UNIL | Université de Lausanne

Unicentre
CH-1015 Lausanne
http://serval.unil.ch

Year : 2021

Essays on Data Democratization & Protection in the Data-driven Enterprise

Labadie Clément

UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES

DÉPARTEMENT DES SYSTÈMES D’INFORMATION

ESSAYS ON DATA DEMOCRATIZATION & PROTECTION IN THE DATA-DRIVEN ENTERPRISE

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

Clément LABADIE

Directrice de thèse
Prof. Christine Legner

Jury

Prof. Felicitas Morhart, présidente
Prof. Michalis Vlachos, expert interne
Prof. Öykü Işik, experte externe
Prof. Felix Wortmann, expert externe

LAUSANNE
2021

UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES

DÉPARTEMENT DES SYSTÈMES D'INFORMATION

ESSAYS ON DATA DEMOCRATIZATION & PROTECTION IN THE DATA-DRIVEN ENTERPRISE

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

Clément LABADIE

Directrice de thèse
Prof. Christine Legner

Jury

Prof. Felicitas Morhart, présidente
Prof. Michalis Vlachos, expert interne
Prof. Öykü Işik, experte externe
Prof. Felix Wortmann, expert externe

LAUSANNE
2021

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

IMPRIMATUR

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 Clément LABADIE, titulaire d'un bachelor en Droit de l'Université de Neuchâtel, et d'un master en Droit, Criminalité et Sécurité des Technologies de l'Information de l'Université de Lausanne, en vue de l'obtention du grade de docteur ès Sciences en systèmes d'information.

La thèse est intitulée :

ESSAYS ON DATA DEMOCRATIZATION & PROTECTION IN THE DATA-DRIVEN ENTERPRISE

Lausanne, le 22 juin 2021

Le doyen

Jean-Philippe Bonardi

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. Öykü IȘIK

Professor at the International Institute for Management Development (IMD), Lausanne, Switzerland.

External member of the thesis committee

Prof. Felix WORTMANN

Professor at the School of Management of the University of St. Gallen (HSG), Switzerland.

External member of the thesis committee

Prof. Felicitas MORHART

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 (HEC)

PhD in Information Systems

I hereby certify that I have examined the manuscript submitted by

Clément LABADIE

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or other committee members requested during the doctoral colloquium have been addressed to my satisfaction.

Signature: ______________________ Date: 17/6/2021

Prof. Christine LEGNER

Thesis supervisor

University of Lausanne

Faculty of Business and Economics (HEC)

PhD in Information Systems

I hereby certify that I have examined the manuscript submitted by

Clément Labadie

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or other committee members requested during the doctoral colloquium have been addressed to my satisfaction.

Signature: ______________________ Date: June 17 2021

Prof. Michalis Vlachos

Internal member of the thesis committee

University of Lausanne

Faculty of Business and Economics (HEC)

PhD in Information Systems

I hereby certify that I have examined the manuscript submitted by

Clément LABADIE

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or other committee members requested during the doctoral colloquium have been addressed to my satisfaction.

Signature: __Öykü___ Date: ____14/06/2021___

Prof. Öykü IŞIK

External member of the thesis committee

University of Lausanne

Faculty of Business and Economics (HEC)

PhD in Information Systems

I hereby certify that I have examined the manuscript submitted by

Clément LABADIE

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or other committee members requested during the doctoral colloquium have been addressed to my satisfaction.

Signature: Felix W______________ Date: 13.06.2021

Prof. Felix WORTMANN

External member of the thesis committee

Doctoral Thesis

ESSAYS ON DATA DEMOCRATIZATION & PROTECTION IN THE DATA-DRIVEN ENTERPRISE

Clément Labadie

Department of Information Systems,

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

Lausanne, 2021

Acknowledgements

This work is dedicated to the memory of Eve Andrieu, whose guidance set me on the path leading to this moment, and will continue to shape the road ahead.

Special thanks to Prof. Christine Legner for taking a chance on a not-so-obvious candidate for a PhD in Information Systems, and for her resolve in encouraging its realization. Thank you to Professors Öykü Işik, Felix Wortmann and Michalis Vlachos, for dedicating time to evaluating and discussing my work.

My sincere gratitude goes out to all the people who have supported me during this journey.

To my partner, Pierre Guichon – thank you for bringing this moment within reach with your patience, care and attention.

To my parents, Christine & Michel Labadie – thank you for your unconditional support and for never interfering with my academic choices, however unconventional or confusing. You have made this moment possible.

To my PhD family, Dr. Dina Elikan, Dr. Gianluca Basso and Francesca Gregorio – you are the best, and yours is proof that friendships are not only measured in years.

To my esteemed colleagues, Martin Fadler, Léonore Cellier and Dr. Dimitri Percia David – thank you for your great knowledgeability, and for knowing to be great peers.

To the entire CDQ team and CC members. To Dr. Dimitrios Gizanis, Dr. Tobias Pentek, Dr. Martin Böhmer, Dr. Markus Eurich, Maria Hameister and Eva Weller in particular – thank you for your considerate guidance and support.

To my teammates, Dr. Thomas Boillat, Dr. Dana Naous, Dr. Johannes Schwarz, Dr. Louis Vuilleumier, Dr. Gaël Bernard, Bastien Wanner, Andreas Lang, Matthieu Harbich, Pavel Krasikov, Valerianne Walter and Hippolyte Lefebvre – thank you for having made life in Dorigny so joyous through our many chats, laughs, coffees and lunches.

To my student assistants, Lama El Zein, Stephanie Arreguit O'Neill, Lev Velykoivanenko, Maxime Lucie Bayle and Gabin Flourac – thank you for your invaluable help and reliability, which freed up both my time and mind.

“Laws and Principles are not for the times
when there is no temptation.”

– Charlotte Brontë, *Jane Eyre*
(originally published under the masculine pseudonym of “Currer Bell”)

Abstract

In the data-driven enterprise, large quantities of data are automatically processed to detect patterns, generate insights, and fuel business processes and business models alike. However, enterprises striving to become data-driven face the following challenges. First, while a broadening audience of enterprise stakeholders need to work with data, they largely remain lodged in organizational and technical silos. Second, organizations must comply with an increasing number of data protection regulations to exploit personal data, which is key to successful data-driven business strategies (e.g., customer relationship management, know your customer, 360-degree customer view). Set in a consortium research environment, this dissertation comprises six essays organized in two research streams that explore the seemingly contradictory objectives of data protection compliance and increased data usage in enterprises. The first research stream investigates enterprise data catalogs (EDC) as emerging platforms that support data democratization and implement the FAIR principles (i.e., findable, accessible, interoperable, reusable) in the enterprise context. Based on situational inquiry (i.e., insights from over 10 company-specific EDC initiatives) and materialized instantiations (i.e., EDC solutions and pilot implementations), we identify emerging data-related roles and analyze how enterprise data catalogs support their typical data needs and usages. We synthesize our findings in a reference model and outline implementation approaches. From an academic perspective, our findings extend existing literature by exploring the ways and processes towards data democratization and by providing a comprehensive conceptualization of enterprise data catalogs from three architecture views. The second research stream examines the interplay between data protection regulations and data processing practices. We link compliance requirements to data management capabilities, by proposing a capability model for data protection and operationalizing it through a personal data life cycle model enriched with a data model and business rules. From an academic perspective, we contribute to the regulatory compliance management research domain by unifying data protection and data management requirements. In collating the outcomes of these two research streams, we find that creating transparency on enterprise-wide data assets is a cornerstone of both data protection compliance and data democratization. Our contributions inform the redesign of data management practices and call attention to the need for diligent data documentation as essential building block of the conformable and democratized data-driven enterprise.

