, in his role as a data citizen, a marketing manager wants to understand how a certain forecast was calculated, whereas a data analyst requires domain knowledge about the specific data they are working with. We use user stories to describe how their specific tasks could be better executed by using an EDC. Each user story is associated with relevant function groups and metadata objects, which we will present in detail in the following sections.

The pairings of roles and user stories outline three key aspects of EDCs from a user perspective. First, an EDC is expected to put forward data assets and increase data transparency. Second, this

transparency should apply not only to datasets themselves but also to the way they are used (e.g., overseeing data flows across business units, business processes, applications & systems) and to potential issues related to usage (e.g., regulatory constraints, internal guidelines). The concept of lineage seems to apply to virtually all identified scenarios in that it is crucial to understanding data usage patterns and flows between systems to use it properly and identify potential issues and their root cause. Third, collaboration between users appears to be an important value driver for EDCs either in terms of exposing ownership and responsibilities or by enabling communication between various stakeholders (e.g., sharing/social or task management features).

4.2 Function View

As outlined in section 2.3, platforms for data democratization must contain functions for creating, searching, and using data and metadata, which is highlighted in particular in the DS-related literature. For EDCs, we build on these functions and used function trees (Scheer 2001, pp. 21-38) to structure the functions hierarchically in two layers of function groups and functions.

From our market analysis and implementation projects, we identify an EDC's functional scope as comprising functions to register data, retrieve and use data, and assess and analyze data. Hence, an EDC should provide a data inventory (for data supply) and a data discovery (for data demand) as basic function groups. Other function groups should support individual user roles in data governance, data assessment, data analytics, and administration alongside appropriate function groups for visualization, automation & ML, and data collaboration. In the following, we describe each of the function groups.

Table 43. Function view: function groups and functions

Function group	Description	Function
Data inventory	Helps to register and document data either manually or automatically.	Data registration Metadata management Business glossary Data dictionary Data provenance Data ingestion/crawling Upload/link content
Data discovery	Supports users in finding and obtaining data in a guided way.	Search Dataset recommendation Data access Data subscription Data delivery
Data analytics	Facilitates the work of data analysts and scientists with specific functionalities.	Data story Data application repository Data query Data lake monitoring
Data assessment	Supports in evaluating and measuring data according to specific metrics.	Data usage Data quality Data risk Data valuation Data profiling Data lineage
Data collaboration	Enhances the collaboration of user roles when maintaining, documenting, or using data.	Tagging Rating Following/updates Commenting Messaging/user chat Sharing
Data governance	Facilitates typical data governance procedures and task allocation.	Role & responsibility management Workflow Rule & policy Data access management
Data visualization	Helps in visualizing data and metadata based on different dimensions.	Graphs Diagrams In-table visualization Dashboards/cockpit Data flow/network visualization
Automation & machine learning	Supports either in automating or facilitating certain tasks of other function groups.	Automated scanning/ingestion Automated classification/tagging Normalization/data similarity Data unification Usage pattern analysis Recommendation
Administration	Helps in managing users and configuring the optimal use of the solution.	Configuration User management

With the **Data inventory** function group, data can be registered and documented either manually by user roles or automatically through an exchange with source systems. Thus, an EDC uses a pre-defined metadata model (see section 4.2) that describes data for both technical and non-technical user roles and makes it possible to normalize data descriptions across systems. An EDC combines metadata concepts such as business glossaries and data dictionaries to document data on all levels – in the form of conceptual, logical, and physical data models – to support technical and non-technical user roles alike. This allows an EDC to act as a data context service in data lake environments, where data is stored in various formats and types, to deliver a unified view on data (Hellerstein et al. 2017). For instance, with their Anzo® Smart Data Lake, Cambridge Semantics provides an EDC solution that uses a semantic graph model to document data and the data relations from a physical to a conceptual level.

With the **Data discovery** function group, users can find and obtain data in a guided way. DLs and DSs comprise search, browse, and discovery functionalities to find relevant data (Borgman 2003; Franklin et al. 2005). In the suggested DS system by Google, the usage of datasets is traced across systems and applications. This enables users to discover how datasets were used and have changed over time (Halevy et al. 2016). Similarly, an EDC's most basic functionality is a search function that uses metadata to match a user request with the related data resources. In a more advanced setup, a user role receives proactive recommendations for data based on her/his user profile and activity logs. In addition to the search functionalities, an EDC provides features for obtaining data. For instance, a user receives access permissions to obtain data by entering a data subscription while respecting access rights and data license conditions. The solutions on the market vary among their data discovery functionalities. One of the most advanced is the company Collibra's solution. Here, users can discover data by either searching or receiving a recommendation. Once a relevant dataset is found, a user requests access to the data. This process behaves in a way similar to a checkout in an e-commerce shop: The user adds their data of interest to a shopping cart, which triggers a workflow after checkout in which the corresponding data official grants or denies access to the requested data.