Contents

Introduction

Research stream 1

Research stream 2

Introductory Paper on

Essays on Data Democratization and Protection in the Data-driven Enterprise

Clément Labadie

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

Table of contents

List of figures

List of tables

1 Introduction

In 2021, data are produced at an unprecedented rate[1] and have been recognized as the currency of the digital economy (Szczepański 2020). The promise of the information society, where information technologies play a fundamental role and where exchanging data is paramount, has materialized as the next societal evolutionary steps in industrial nations (Beniger 1986) – the many (supra-)governmental initiatives surrounding it are a testimony of this shift[2], such as the European GAIA-X project, which aims at developing a federated data infrastructure to ensure data sovereignty in Europe (Federal Ministry for Economic Affairs and Energy (BMWi) 2020), and the European Union's (EU) Data Strategy, that defines a comprehensive policy framework to position the EU as "a leading role model for a society empowered by data to make better decisions – in business and the public sector" (European Commission 2020). This new paradigm has been fueled by the accelerated development of computing power and networking bandwidth, as well as by a rise in their use in the general public (e.g., wide-spread adoption of the Internet and increased use of social media and other digital services, through always-on connected devices). As a result of the convergence of these developments, data is increasingly pervasive, and new technologies, especially artificial intelligence and machine learning, are being developed to leverage those fast-growing data volumes.

In the corporate world, the shift translates into a push for organizations to digitalize their operations and leverage their data resources, in an attempt to generate additional business value from them (Dallemulle and Davenport 2017; Wixom and Ross 2017). With this objective on the horizon, new data-driven business models have emerged, based on the provision of digital services relying on large-scale exploitation of data (Schüritz et al. 2017), and existing offerings and processes are infused with data (Bharadwaj et al. 2013; Matt et al. 2015; Wixom et al. 2013). These data-driven opportunities are made possible by technologies that enable organizations to extract as much value as possible from their data resources (Abbasi et al. 2016; George et al. 2014). Self-service business intelligence (BI – Işık et al. 2013) and Big data & analytics (BDA, including artificial intelligence & machine learning – Fadler and Legner 2020a) are prime examples thereof, which many organizations attempt to integrate into their practices. While academic literature is rich with contributions outlining how these technologies can be leveraged to augment data processing capabilities in enterprises, the ability to translate them into actual

[1] In a 2017 report, IBM estimated that 90% of the world's data at the time had been created within the two previous years alone (IBM 2017).

[2] Other examples include the European Union's *eEurope* action plan, *i2010* strategy, *Digital Single Market* initiative & *Digital Strategy*, and the United Nations General Assembly's *resolution A/RES/60/252*.

increased business value is less clear, and many data-driven initiatives are unable to move forward from a project state (Grover et al. 2018).

From an internal perspective, organizations accelerate data-driven efforts by making their data resources available enterprise-wide and breaking down "exclusionary data structures" (Schlagwein et al. 2017). They open them up to a broad audience of business stakeholders and establishing an enterprise data culture (Hyun et al. 2020), a phenomenon referred to as data democratization (Awasthi and George 2020).

From an external perspective, regulatory pressure is ever increasing for large organizations. In particular, data protection regulations aim at mitigating the risks of legitimate custodians of personal data using it for unauthorized purposes (Burt 2019; Meier 2011). As data use becomes pervasive in enterprises, older data protection regulations pre-dating the digital era were in need of a refresh (De Hert and Papakonstantinou 2012, 2016; Mitrou 2017; Nicolaidou and Georgiades 2017). The recent EU-GDPR was the first significant regulatory framework update meant to address this new data-driven reality, and it inspired similar initiatives around the world. These new regulations strengthen existing data protection rights and require organizations to expose clear and extensive information about their processing activities. Hence, they must establish capabilities to systematically document all personal data processing purposes and ensure that they can demonstrate compliant processing.

Considering these developments, this thesis aims to examine how companies address two distinct, yet interrelated phenomena: data democratization and data protection compliance. In doing so, we ask the following overarching research question: what are capabilities that organizations need to address to increase usage of data resources enterprise-wide, while complying with data protection regulatory requirements? We explore each phenomenon in a dedicated research stream.

The first research stream investigates enterprise data catalogs as emerging platforms that enable data democratization in the enterprise context. Current research has started exploring data democratization (Awasthi and George 2020) and on outline its potential benefits (Hyun et al. 2020), but stays at the conceptual level. It has not yet reached the state of investigation the means, and specifically, the platforms that support data democratization. Enterprise data catalog are emerging platform that advance prior metadata concepts, but their scope and role in enterprise system landscapes has yet to be understood. To address this gap, we analyze the enterprise data catalog concept to uncover its constitutive elements in a reference model, and anchor it in the existing body of knowledge on data platforms, metadata management and data governance. Following a design science research paradigm, we identify emerging data-related roles, analyze how enterprise data catalogs support their typical data needs and usages, and outline implementation approaches.

The second research stream examines the interplay between data protection regulations and data processing practices. In order to link compliance requirements to data management capabilities, we propose a capability model for data protection and concretize it through a personal data life cycle model enriched with data model extensions and business rules. This stream contributes to privacy research in IS (Bélanger and Crossler 2011) and regulatory compliance management (RCM, El Kharbili 2012). It addresses the lack of RCM-related contributions that address data protection regulations (Abdullah et al. 2009) and provide guidance to concretize strategic compliance objectives (Cleven and Winter 2009).

This introductory paper provides an overview of the dissertation – it presents the inputs and outcomes of each research stream and discusses their relationship in the context of the data-driven enterprise. The remainder of the paper is structured as follows. Section 2 sets the contextual and theoretical background for the thesis. It introduces foundational concepts in the areas of enterprise data management and privacy research and highlights the gaps that this thesis intends to bridge. Section 3 provides a high-level view on the thesis objectives and constituent research streams and describes the particular consortium research setting in which it was conducted. Each research stream is then presented individually in sections 4 and 5, in terms of key background aspects, methodologies, contributions and limitations. Finally, section 6 consists of a critical summary of findings, discussing their implications and limits, and elaborating on the interplay between data democratization and data protection.

2 Background

2.1 The data-driven enterprise

2.1.1 The potential of data exploitation

Recent statistics highlight that enterprises tend to accumulate large quantities of data, and a small fraction of it is actually being put to use[3]. The motivation behind the data-driven enterprise is to tap into this hidden potential of data, by leveraging advances in data-processing technologies to generate business impact (Chen et al. 2012).

Enterprises can realize additional business value through data-driven ways of working, (1) by enriching products and services with data and (2) by using data to improve internal processes (Wixom et al. 2013; Wixom and Ross 2017):

- (1) By "wrapping information around products" (Wixom and Ross 2017), enterprises can seize opportunities to drive innovation (Duan et al. 2020) with data, leading to the creation of new products and services that rely on data exploitation to provide additional value compared to standard offerings, e.g. data-driven business models (Brownlow et al. 2015; Schüritz et al. 2017), data-augmented products / internet-of-things (Wortmann and Flüchter 2015).
- (2) By leveraging business analytics capabilities, enterprises can promote the pervasive and dynamic use of their data resources (Wixom et al. 2013) to increase their agility (Park et al. 2017) in making rapid and informed business decisions (Bharadwaj et al. 2013; Ghasemaghaei et al. 2018).