With the **Data analytics** function group, an EDC provides specific functionalities to support the work of data analysts and scientists. When connected to code repository solutions such as GitHub, these roles can maintain their analytics application repository with the used data. In the EDC, the functionalities to process data, as well as dataset characteristics, can be documented – e.g., a certain way to preprocess data or peculiarities about a dataset. This form of documentation not only enhances collaboration but also increases the efficiency of analytics projects by having direct access to reusable analytical data and avoiding certain pitfalls when

working with datasets. Once an analytics application has been successfully implemented in the organization, writing, and publishing (through an EDC), a data story on how data was used can inspire other teams and stimulate analytics use in other departments, for example. Data stories also help onboard employees more quickly because they describe analytics applications more comprehensively than a mere code repository. This function group also supports data-intensive research activities. The FAIR principles introduced earlier emphasize not only data but also "the algorithms, tools, and workflows that led to that data" (Wilkinson et al. 2016, p.1). This means that data (as input for not only the analytics application but also the analytics application and the workflows needed for its organizational implementation) should be findable, accessible, interoperable, and reusable. EDC solutions on the market vary in their support of analytics-oriented user roles. The company Alation's solution makes it possible to write queries and to delve into sample datasets within the tool itself. This gives data scientists an efficient way to discover relevant datasets while leveraging the advantages of a data lake environment.

With the **Data assessment** function group, functionalities are provided that help evaluate and measure data according to specific metrics like business value and quality. While a data profiling functionality provides generic descriptive statistics on datasets, other functionalities may enable more targeted assessments regarding their quality, risk, value, or use: "*value may be based on multiple attributes, including usage type and frequency, content, age, author, history, reputation, creation cost, revenue potential, security requirements, and legal importance*" (Short and Todd 2017, p. 18). Although data valuation is a rather new field of research, there is a rich body of knowledge on data quality assessment (Batini et al. 2009; Pipino et al. 2002; Wang and Strong 1996). Here, data quality can be assessed through quantitative and qualitative measures that can be supported through an EDC. For instance, in SAP's Data Hub & Information Steward, data quality can be assessed and monitored using dashboards and scorecards.

With the **Data collaboration** function group, user roles can collaborate when maintaining, documenting, or using data. Besides commenting, users can also collaborate by rating or tagging datasets. These are typical functionalities used to curate content in modern platform environments like Facebook but are also considered in DLs. DLs *"should be collaborative, allowing users to contribute knowledge to the library, either actively through annotations, reviews, and the like, or passively through their patterns of resource use"* (Lagoze et al. 2005). The solutions on the market provide varying functions for data collaboration. Zaloni's Data Management Platform provides a workspace feature where users can share their work results on data. Such a function enhances efficiency, especially in analytics projects where work is often performed in cross-functional teams. In Collibra's EDC solution, users can leave comments on datasets and

mention other users. With this functionality, the required data work is identified early in the process and directly assigned to an official. Thus, data quality issues can be resolved more quickly.

With the **Data governance** function group, an EDC facilitates typical data governance procedures. Effective data governance is important to ensure value creation from data and analytics investments (Grover et al. 2018) and includes knowing who is responsible for a dataset over its lifecycle and having a structured workflow for managing data requests so that efficient access to quality data is guaranteed. By bringing together different user roles, an EDC can support such organizational tasks while facilitating data governance initiatives and ensuring that data stays *"fit for use."* This functional requirement is also being emphasized in DL research with the notion of data curations, defined as *"[t]he activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and reuse"* (Lord et al. 2004, p.1). Hence, the curator role maintains the DL's content over its lifecycle. While the content in DL is rather homogenous and usually publicly available, data in the enterprise context is rather heterogenous, involves more complex rules, and is often confidential. Hence, EDC solutions on the market provide different functionalities to facilitate data governance, such as nominating roles, assigning responsibilities, and establishing workflows for data throughout the company. As an example, Collibra's EDC solution provides a workflow to support the data authorization process. In IBM's InfoSphere Information Governance Catalog, governance policies can be documented and rules put in place to guide how data should be managed and used.

With the **Data visualization** function group, user roles can visualize data values, key metrics, data dependencies, and metadata about data by using dashboards or cockpits and data flow or network visualizations. This function group facilitates the other function groups and helps user roles with decision making. In Collibra, the lineage of data can be visualized to gain transparency in how data flows between systems. All the data in Informatica's EDC can be visualized through the company Tableau's solutions as third-party integration.

With the **Automation & machine learning** function group, other function groups are supported by either automating or facilitating certain tasks (e.g., data assessment, recommendations). This automation can be done by using a rules-based or a learning-based approach. In Zaloni's solution, complete workflows can be automated using rules. In Alation's EDC solution, a learning-based approach is used to recommend which tables users should join when they start typing a query.

With the **Administration** function group, typical functionalities are provided that help application managers in managing users and configuring the optimal use of the EDC solution.

4.3 Data View

According to section 2.3, a platform for data democratization comprises a data inventory and relies on metadata describing various aspects of the data (incl. relationships). Describing data through metadata increases data reusability and has been highlighted in the DL literature (Borgman et al. 2015). Therefore, the EDC reference model's data view is expressed in the form of a metadata model (Kerhervé and Gerbé 1997) and comprises metadata objects that are to be documented with their relationships. Our proposed model addresses the following requirements, which were identified in focus groups: First, it should align the different perspectives on data – specifically, the business-oriented and the system-oriented perspectives. Second, it should support data democratization and provide data documentation for typical data consumers (both experts and non-experts: data citizens, data analysts, data protection officers, data architects, data stewards, and data owners). To reconcile both business- and system-oriented perspectives on data, metadata objects follow data modeling guidelines and are organized into three layers (Batini et al. 1986; Tsichritzis and Klug 1978): conceptual, logical, and physical. As the business alignment of the model was a critical requirement, in line with the goals of data democratization, the conceptual layer was broken down into specific views to address governance and analytics considerations, in addition to classical business concepts.

Table 44. Data view: metadata model layers, views, and objects

Modeling layer	Model view	Metadata object
Conceptual layer	Business process view	Business process
		Business capability
		Business domain
	Business terminology view	Business term
	Analytics view	Metric
		KPI
		Report
	Governance view	Actor
		Role
		Board/council
		Regulations & guidelines
Logical layer	Logical data view	Application
		Transformation
		Data domain
		Business object
		Business object attribute
		Value domain
Physical layer	Physical data view	Data object
		Data object attribute
		Data structure
		System
		Interface

The conceptual layer depicts a high-level, business understanding of the data and includes several views that are specific to the enterprise context. They comprise the different usage contexts that depict where and how data is created and used in the enterprise (i.e., governance, business process, analytics, and the related business terminology):

- using the documentation of business domains, capabilities, and processes, the business process view describes where and how data is used inside an organization. *Business processes* represent how an enterprise performs its activities and are enabled by *business capabilities*, which consist of a combination of technological, informational, and organizational resources and represent what a company does (Bharadwaj 2000; Grant 1991). The business domain represents the strategic business areas of a company and reflects its strategic goals.
- the business terminology view documents *business terms*, referring to business objects and their attributes, to provide users with definitions and guidelines on data – it documents key terms in a way that business users can understand.

- the analytics view refers to metrics, key performance indicators, and reports. Metrics provide quantifiable measure reflecting the state of the enterprise. They are the basic key performance indicators (KPIs). Finally, reports organize and present metrics and/or KPIs in human-readable form that enables visualization on different dimensions.
- the governance view integrates individuals (*actors*) and their responsibilities and *roles* in the enterprise (Khatri and Brown 2010; Weber et al. 2009). It also depicts internal (e.g., standards) and external (e.g., laws) guidelines, as well as advisory groups that may influence the way data is managed and used (El Kharbili 2012).

The logical layer reflects the information systems view on data and constitutes an abstraction layer between the storage instances of data on the physical layer and their business meaning on the conceptual layer (Kumpati 1988). It represents a more structured but system-agnostic view of the conceptual model (Tupper 2011), focuses on the detail level of entities and their relationships, and documents the core *data domains*, as well as related business object and their business object attributes, along with the *applications* that create and *transform* them. Among others, it contains single entity definitions (e.g., a "customer" could be mapped to multiple physical instantiations and have various conceptual meanings depending on the specific business context).

The physical layer reflects the implementation view of data and represents the way data is organized and stored in enterprise systems (e.g., databases). In this layer, *systems*, *interfaces,* and *data structures* (e.g., relational database, graph) are documented, along with *data objects* and *data object attributes*, which are the physical projection(s) of business objects and business object attributes, respectively.