These avenues for increased impact have found fruitful concretizations in corporations, and especially in the context of consumer-oriented businesses. On the one hand, successful "digital-native" companies often entirely rely on data processing in the services they provide to consumers. There is also an increasing number of offerings involving physical devices with data-enabled capabilities – they are one facet of the internet-of-things, and have incentivized organizations considered traditional to initiate data-driven initiatives (Pflaum and Gölzer 2018; Porter and Heppelmann 2015). On the other hand, the ability to leverage existing customer data has revived the interest in customer relationship management (Bohling et al. 2006; Chen and Popovich 2003) – in this context, analytics capabilities exploit large quantities of customer

[3] IDC estimates that 43% of data captured by enterprises goes unused (IDC 2020). Forrester also estimates 60 to 73% of enterprise data is not used in analytics activities (Gualtieri et al. 2016).

information from multiple sources (e.g. customer master and transactional data enriched with history and behavioral data from web-based customer frontends) in order to establish more exhaustive and finer grained customer profiles, which are referred to as the "360 degrees view of the customer" (Bose 2009), the ultimate goal being the ability to predict customer behavior (Kitchens et al. 2018).

In order to realize these modalities of business impact generation, a data-driven enterprise needs to consider its data resources as a strategic asset (Legner et al. 2020), which requires that said data resources are fit-for-use (Grover et al. 2018).

2.1.2 Data management evolution and challenges

Data management has been a long standing topic in both research and practice, and the meaning of fit-for-use has changed along with the role of data for businesses, over three main phases (Legner et al. 2020). Data started with being treated in isolated databases in the 1980s, to become a strategic business enabler in the 2010s (s. Table 1). While data quality was the key topic of early data management efforts to enable data availability in individual business functions, the need to scale this availability at an enterprise-wide level emerged in the 1990s, during the second phase, as integrated and analytical systems started to appear. In the last phase, as enterprises strive to exploit the untapped value potential of data, they first need to ensure that their data resources are accessible to a wider range of stakeholders.

Table 1. The evolution of data management in enterprises (adapted from Legner et al. 2020)

	Phase 1: **Data administration (since the 1980s)**	**Phase 2:** **Quality-oriented data management (since the 1990s)**	**Phase 3:** **Extensions to strategic data management (since the 2010s)**
Business context			
Roles of data	Data as prerequisite for application development and as an enabler of automation in business functions	Data as enabler of enterprise-wide business process and decision-making	Data as enabler of a firm's business models and value propositions
Data resources	Databases for automated data processing in specific enterprise functions (e.g., accounting systems and inventory systems) Structured data	Integrated information systems, e.g., enterprise resource planning systems (ERP) Data warehouses, business intelligence (BI) Mainly internal, structured data	Integrated and connected information systems Data lakes and advanced analytics platforms Large volumes of internal and external data (*big data*), comprising structured and unstructured data sources
Data-related concerns	Data model quality Data availability Data re-use	Enterprise-wide data integration Data quality	Business value and impacts Data compliance Data privacy & security

Data management context			
Perspectives on data management	Data administration (focus on databases)	Quality-oriented data management (focus on data as an enterprise resource)	Strategic data management (focus on data-driven innovation)
Management approach	Database management	Resource management Quality management	Strategic management

These considerations have been conceptualized in the academic community under the umbrella of the FAIR principles (Wilkinson et al. 2016 - Findable, Accessible, Interoperable, Reusable). These principles promote wide-spread data availability and usage. In the enterprise, this push for a wider scope of availability is referred to as data democratization, and build on enterprise-wide availability of data, by putting it in the hands of an array of employees spanning beyond those traditionally considered as data experts (Awasthi and George 2020), and by fostering a data culture (Dubey et al. 2019).

In reality, even though organizations recognize the strategic aspect of data, many of them are still in the process of eliminating technical and organizational silos that prevent enterprise-wide access to data resources (Hai et al. 2016; Roszkiewicz 2010), thus hindering data analytics activities. From a technical perspective, as large organizations typically rely on highly distributed system landscapes, data resources are stored in a multiplicity of storage systems that are not fully integrated. Even though paradigms such as master data management (Cleven and Wortmann 2010; Vilminko-Heikkinen and Pekkola 2017) are examples of attempts at resolving these issues, they are challenging initiatives that have often been unsuccessful (Marsh 2005). This is partly due to the little perceived value of centralization over autonomy by business users (Van Alstyne et al. 1995).

From an organizational perspective, as corporate structures are typically organized around the division of labor concept, data tends to be produced and consumed within the boundaries of individual business units. Overcoming these organizational silos has been one of the objectives of data governance (Khatri and Brown 2010; Weber et al. 2009), which “defines roles, and […] assigns responsibilities for decision areas to these roles” (Weber et al. 2009). The changing role of data in business also calls for rethinking the concept of data ownership, especially in the context of large-scale analytics (Fadler and Legner 2020b). Related initiatives tend to face similar reservations as master data management ones.

The evolution of data management resonates with the discourse on the openness of information (Schlagwein et al. 2017). Specifically, there are four key principles behind the concept of openness (Schlagwein et al. 2017), with regards to a given resource, in our case, enterprise data:

- *transparency*, according to which enterprise data should be exposed and open to examination,
- *access*, according to which there should exist modalities to exercise this examination, i.e., by being provided with tools to access the data,
- *participation*, according to which a set of individuals, or users, should be able to be afforded to use such tools,
- *democracy*, according to which this set of users should be as inclusive as possible.

2.2 Two different angles on data privacy

2.2.1 The Information Systems perspective on data privacy

As data becomes a strategic asset in enterprises, and is made available to larger groups of stakeholders, threats of insider misuse and breach arise, which are related to information privacy concern (Bertino 2012).

The concept of privacy is a millennia-old topic - Greek philosopher Aristotle described the difference between public and private life as early as the 6th century BC. Specifically, there are four definitional approaches to privacy (Smith et al. 2011):

- Privacy as control (cognate-based definition): privacy is defined as the possibility of selectively controlling access to information.
- Privacy as a state (cognate-based definition): privacy is defined as the resulting state of limited access to information.
- Privacy as a commodity (value-based definition): privacy emerges as a commodity that is assigned a value and can be traded off by cost-benefit calculations.
- Privacy as a right (value-based definition): privacy emerges as guarantees granted by given legal systems, for individuals to be left alone (Warren and Brandeis 1890) and to develop as autonomous selves (Bowie and Jamal 2006).

Table 2 summarizes these approaches to privacy, and links them to IS research topics (according to (Bélanger and Crossler 2011), which we introduce in the following paragraphs.

Table 2. Linking privacy aspects to information systems (IS) and the law

Privacy approaches	Control	State	Commodity	Right
Related information privacy topics according to (Bélanger and Crossler 2011)	Privacy tools and technologies	Privacy practices	Privacy concerns	Privacy regulations
Coverage in academic disciplines: examples	IS: encryption, privacy-preserving tools	IS: Fair Information Practices (FIPs)	IS: Privacy calculus	Law: Data protection

To review the state of privacy research in IS, two prominent review papers on information privacy research in IS were published in 2011 (Bélanger and Crossler 2011; Smith et al. 2011). Both papers investigated a large number of publications (i.e., Bélanger & Crossler reviewed 284 journal and conference papers, and Smith et al., 320 journal and conference papers, as well as 128 books), and came to similar conclusions, namely that IS publications generally study privacy as an individual construct, through the lens of information privacy concerns.