5 Contribution, Discussion and Implications

5.1 Contribution: EDC Reference Model

Our study enriches the ongoing scientific discourse on data democratization and provides a grounded definition of the EDC concept. The key contribution is a reference model that conceptualizes EDCs by defining their key components through a triptych of architecture views and their interconnection: the organization view, the function view, and the data view. The *organization view* outlines the user requirements of eight EDC user roles in the form of user stories and links each role to the required function groups and metadata objects. This perspective shows that EDCs act as integrated platforms connecting different user roles (e.g., data scientist, data owner) while efficiently coordinating data management activities (e.g., managing data access) across the company. The *function view* defines nine function groups to support data supply and demand. This part expands the general functions derived from the DS concept (Franklin et al. 2005), which are use-case-agnostic, and transposes them for the enterprise context. In the EDC reference model, each function group is defined from a user perspective and, therefore, puts the required functional capabilities in the enterprise context (e.g., data analytics and data governance). The *data view* outlines supporting metadata objects that enable the FAIR principles (Wilkinson et al. 2016) and is intended to serve as a blueprint for enterprises seeking to design their own, company-specific metadata model in support of providing data documentation for data democratization platforms. By proposing enterprise-specific metadata objects featuring views dedicated to usage and governance contexts and grouped in conceptual, logical, and physical layers, it goes beyond existing metadata standards that contain flat lists of attributes.

As generic and abstract design knowledge, the EDC reference model explicates the (implicit) design knowledge that we derived from situational inquiry (i.e., insights from company-specific EDC initiatives) and materialized instantiations (i.e., EDC solutions and pilot implementations). As a recommended practice, it is intended to form the basis for assessing vendor solutions and creating company-specific situational designs (instantiation).

5.2 Discussion: EDC's role in Future IT Landscape

5.2.1 Distinctive characteristics of EDCs as platforms for data democratization

The EDC reference model anchors these emerging platforms for data democratization in enterprises to related concepts, including the digital library and the dataspace (overall concept), metadata management (data view with metadata objects), and data governance (organization view with user roles). Thus, we see the EDC as an evolutionary concept of metadata management because it aggregates existing metadata concepts (i.e., data dictionary, business glossary, and metadata repository) to provide a holistic viewpoint on data and connect technical and business-oriented user roles (see Table 45). From a functional perspective, data dictionaries, business glossaries, and metadata repositories serve the purpose of a data inventory, as they provide documentation for all data or business objects. Business glossaries and metadata repositories also support governance efforts, as they provide additional information (e.g., definition, metrics) on the data. Metadata repositories can also support data discovery functions by acting as an index for documented data. However, data dictionaries, business glossaries, and data repositories are focused tools that cater to specific categories of users and operate at a defined information layer. In comparison to these concepts, EDCs facilitate data democratization for a broad audience within organizations. This highlights the key differentiator of data catalogs, which extend preceding metadata management solutions from a functional perspective by enriching data documentation capabilities with data usage capabilities, thus catering to the needs of a broader variety of users.

Our analysis also shows that EDCs provide core functionalities that enable the FAIR principles (i.e., data inventory, data discovery and delivery, data governance, and data visualization) by ensuring that employees can find, access, and understand the data and put it to use. In addition, EDCs offer functionalities that enable the direct use of data resources (i.e., data analytics, data assessment, automation & ML), as well as interactions between users (i.e., data collaboration, which has high coverage and priority) within the platform itself. These two aspects even go beyond the FAIR principles and constitute specificities of data democratization in the enterprise context.

Table 45. EDCs compared with other metadata management concepts

	Data dictionary	Business glossary	Metadata repository	EDC	Governance EDC	Analytics EDC
Roles						
Data citizen						
Data owner						
Data steward						
Chief data officer						
Data analyst/ scientist						
Compliance officer						
Data architect						
Solution architect						
Function groups						
Data inventory						
Data discovery						
Data analytics						
Data assessment						
Data collaboration						
Data governance						
Data visualization						
Automation & ML						
Information layer						
Conceptual						
Logical						
Physical						

5.2.2 EDC's archetypes

Our findings highlight the wide range of data catalogs. These include their capabilities to act as a front end for managing enterprise-wide data assets, satisfy the needs of a variety of technical and business users, and facilitate collaboration. However, EDCs are far from being uniform solutions. While most of the EDC solutions on the market offer basic functionalities to inventory, govern, and discover data, none of them cover all the function groups. In fact, the inventory function is the common denominator. Most of the analyzed EDCs are completely standalone solutions, while certain solutions (e.g., Ab Initio, Informatica, Talend, and SAP)[4] require a combination of several components and tools from the various product portfolios.

[4] For example, in the case of Talend, the "data catalog" is part of the "govern" capability, alongside "data quality," "data preparation," "data stewardship," and "data inventory."

The analysis and comparison of EDC solutions and implementation projects provide interesting insights as they help identify specific patterns (archetypes) of EDCs.

- **Analytics-oriented EDC:** Some EDC solutions primarily focus on the management of data lakes and thereby target data analysts/scientists as user roles. These solutions take advantage of machine-learning technology to build up a data inventory by scanning, collecting, and describing data in a highly automated fashion. In addition, these tools offer analytics functions to support the management of data lake environments. Solutions in this category include the Anzo Smart Data Lake 4.0 (Cambridge Semantics), Enterprise Data Catalog (Informatica), Smart Data Catalog (Waterline), and Zaloni Data Management Platform.
- **Governance-oriented EDC:** Other EDC solutions focus on data collaboration and data governance. These tools primarily aim to support data management workflows. With these tools, the data inventory is built up through manual action on the part of the EDC users. Solutions in this category include Adaptive Metadata Manager, Collibra Data Governance Center, Information Value Management (Datum), IBM InfoSphere Information Governance Catalog, Axon Data Governance (Informatica), and SAP Information Steward.

This analysis demonstrates the broad range of data catalog solutions and their roles in future IT landscapes. Thus, it underpins the importance of the reference model for setting the scope for an EDC in terms of the target group and functional scope and for comparing and assessing the different solutions.

5.3 Limitations and Outlook on Future Research

As with any empirical work, this study has limitations. Since EDCs are a novel concept, most of the enterprises were still in the early phase of adoption. Moreover, EDC vendors extend their functionalities. Therefore, we strongly encourage future research on EDC to validate and improve the reference model but also to investigate the analytics-oriented and governance-oriented EDCs. Building on our research, we see interesting avenues for future studies: Since enterprises increasingly source external data, potential integrations of EDCs with open data portals and data marketplaces seem to be a promising research direction. Further research could also explore how data valuation approaches could complement the existing assessment functionality.

6 References

Ahlemann, F., and Riempp, G. 2008. "RefModPM: A Conceptual Reference Model for Project Management Information Systems," *Wirtschaftsinformatik* (50:2), pp. 88–97.

Awasthi, P., and George, J. J. 2020. "A Case for Data Democratization," in *Proceedings of the 26th Americas Conference on Information Systems (AMCIS)*, Virtual Conference, August 10, p. 23.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. 2009. "Methodologies for Data Quality Assessment and Improvement," *ACM Computing Surveys* (41:3), pp. 1–52. (https://doi.org/10.1145/1541880.1541883).

Batini, C., Lenzerini, M., and Navathe, S. B. 1986. "A Comparative Analysis of Methodologies for Database Schema Integration," *ACM Computing Surveys* (18:4), pp. 323–364. (https://doi.org/10.1145/27633.27634).

Belissent, J., Leganza, G., and Vale, J. 2019. "Determine Your Data's Worth: Data Plus Use Equals Value," Consortium Report, Consortium Report, Forrester Research, February 5. (https://www.forrester.com/report/Determine+Your+Datas+Worth+Data+Plus+Use+Equals+Value/-/E-RES127541).

Bharadwaj, A. S. 2000. "A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Investigation," *MIS Quarterly* (24:1), pp. 169–196. (https://doi.org/10.2307/3250983).

Borgman, C. L. 2003. From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World, MIT Press.

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., and Traweek, S. 2015. "Knowledge Infrastructures in Science: Data, Diversity, and Digital Libraries," *International Journal on Digital Libraries* (16:3–4), pp. 207–227. (https://doi.org/10.1007/s00799-015-0157-z).

Bowne-Anderson, H. 2018. "What Data Scientists Really Do, According to 35 Data Scientists," *Harvard Business Review Digital Articles*, pp. 2–5.

Calhoun, K. 2014. Exploring Digital Libraries: Foundations, Practice, Prospects, London, UK: Facet Publishing.

Chaki, S. 2015. "Pillar No. 7: Metadata Management," in *Enterprise Information Management in Practice: Managing Data and Leveraging Profits in Today's Complex Business Environment*, S. Chaki (ed.), Berkeley, CA: Apress, pp. 115–127. (https://doi.org/10.1007/978-1-4842-1218-9_10).

Chang, T.-H., Fu, H.-P., Ou, J.-R., and Chang, T.-S. 2007. "An ARIS-Based Model for Implementing Information Systems from a Strategic Perspective," *Production Planning & Control* (18:2), Taylor & Francis, pp. 117–130. (https://doi.org/10.1080/09537280600913447).

Dallemulle, L., and Davenport, T. 2017. "What's Your Data Strategy?," *Harvard Business Review* (May-June), pp. 112–121.

De Simoni, G., Dayley, A., and Edjlali, R. 2018. "Magic Quadrant for Metadata Management Solutions," Consortium Report, Consortium Report, Gartner.

De Simoni, G., White, A., Jain, A., and Dayley, A. 2018. "Market Guide for Information Stewardship Applications," Consortium Report, Consortium Report, Gartner.

Díaz, A., Rowshankish, K., and Saleh, T. 2018. "Why Data Culture Matters," *The McKinsey Quarterly* (3), p. 37.

Dublin Core Metadata Initiative. (n.d.). "DCMI: DCMI Metadata Terms." (https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, accessed August 25, 2019).