Bélanger and Crossler's (2011) analysis specifically highlights relevant research gaps. First, IS researchers study privacy from a large variety of vantage points and constructs. Yet action and designed-focused research only represents ca. 4% of selected literature items, when most contributions focus on analyzing, explaining, or predicting information privacy concerns and attitudes. Second, research revolving around information privacy concerns (e.g., willingness to transact online), e-business impacts (e.g., willingness to share information with e-merchants and/or e-government), information privacy attitudes (e.g., reaction to privacy invasive technologies), which usually study privacy constructs at the individual level, accounts for ca. 60% of selected literature items. Third, research revolving around information privacy practices, which is the only topic that appears to study these constructs at the level of individuals and organizations alike, accounts for ca. 32% of selected literature items. Smith et al. (2011) draw similar conclusions regarding the level of analysis, with the individual level being the most investigated by a large margin, thus calling for more research to be conducted at the organizational level of information privacy.

2.2.2 The legal perspective on data privacy

Data protection is closely related to the concept of privacy and constitutes the legal view on information privacy (s. Table 2 above). The idea of a right to privacy emerged in the late 19th century, in the minds of Boston-based lawyers Samuel Warren and Louis Brandeis, who described it as a "right to be left alone" (Warren and Brandeis 1890). It was also at a later point in time, during the second half of the 20th century, that modern data protection regulations appeared, first in the German *Land* of Hessen in 1970, and throughout the 1980's and 1990's (e.g., France in 1978, the German Federal Republic in 1979, the United Kingdom in 1984 and Switzerland in 1992).

The emergence of Information Systems in enterprises is usually dated back from the mid-1960's until the mid-1970's (Hirschheim and Klein 2012), and it is no coincidence that the first legislative efforts in the field of data protection were instigated during that same time period, which, according to Westin's (2003) classification, corresponds to the "first era of contemporary privacy development". Before information processing was computerized, fewer organizations held personal data about individuals and performing data analysis manually was a complex task (Meier, 2011) – furthermore, public trust in governments and businesses was high, and data collection was thus not perceived as a threat (Westin, 2003). However, in the absence of proper regulation and with the rise of data collection and automated data processing, individuals started to face challenges associated with an absence of transparency regarding who was processing their data, in which ways, and for what reasons (Meier 2011, p. 63).

These challenges were still relevant in the early 2010s and were even amplified by the development of advanced data processing technologies and the emergence of the data-driven enterprise. To address these challenges, the European Union initiated a rework of its data protection framework that resulted in the adoption of the General Data Protection Regulation (EU-GDPR) in 2016, which entered into force in May 2018. At the European scale, the legal landscape was fragmented, with each member State having their own data protection regulation meant to enforce the guiding principles of the European Data Protection Directive of 1995. They proved insufficient, as organizations could obtain extended consent by embedding broad data processing agreements in their general terms and conditions. As a result, individuals could be asked to provide a "data processing blank check", in the sense that they were forced into opting into such general agreements in order to benefit from an organization's products or services. The updated regulatory framework requires that organizations provide transparency on their data processing activities and let individuals control the data they share and the way it is used with a high level of granularity (De Hert and Papakonstantinou 2012, 2016). Similar requirements have been taken over by other regulatory projects around the world, most notably in in the United States (Rubio 2019), who do not have a long-standing data protection legal tradition. The California Consumer Privacy Act (US-CCPA), which became effective in January 2020, adapts many of the principles that were introduced with EU-GDPR. In Switzerland, the revised Federal Data Protection Act, which was finalized in October 2020, also aligns with EU-GDPR to a large extent (Métille and Raedler 2017).

These new regulatory instruments require changes in the way organizations process personal data, as they now have to document and disclose details about each of their personal data processing activities, and make sure that they can demonstrate compliant processing (De Hert and Papakonstantinou 2016; Mitrou 2017; Nicolaidou and Georgiades 2017).

2.3 Research opportunity

Organizations adopting a data-driven mindset and aiming at generating must open up their data resources beyond technical and organizational silos in order to benefit from new data opportunities. This also implies extending data-driven ways of working to a broad audience of employees, spanning beyond data experts, and addressing the lower levels of data proficiency. The simultaneous introduction of strengthened data protection standards highlights the lack of common ground between the legal and IS domains. From an organizational perspective, as data-related enterprise roles evolve beyond their traditionally accepted boundaries, data management and compliance remain, to this day, separate conversations.

On the one hand, although the benefit of promoting a data culture is starting to be recognized (Awasthi and George 2020; Hyun et al. 2020), few contribution exist on concrete ways to realize data democratization. Academic research extensively investigates the technical aspects and business benefits of advanced data processing methods, but little emphasis is put on making sure that the necessary data is findable, accessible and understandable. These objectives are in line with the evolution of data management practices, and we see a promising avenue in elaborating on how emerging data platforms can support data democratization.

On the other hand, organizations must evaluate their data management practices in light of strengthened data protection regulatory requirements. Yet, legal analysis deviates from the interests and proficiency of most IS researchers (Abdullah et al. 2009; Bélanger and Crossler 2011), and there is a need to analyze new legal requirements and elaborate on corrective measures that concretize strategic compliance objectives (Cleven and Winter 2009).

3 Dissertation overview

3.1 Research objectives

Considering the identified gaps, this thesis attempts to answer the following question:

What are capabilities that organizations need to build to increase usage of data resources enterprise-wide, while complying with data protection regulatory requirements?

As depicted in Figure 1, the thesis comprises two research streams that explore the drivers behind these two objectives (i.e., data democratization and data protection), in order to conceptualize their organizational and technical impact on data management practices.

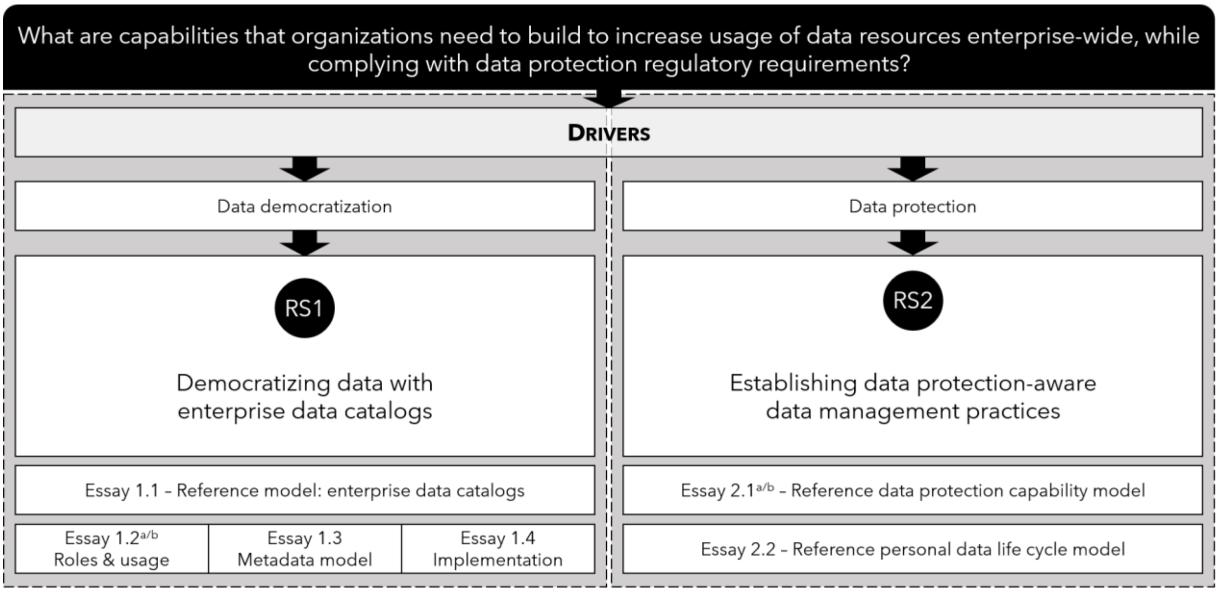

What are capabilities that organizations need to build to increase usage of data resources enterprise-wide, while complying with data protection regulatory requirements?
DRIVERS
Data democratization
Data protection
RS1
Democratizing data with enterprise data catalogs
RS2
Establishing data protection-aware data management practices
Essay 1.1 – Reference model: enterprise data catalogs
Essay 2.1a/b – Reference data protection capability model
Essay 1.2a/b Roles & usage
Essay 1.3 Metadata model
Essay 1.4 Implementation
Essay 2.2 – Reference personal data life cycle model