Duncan, A. D., Laney, D., and De Simoni, G. 2016. "How Chief Data Officers Can Use an Information Catalog to Maximize Business Value From Information Assets," Consortium Report, Consortium Report, Gartner.

El Kharbili, M. 2012. "Business Process Regulatory Compliance Management Solution Frameworks: A Comparative Evaluation," in *Proceedings of the 8th Asia-Pacific Conference on Conceptual Modelling (APCCM)* (Vol. 130), Melbourne, Australia, January, pp. 23–32. (http://dl.acm.org/citation.cfm?id=2523782.2523786).

Fadler, M., and Legner, C. 2020. "Building Business Intelligence & Analytics Capabilities - A Work System Perspective," in *Proceedings of the 41st International Conference on Information Systems (ICIS)*, Hyderabad, India, December 13, p. 2615. (https://aisel.aisnet.org/icis2020/governance_is/governance_is/14).

Fettke, P., and Loos, P. 2003. "Classification of Reference Models: A Methodology and Its Application," *Information Systems and E-Business Management* (1:1), pp. 35–53. (https://doi.org/10.1007/BF02683509).

Fox, E. A., and Sornil, O. 2003. "Digital Libraries," in *Encyclopedia of Computer Science* (4th ed.), Chichester, UK: John Wiley and Sons Ltd., pp. 576–581. (http://dl.acm.org/citation.cfm?id=1074100.1074337).

Frank, U. 1999. "Conceptual Modelling as the Core of the Information Systems Discipline - Perspectives and Epistemological Challenges," in *Proceedings of the 5th Americas Conference on Information Systems*, Milwaukee, USA, p. 3.

Frank, U. 2014. "Multilevel Modeling: Toward a New Paradigm of Conceptual Modeling and Information Systems Design," *Business & Information Systems Engineering* (6:6), pp. 319–337. (https://doi.org/10.1007/s12599-014-0350-4).

Frank, U., Strecker, S., Fettke, P., Vom Brocke, J., Becker, J., and Sinz, E. 2014. "The Research Field "Modeling Business Information Systems"," *Business & Information Systems Engineering* (6:1), pp. 39–43.

Franklin, M., Halevy, A., and Maier, D. 2005. "From Databases to Dataspaces: A New Abstraction for Information Management," *ACM SIGMOD Record* (34:4), pp. 27–33. (https://doi.org/10.1145/1107499.1107502).

George, G., Haas, M. R., and Pentland, A. 2014. "Big Data and Management," *Academy of Management Journal* (57:2), pp. 321–326. (https://doi.org/10.5465/amj.2014.4002).

Goetz, M., Leganza, G., and Hennig, C. 2020. "Now Tech: Machine Learning Data Catalogs, Q4 2020," Consortium Report, Consortium Report, Forrester Research. (https://www.forrester.com/report/Now%20Tech%20Machine%20Learning%20Data%20Catalogs%20Q4%202020/-/E-RES157529).

Goetz, M., Leganza, G., Hoberman, E., and Hartig, K. 2018. "The Forrester Wave™: Machine Learning Data Catalogs, Q2 2018," Consortium Report, Consortium Report, Forrester Research.

Grant, R. M. 1991. "The Resource-Based Theory of Competitive Advantage: Implications for Strategy Formulation," *California Management Review* (33:3), p. 114.

Grover, V., Chiang, R. H. L., Liang, T.-P., and Zhang, D. 2018. "Creating Strategic Business Value from Big Data Analytics: A Research Framework," *Journal of Management Information Systems* (35:2), pp. 388–423.

Hai, R., Geisler, S., and Quix, C. 2016. "Constance: An Intelligent Data Lake System," in *Proceedings of the 2016 International Conference on Management of Data*, New York, NY, USA: ACM, pp. 2097–2100. (https://doi.org/10.1145/2882903.2899389).

Halevy, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., and Whang, S. E. 2016. "Goods: Organizing Google's Datasets," in *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, San Francisco, California, USA: ACM Press, pp. 795–806. (https://doi.org/10.1145/2882903.2903730).

Hellerstein, J. M., Sreekanti, V., Gonzalez, J. E., Dalton, J., Dey, A., Nag, S., Ramachandran, K., Arora, S., Bhattacharyya, A., Das, S., Donsky, M., Fierro, G., She, C., Steinbach, C., Subramanian, V., and Sun, E. 2017. "Ground: A Data Context Service," in *Proceedings of the 8th Conference on Innovative Data Systems Research (CIDR)*, Chaminade, California, USA, January 8, p. 12.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75–105.

Hyun, Y., Kamioka, T., and Hosoya, R. 2020. "Improving Agility Using Big Data Analytics: The Role of Democratization Culture," *Pacific Asia Journal of the Association for Information Systems* (12:2). (https://doi.org/10.17705/1pais.12202).