Figure 1. Overview of research streams and questions

The first research stream aims to outline the enablers of data democratization in enterprises, in terms of data documentation, user roles and supporting platforms. Given the increasing popularity of enterprise data catalogs[4], it aims at clarifying their role as key platforms for data democratization. It comprises four essays that develop a reference model for enterprise data catalogs – they focus on outlining key constituents of enterprise data catalogs, by defining the required data documentation through metadata, the existing and emerging data-related roles in the data-driven enterprise and their specific data needs, as well as outlining data catalog implementation approaches.

[4] In 2020, Forrester Research denotes 27 existing solutions from established vendors (Goetz et al. 2020).

The second research stream aims to close the gap between data protection regulations and the supporting data management practices. It comprises two essays that translate data protection regulatory requirements into a capability model and propose a concretization of regulatory requirements through a data life cycle model, including data model extensions and business rules.

With this thesis, we seize the opportunity to establish a link between these two topics, by analyzing and confronting them in the context of the data-driven enterprise.

3.2 Research setting

This research has been carried out in the context of a consortium research project (Österle and Otto 2010, s. Figure 2) in the field of data management, the Competence Center Corporate Data Quality (CC CDQ). It convenes data management expert from around 20 multinational organizations, active in diverse industrial areas[5], and a team of researchers, in order to address relevant research problems. Consortium research programs are part of the collaborative practice research tradition. It is based on a collaboration between researchers and practitioners and focalizes on serving both groups' interests, by adding to the bodies of knowledge of involved professional and scientific communities alike, as well as by advancing practices in the area of interest (Mathiassen 2002). In that sense, they are a type of practice research (Feldman & Orlikowski, 2011; Goldkuhl, 2012) and have proven to be fruitful when tackling wicked topics (i.e., novel, interdisciplinary, and loosely defined), such as ours, as they provide a favorable environment for the accumulation of knowledge from design-oriented research (Vom Brocke and Buddendick 2006; Winter 2008; Legner et al. 2020).

The CC CDQ is composed of a diversity of organizations – while most of them are multi-national corporations, they vary in size, industry, as well as in their level of data management maturity. Although all representatives from these organizations are experienced data management professionals, their specific job profiles and experiences differ – for instance, working groups typically involve management executives and engineering experts alike. In other words, the CC CDQ gathers participants with similar concerns, albeit bringing various perspectives and experiences to the table. This helps to avoid group uniformity, and promotes the generalizability of the findings. Hence, despite some commonalities, there is natural variation within the group, which allows to study a variety of cases. To further extend the contexts of study and

[5] Over 10 industries are represented in the CC CDQ, e.g., pharmaceuticals, retail, engineering, telecommunications, fast-moving consumer goods, software, automotive, chemistry.

generalizability of findings, research activities also entailed contacts with a multi-pronged network of stakeholders outside of the consortium research program, including other organizations, researchers, as well as corporate and academic experts.

The drivers behind our two research streams, i.e., data democratization and data protection, set in the context of the data-driven enterprise are novel, highly interdisciplinary topics, that draw from computer science, social science, law, IS and management. As this thesis investigates emerging topics in the information systems domain with an enterprise focus, our outcomes primarily consist reference models, which are a specific type of conceptual models (Frank et al. 2014; Vom Brocke 2007) suitable to design and plan complex systems while fostering communication with prospective users and providing a sound basis for system implementation (Frank 1999, p. 695). Reference modeling has been successfully used for the development of enterprise-specific models (Fettke and Loos 2003, p. 35). In particular, they have enabled knowledge accumulation in the data management domain, in the form of data management frameworks and models (Batini et al. 2009; Madnick et al. 2009; Legner et al. 2020). It also integrates well within a design science research design, as reference models are usually developed iteratively, through design and evaluation cycles.

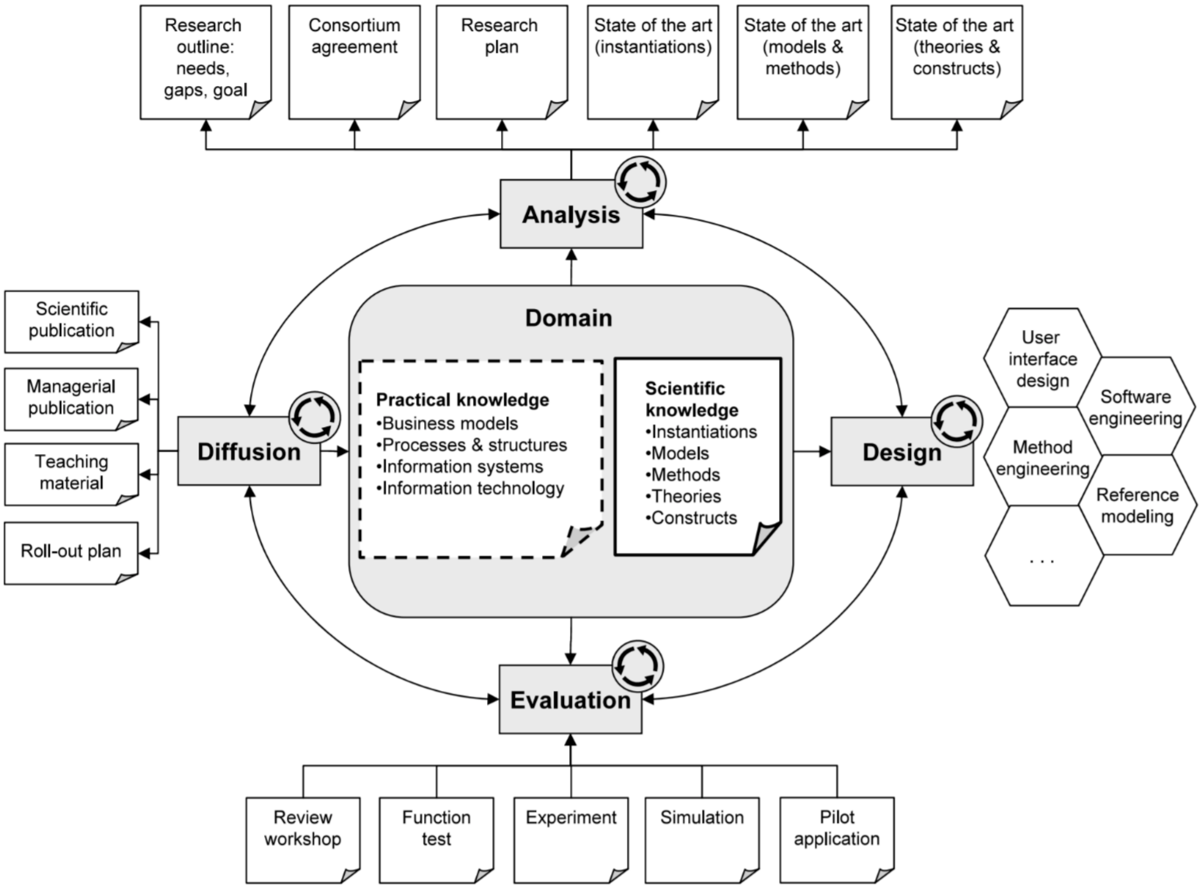


Figure 2. Consortium research overview (adapted from Österle & Otto, 2010).