International Organization for Standards / International Electrotechnical Commission (ISO/IEC). 2013. International Standard ISO/IEC 11179-3. Information Technology - Metadata Registries (MDR) - Part 3: Registry Metamodel and Basic Attributes. (https://standards.iso.org/ittf/PubliclyAvailableStandards/c050340_ISO_IEC_11179-3_2013.zip).

Kahn, R., and Wilensky, R. 1995. "Kahn/Wilensky Architecture: A Framework for Distributed Digital Object Services." (http://www.cnri.reston.va.us/k-w.html, accessed July 18, 2019).

Kerhervé, B., and Gerbé, O. 1997. "Models for Metadata or Metamodels for Data?," in *Proceedings of the 2nd IEEE Metadata Conference*, Silver Spring, Massachusetts, USA, September.

Khatri, V., and Brown, C. V. 2010. "Designing Data Governance," *Communication of the ACM* (53:1), pp. 148–152.

Kumpati, M. 1988. Database Management System with Active Data Dictionary.

Labadie, C., Legner, C., Eurich, M., and Fadler, M. 2020. "FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs," in *Proceedings of the 22nd IEEE Conference on Business Informatics (CBI)* (Vol. 1), Antwerp, Belgium, June 22, pp. 201–210. (https://doi.org/10.1109/CBI49978.2020.00029).

Lagoze, C., Krafft, D. B., Payette, S., and Jesuroga, S. 2005. What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL, p. 23.

Lee, Y., and Strong, D. 2004. "Knowing-Why about Data Processes and Data Quality," *Journal of Management Information Systems* (20:3). (http://www.jstor.org/stable/40398639).

Legner, C., Pentek, T., and Otto, B. 2020. "Accumulating Design Knowledge with Reference Models: Insights from 12 Years of Research on Data Management," *Journal of the Association for Information Systems* (21:3).

Lord, P., Macdonald, A., Lyon, L., and Giaretta, D. 2004. "From Data Deluge to Data Curation," in *In Proc 3th UK E-Science All Hands Meeting*, pp. 371–375.

Otto, B., Hompel, M. ten, and Wrobel, S. 2019. "International Data Spaces," in *Digital Transformation*, R. Neugebauer (ed.), Berlin, Heidelberg: Springer, pp. 109–128. (https://doi.org/10.1007/978-3-662-58134-6_8).

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24), pp. 45–77. (https://doi.org/10.2753/MIS0742-1222240302).

Peyret, H., Cullen, A., Kramer, A., and Bartlett, S. 2017. "The Forrester Wave™: Data Governance Stewardship And Discovery Providers, Q2 2017," Consortium Report, Consortium Report, Forrester Research.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. 2002. "Data Quality Assessment," *Commun. ACM* (45:4), pp. 211–218. (https://doi.org/10.1145/505248.506010).

Poole, J., Chang, D., Tolbert, D., and Mellor, D. 2002. *Common Warehouse Metamodel. An Introduction to the Standard for Data Warehouse Integration.*, New York, NY, USA: John Wiley & Sons, Inc.

Prat, N., Comyn-Wattiau, I., and Akoka, J. 2015. "A Taxonomy of Evaluation Methods for Information Systems Artifacts," *Journal of Management Information Systems* (32:3), pp. 229–267. (https://doi.org/10.1080/07421222.2015.1099390).

Roszkiewicz, R. 2010. "Enterprise Metadata Management: How Consolidation Simplifies Control," *Journal of Digital Asset Management* (6:5), pp. 291–297. (https://doi.org/10.1057/dam.2010.32).

Russom, P. 2017. "The Data Catalog's Role in the Digital Enterprise: Enabling New Data-Driven Business and Technology Best Practices," Consultancy Report, Consultancy Report, TDWI. (https://tdwi.org/research/2017/11/ta-all-informatica-the-data-catalogs-role-in-the-digital-enterprise).

Sallam, R., Sicular, S., den Hamer, P., Kronz, A., Schulte, W. R., Brethenoux, E., Woodward, A., Emmott, S., Zaidi, E., Feinberg, D., Beyer, M., Greenwald, R., Idoine, C., Cook, H., De Simoni, G., Hunter, E., Ronthal, A., Tratz-Ryan, B., Heudecker, N., Hare, J., and Clougherty Jones, L. 2020. "Top 10 Trends in Data and Analytics, 2020," Consortium Report, Consortium Report, Gartner Research. (https://www.gartner.com/en/doc/718161-top-10-trends-in-data-and-analytics-2020).

Scheer, A.-W. 2001. *ARIS — Modellierungsmethoden, Metamodelle, Anwendungen*, (4th ed.), Berlin Heidelberg: Springer-Verlag. (https://www.springer.com/la/book/9783540416012).

Scheer, A.-W., and Schneider, K. 2006. "ARIS — Architecture of Integrated Information Systems," in *Handbook on Architectures of Information Systems*, International Handbooks on Information Systems, P. Bernus, K. Mertins, and G. Schmidt (eds.), Berlin, Heidelberg: Springer, pp. 605–623. (https://doi.org/10.1007/3-540-26661-5_25).

Sen, A. 2004. "Metadata Management: Past, Present and Future," *Decision Support Systems* (37:1), pp. 151–173. (https://doi.org/10.1016/S0167-9236(02)00208-7).

Short, J. E., and Todd, S. 2017. "What's Your Data Worth?," *MIT Sloan Management Review* (58:3), p. 5.

Sonnenberg, C., and vom Brocke, J. 2012. "Evaluations in the Science of the Artificial – Reconsidering the Build-Evaluate Pattern in Design Science Research," in *Design Science Research in Information Systems. Advances in Theory and Practice*, Lecture Notes in Computer Science, K. Peffers, M. Rothenberger, and B. Kuechler (eds.), Springer Berlin Heidelberg, pp. 381–397.

Stanford. 1999. "The Stanford Digital Libraries Technologies Project." (http://diglib.stanford.edu:8091/, accessed July 19, 2019).

Staples, T., Wayland, R., and Payette, S. 2003. "The Fedora Project: An Open-Source Digital Object Repository Management System," *D-Lib Mag.* (9). (https://doi.org/10.1045/april2003-staples).

Tsichritzis, D., and Klug, A. 1978. "The ANSI/X3/SPARC DBMS Framework Report of the Study Group on Database Management Systems," *Information Systems* (3:3), pp. 173–191. (https://doi.org/10.1016/0306-4379(78)90001-7).

Tupper, C. D. 2011. "Model Constructs and Model Types," in *Data Architecture*, C. D. Tupper (ed.), Boston: Morgan Kaufmann, pp. 207–221. (https://doi.org/10.1016/B978-0-12-385126-0.00011-5).

Uhrowczik, P. P. 1973. "Data Dictionary/Directories," *IBM Systems Journal* (12:4), pp. 332–350. (https://doi.org/10.1147/sj.124.0332).

Upadhyay, P., and Kumar, A. 2020. "The Intermediating Role of Organizational Culture and Internal Analytical Knowledge between the Capability of Big Data Analytics and a Firm's Performance," *International Journal of Information Management* (52), p. 102100. (https://doi.org/10.1016/j.ijinfomgt.2020.102100).

Vom Brocke, J. 2007. "Design Principles for Reference Modeling: Reusing Information Models by Means of Aggregation, Specialisation, Instantiation, and Analogy," in *Design Principles for Reference Modeling: Reusing Information Models by Means of Aggregation, Specialisation, Instantiation, and Analogy*, IGI Global.

Wallace, D. P., and Van Fleet, C. 2005. "The Democratization of Information? Wikipedia as a Reference Resource," *Reference & User Services Quarterly* (45:2), pp. 100–103.

Wang, R., and Strong, D. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), pp. 5–33.

Weber, K., Otto, B., and Österle, H. 2009. "One Size Does Not Fit All---A Contingency Approach to Data Governance," *Journal of Data and Information Quality* (1:1), pp. 1–27.

Wilcox, D. 2018. "Supporting FAIR Data Principles with Fedora," *LIBER Quarterly* (28:1), pp. 1–8. (https://doi.org/10.18352/lq.10247).

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* (3:1), pp. 1–9. (https://doi.org/10.1038/sdata.2016.18).

Winter, R., and Schelp, J. 2006. "Reference Modeling and Method Construction: A Design Science Perspective," in *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC '06, New York, USA: ACM, pp. 1561–1562. (https://doi.org/10.1145/1141277.1141638).

Wixom, B., and Ross, J. 2017. "How to Monetize Your Data," *MIT Sloan Management Review* (58:3).

World Wide Web Consortium (W3C). (n.d.). "Data Catalog Vocabulary (DCAT)." (https://www.w3.org/TR/vocab-dcat/, accessed August 25, 2019).

Zaidi, E., De Simoni, G., Edjlali, R., and Duncan, A. D. 2017. "Data Catalogs Are the New Black in Data Management and Analytics," Consultancy Report, Consultancy Report, Gartner, December 13. (https://www.gartner.com/doc/reprints?id=1-4MKJU2Y&ct=171220&st=sb&submissionGuid=12d68804-ceec-454e-b412-a66bdff38e2e).

Zeng, J., and Glaister, K. W. 2018. "Value Creation from Big Data: Looking inside the Black Box," *Strategic Organization* (16:2), SAGE Publications, pp. 105–140. (https://doi.org/10.1177/1476127017697510).