In that sense, a consortium research program constitutes a favorable environment for design science research, as it provides opportunities for frequent and early iterations with practitioners (Sonnenberg and vom Brocke 2012) to derive innovative artifacts that effectively solve identified organizational problems (Hevner et al, 2004).

Our research process is based on Österle & Otto's (2010) proposed method for researcher-practitioner collaboration in design-oriented IS research and uses Design Science Research (Peffers et al. 2007) as main underlying methodology.

Through the consortium, participant organizations were involved in all stages of the artefact development process. Together with literature review and an analysis of legal regulations, participant feedback was gathered to identify problems and motivate our research outcomes, as well as for defining the objective of planned solutions. The entirety of research stream 1 was elaborated with a stable subset of participating companies and representatives, who were involved throughout the research activities – they provided direct input in design phases, applied the resulting artefacts in demonstration, and provided feedback for evaluation purposes. In the development of research stream 2, there was more variation in participant companies and representatives, and outside stakeholders were also involved to provide feedback along the various steps of the research process. We collected participant feedback by documenting in the following ways (Legner et al. 2020):

- Plenary discussions: used in analysis and design research phases, they enable the review and confirmation of requirements and artefacts. They were conducted for both research streams.
- Focus groups: conducted for both research streams. They took place either in the form of in-person events or online meetings. They typically feature a blend of academic input from the research team (following desk research, s. below), a related discussion with participants, as well as collaborative, design-oriented exercises. They were used in all research phases (i.e., analysis, design and evaluation).
- Expert interviews: consist of one-to-one discussions between subject matter experts in participant organizations and a member of the research streams. They are used for similar purposes and at similar phases that focus groups.
- Projects: consist of intervention of researchers in naturalistic, corporate settings to instantiate and evaluate research results. They were used in research stream 2.

- Case study: consists of qualitative research applied for the exploration and explanation of problems and solution designs. They were used in research stream 1 in design and evaluation phases.
- Survey: consist of data collection through semi-structured or structured questionnaires. They were used in both research streams for evaluation purposes.
- Desk research: traditional research activities (e.g., literature review) conducted by researchers to ground artefact development in relevant academic and executive bodies of knowledge. They were used in both research streams in analysis and design phases.

All essays are based on design science research methodology, except for essay 2.4, which consists of an explorative study based on mixed methods. Table 3 provides a consolidated view of research streams, questions and methods, and outline how each of the constituting essays contribute to answering the thesis's overarching question.

Table 3. Dissertation structure and research streams

Essay	Research question(s)	Research method(s)	Key contributions	Publication status
Research stream 1: Democratizing data with enterprise data catalogs				
Essay 1.1: 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 following (Peffers et al. 2007)	Reference model for enterprise data catalogs	Journal manuscript: Submission for *Business Information Systems & Engineering (BISE)*
Essay 1.2a/b: Data Democratization in Practice: Fostering Data Usage with Data Catalogs	Which data-related roles emerge in the context of data democratization? How do data catalogs support these roles and their typical data needs and usages?	Design science research, following (Peffers et al. 2007)	Role model for data catalogs, with usage and functionality needs, and illustrative collaborative usage scenarios (vignettes)	a: Communications of the *20th Symposium of the Association Information et Management* (2020) – Best Paper Award b: Extended journal manuscript
Essay 1.3: Empowering Data Consumers to Work with Data: Data Documentation for the Enterprise Context	How to organize data documentation to support data discovery and data use in the enterprise context?	Design science research, following (Peffers et al. 2007)	Metadata model for enterprise data documentation in the context of data democratization	Proceedings of the *15th International Conference on Wirtschaftsinformatik* (2020)
Essay 1.4: FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs	How can data catalogs support the FAIR principles in the enterprise context?	Explorative research with mixed methods (Kaplan and Duchon 1988; Venkatesh et al. 2013): Taxonomy development following (Nickerson et al. 2013), case studies	Taxonomy of enterprise data catalog implementation and illustrative case studies	Proceedings of the *22nd IEEE Conference on Business Informatics* (2020)
Research stream 2: Establishing data protection-aware data management practices				
Essay 2.1a/b: Understanding Data Protection Regulations from a Data Management Perspective: A Capability-Based Approach to EU-GDPR	What data management capabilities need to be built in order to address EU-GDPR's requirements?	Design science research, following (Peffers et al. 2007)	Reference data protection capability model	a: Proceedings of the *14th International Conference on Wirtschaftsinformatik* (2019) b: Extended journal manuscript
Essay 2.2: Personal Data Management Inside and Out – Integrating Data Protection Requirements in the Data Life Cycle	What is the impact of data protection regulations on the personal data life cycle? How could data life cycle models be amended in order to address regulatory requirements for data protection?	Design science research following (Peffers et al. 2007)	Reference personal data life cycle model	Journal manuscript: Published in *Enterprise Modelling and Information Systems Architecture – International Journal of Conceptual Modeling* (*EMISAJ*, 2020)

4 Research stream 1: Democratizing data with enterprise data catalogs

4.1 Background

4.1.1 Democratizing enterprise data

When it comes to the democratization of data in the enterprise, it has been defined as "*the act of opening organizational data to as many employees as possible, given reasonable limitations on legal confidentiality and security*" (Awasthi and George 2020). In a similar spirit, researchers have highlighted the importance of promoting a culture of data use (Upadhyay and Kumar 2020) in enterprises.

While organizations may possess a rich pool of data resources, perhaps with a high level of quality, the mere existence of such resources does not automatically translate in added business value. IS researchers have posited that an organization's ability to derive benefits from data resources and analytical insights is depending on its ability to nurture an inclusive data culture (Upadhyay and Kumar 2020), i.e., one that casts a wide next among enterprise stakeholders. The rationale behind this relationship is that firm performance is positively influenced by the transformation of tacit knowledge from data into explicit knowledge, which requires the development of analytical capabilities (Grover at al. 2018; Upadhyay and Kumar 2020). Further studies show that establishing a data culture is a key driver in executing such analytics capabilities (Zheng 2005), and that it acts a catalyzer for the conversion of analytics-related investments to business value (Grover et al. 2018). In addition, establishing a data culture is inherently linked with providing transparency and access to data resources, which has also been identified as mediating factor between analytics capabilities and value generation (Grover et al. 2018). The combination of a data-driven culture and of increased transparency on data resources has been referred to as "democratization culture", as a component of data democratization (Hyun et al. 2020).

As a result of these findings, while a judicious use of data constitutes a competitive advantage (George et al. 2014) , such use needs to be enabled enterprise-wide, by instilling a data culture among and providing access to data resources to enterprise stakeholders that are not inherently, by education or job description, data experts. In that sense, data democratization can be viewed as an enabler of data-driven opportunities, as the promotion of a democratization culture can correlate positively with increased analytics usage and more agile enterprise decision making

(Hyun et al. 2020). Additionally, the various definitions of democratization proposed by the literature express conceptualizations of the processes and related guiding principles towards removing obstacles to a resource (Schlagwein et al. 2017).

However, these conceptualizations do not specify how this democratization process should be realized and which specific tools could support it[6]. In the following, we present the FAIR principles, that support the idea of data democratization, as well as enterprise data catalogs as a suitable candidate to implement these principles in the enterprise context.

4.1.2 The FAIR principles and the enterprise context

In research, a 2016 paper co-authored by more than fifty researchers unveiled the FAIR principles (Wilkinson et al. 2016), according to which data should be findable, accessible, interoperable and reusable – they "describe distinct considerations for contemporary data publishing environments with respect to supporting both manual and automated deposition, exploration, sharing, and reuse." These principles are rooted in the observation that researchers have to spend significant amounts of time in data gathering activities, due to a general lack of properly documented and indexed data sources. Hence, the authors conceptualize the related obstacles and argue that data resources should be made available through repositories, and come together with a comprehensive documentation by means of metadata, relying on interoperable vocabularies and identifiers. Table 4 shows the detailed specifications of each principle.

Table 4. The FAIR principles in detail (from Wilkinson et al. 2016)

Principle	Specification
Findable	(Meta)data are assigned a globally and persistent
	Data are described with rich metadata
	Metadata clearly and explicitly include the identifier of the data it describes
	(Meta)data are registered and indexed in a searchable source
Accessible	(Meta)data are retrievable by their identifier using a standardized communication protocol
	The protocol is open, free, and universally implementable
	The protocol allows for an authentication and authorization procedure, where necessary
	Metadata are accessible, even when the data are no longer available
Interoperable	(Meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation
	(Meta)data use vocabularies that follow FAIR principles
	(Meta)data include qualified references to other (meta)data
Reusable	Meta(data) are richly described with a plurality of accurate and relevant attributes
	(Meta)data are released with a clear and accessible usage license
	(Meta)data are associated with detailed provenance
	(Meta)data meet domain-relevant community standards

[6] Therein lies the main difference between democratization culture, which is a stage to be reached, and data democratization, which also encompasses the process to reach that stage.

First, the *findable* principle addresses the search activity itself, i.e., figuring out what data exists and where it exists. It is about ensuring that users are provided with the means to find data within given repositories. This requires that data is associated with identifiers and is indexed in a searchable source, which presents them with a suitable description of their contents. Once suitable data sources have been identified, the *accessible* principle requires that they are made available to users, e.g., standard interfaces and openly documented APIs. Here, it is about ensuring that users are provided with the means to access the data within the repository. The *interoperable* principle further requires that data should be encapsulated in standardized, commonly used formats. The *reusable* principle points explicitly to the documentation of the data, which goes beyond the description that is comprised in the finable element. Documentation should not only enable users to find and identify data, but to provide all necessary contextual information so that users can understand them (e.g., detailed description about tables, columns, attributes) and put them to use. Table 4 provides an overview of the FAIR principles and their specification.

As the authors point out, the FAIR principles do not constitute a standard, and they do not prescribe a specific solution to solve the issues that they outline – they are formulated in both domain- and technology-independent terms (Wilkinson et al. 2016) and are meant to be applied to the design and implementation of platforms supporting data exploration, sharing and re-use. In doing so, such platforms would address those obstacles and reduce the amount of time and effort that researchers need to invest in gathering data, thus enabling them to focus on their own contributions. Thus far, the FAIR principles have played an important role in the academic world. Since their introduction, researchers have suggested implementation considerations to guide the design of solutions (Jacobsen et al. 2019), and related scientific initiatives are emerging, such as the Internet of FAIR Data and Services (van Reisen et al. 2019). Despite their popularity in research, the FAIR principles have experienced few thorough applications in other settings, including corporate settings (van Reisen et al. 2019). In the enterprise, they lead to the idea of enterprise data catalogs – these emerging platforms aim at linking data supply and demand and are an extension of existing metadata management concepts. From a business perspective, "*a data catalog maintains an inventory of data assets through the discovery, description and organization of datasets. The catalog provides context to enable data analysts, data scientists, data stewards and other data consumers to find and understand a relevant dataset for the purpose of extracting business value*" (Zaidi et al. 2017a). Data catalogs have also been characterized as data management platforms that "*take metadata management from its backwater silos to a* *centralized cross-platform facility that is feature-rich and comprehensive*" (Russom, 2017, p. 3). Therefore, they aim at exposing enterprise-wide data resources to a wide group of users, and provide them with functionalities to leverage them, thus supporting data democratization.

4.1.3 Evolution of metadata towards enterprise data catalogs

Enterprise data catalogs are a result of the evolution of data documentation and data provisioning concepts, established since the early days of data processing in the 1960s (s. Figure 3). Data documentation is most often associated with metadata, that are commonly defined as "data about data" (Roszkiewicz 2010) and "aim at facilitating access, management and sharing of large sets of structured and/or unstructured data" (Kerhervé and Gerbé 1997). The origins of data documentation trace back to field names and table definitions (Uhrowczik 1973), before system-specific data dictionaries emerged during the 1980s. These data dictionaries were used for basic technical documentation of database tables. With the emergence of enterprise resource planning systems, more emphasis was put on business process integration and the system landscapes grew more and more complex (Kumpati 1988). Over time, it became necessary to plan the data architecture as well as data integration more carefully and link them to the business needs (Stock and Winter 2011).

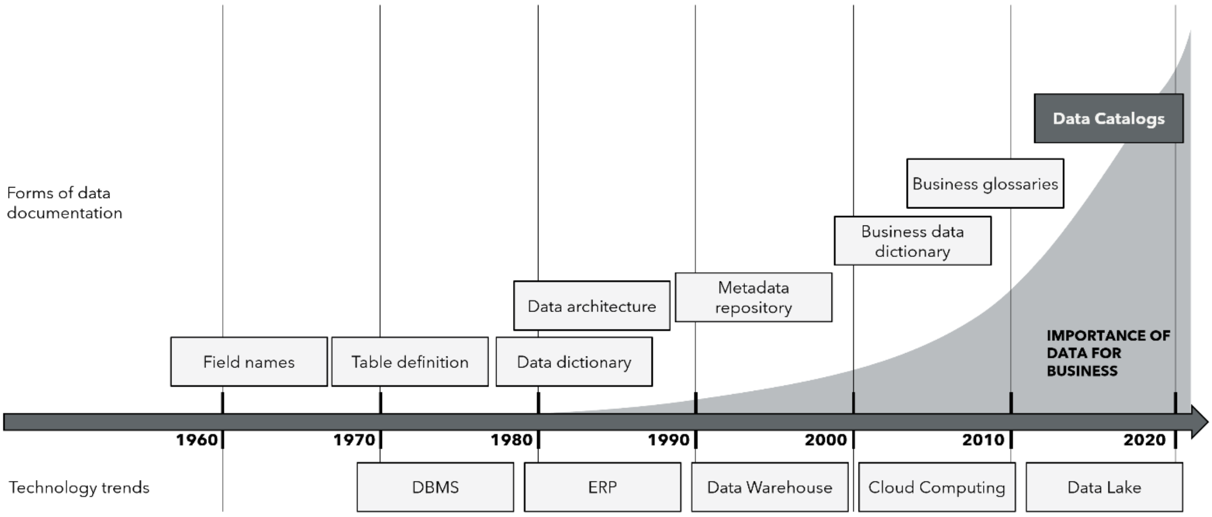


Figure 3. Evolution of metadata technologies towards enterprise data catalogs (adapted from Sen 2004).

This push for a wider-reach of data documentation initiatives has found an echo in the development of commercial metadata management solutions. Even though better data documentation and metadata quality have been shown to result in better corporate decision making (Even et al. 2006; Fisher et al. 2003), data documentation initiatives are complex endeavors that do not always succeed in creating the expected benefits (Shankaranarayanan and Even 2006). While the appearance of the metadata repository concept in the early 1990s gave an impulsion towards the design of enterprise-wide data models (Scheer and Hars 1992), their feasibility proved challenging, which compromised practical applications and uses (Wixom and Watson 2001). Often based on commercial metadata management solutions tightly linked with database management systems, these data documentation initiatives suffered from a lack in adoption and maintenance. This was also due to the fact that these initiatives tended to be decentralized and focused too narrowly on user- or application-specific documentation schemes (Shankaranarayanan and Even 2006) – as summarized by the authors, "*a consequence of adopting a narrow view of metadata while failing to understand the relationships among metadata components is the creation of fragmented "metadata islands." Each island includes metadata of a specific functionality, unaware of and unable to communicate with other islands. Even when system designers and developers understand this complexity, implementing an integrated metadata layer is resource-intensive in terms of money, time, and managerial effort, as well as being a technical challenge.*"

Furthermore, studies from the neighboring knowledge management domain have shown that the adoption of related technologies has been successful only when technology and implementation requirements were in alignment with the culture of the organization (Arpaci 2017; Chen 2010), an aspect that was typically overlooked due to the strong technical focus of traditional metadata management commercial solutions, de facto reducing the target user scope.

Enterprise data catalogs are poised to address these challenges thanks to their emphasis on enterprise-wide integration of data resources and on a broad target user scope. As organizations aim at extracting value from data with advanced analytics technologies, indexing and documenting said data assets, as in a catalog, becomes critical. From this perspective, enterprise data catalogs can be viewed as the next step in the evolution of metadata concepts.

Due to its novel nature, an academic conceptualization of the term "enterprise data catalog" is still missing. The first definition originates from the computer science community. Researchers made the observation that data landscapes evolve from standalone databases to a heterogeneous set of systems to store and analyze data (Franklin et al. 2005). They argue that in spite of being stored across a variety of systems (i.e., multiple versions of the truth), data still need to be managed as though they were stored in a single database (i.e., single version of the truth). Based on these assumptions, they define a data catalog as follows: "A catalog is 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). Coming from the computer science domain, this definition mainly considers infrastructure aspects, but is less explicit when it comes to data usage aspects.

In 2020, enterprise data catalogs are considered a new category of software solutions – they are recognized by Gartner among the top 10 trends in data and analytics for 2020, as an "augmented data management" solution, extending the scope of metadata management capabilities (Sallam et al. 2020). Market analysts denote more than 20 related offerings (Goetz et al. 2020), including products from leading software vendors such as IBM, SAP, Oracle and Informatica. While analysts highlight that enterprise data catalogs enable the discovery and understanding of datasets by different user groups (Zaidi et al. 2017b), the market is still in its infancy, and there is no common understanding of their functional scope, target users and usage purposes.

4.2 Research objectives and approach

The goal of the first research stream is to provide an understanding of the enterprise data catalog concept and to clarify its role as key platform in data-driven enterprise system landscapes. It aims at establishing the key constituents of enterprise data catalogs, by defining related requirements for data documentation, outlining target user roles and usage needs, and understanding implementation characteristics. The emphasis on data democratization and on user requirements is in line with findings from literature that showcase the link between data culture, data transparency and value generation (Grover et al. 2018; Upadhyay and Kumar 2020) and highlight the lack of alignment between such culture and technology requirements in previous, unsuccessful data documentation attempts (Shankaranarayanan and Even 2006; Vnuk et al. 2012).

Due to the lacking conceptualization of EDCs in academic literature and in practice, we set out to develop a reference model as a proxy to define the concept EDC. We argue that EDCs are emerging platforms that enable the realization of the FAIR principles in an enterprise context. By nature, the FAIR principles are meant to facilitate the data gathering process by removing obstacles to finding, accessing, understanding and re-using data, hence supporting data democratization. In the enterprise, as all relevant data resources are stored in the organization's IT systems, platforms that act as abstraction layer of storage systems while providing built-in usage functionalities are suitable candidates to tackle these obstacles. Yet, few academic studies have investigated the operationalization of the FAIR principles in an enterprise context, and EDCs are yet to be understood as an emerging category and cornerstone of future software IT landscapes.

To investigate the challenges, goals and implementation characteristics, we formed a dedicated group as a subset of the consortium research program. It was composed of 17 representatives from 13 multi-national participant organizations, all of which had started EDC implementation initiatives. The experts that joined the group were overseeing implementation initiatives or were closely involved with key implementation aspects. Thus, in line with the focus group method, the participants had a distinct knowledge of data catalog implementation and had been involved in their companies' adoption initiatives (Bryman and Bell, 2007, p. 511; Creswell, 2009, p. 181). Table 5 provides an overview of participant companies and experts. The group worked together for approximately 24 months, during which three full-day meetings took place, alongside 10+ shorter focus group sessions, web conferences and one-to-one interviews.

Table 5. Overview of companies participating in the data catalog research group.

Company	**Industry**	**Revenue range**	**Expert (title)**	**Catalog purpose**	**Status**
A	Adhesives	1 – 50 B €	Lead Data Architect	Metadata management	Rollout and onboarding
B	Automation	1 – 50 B €	Head of Corporate Master Data Management	Support for data governance	Tool selection
C	Chemistry	50 – 100 B €	Data Catalog Product Owner and Enterprise Architect	Support for data governance and analytics	Rollout and onboarding
D	Fashion and jewelry	1 – 50 B €	Data glossary manager	Data glossary	Rollout and onboarding
E	Information technology	1 – 50 B €	Solution Advisor Expert	Support for data governance, metadata management and data analytics	Continuous usage and maintenance
F	Manufacturing	1 – 50 B €	Corporate Data Management	Metadata management	Rollout and onboarding
G	Manufacturing	50 – 100 B €	Enterprise Architect for IoT and Digitalization	Metadata management	Pilot
H	Packaging	1 – 50 B €	Global Master Data Driver	Support of data governance, analytics, inventory and automation	Scoping and tool selection
I	Pharmaceuticals	1 – 50 B €	Associate Director Information Management	Support for data analytics	Implementation in progress
J	Pharmaceuticals	1 – 50 B €	Business Data Analyst, Global Data Team, Ops IT	Support for data governance and data analytics	Rollout and onboarding
K	Retail	100+ B €	Team Lead Master Data Management	Support for data governance	Continuous usage and maintenance
L	Tobacco	50 – 100 B €	Manager Enterprise Data Governance System	Support for data governance	Continuous usage and maintenance
M	Sportswear	1 – 50 B €	Director of Tech Consultancy, Lead Solution Architect	Support for data analytics	Rollout and onboarding

4.3 Contributions

4.3.1 Reference model for enterprise data catalogs

To the best of our knowledge, the EDC concept is mainly discussed among practitioners (Russom 2017; Zaidi et al. 2017a) and a rigorous definition as well as conceptualization is missing. Therefore, essay 1.1 elaborates on the general understanding of enterprise data catalogs, and addresses the following questions: *What are the main constituents of an Enterprise Data Catalog as emerging platforms for data democratization?*

Based on our review of prior literature, as well as existing EDC solutions, we position 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).

Libraries have always played an important role in democratizing information to a large audience (Wallace and Van Fleet 2005). Based on this finding, we identify two prior concepts which support the data democratization process and pursue purposes similar to those of enterprise data catalogs. First, the digital library which supports the FAIR principles in research communities (Wilcox 2018). Second, the dataspace which aims to make interrelated data findable and accessible across distributed databases (Franklin et al. 2005). While digital libraries have a strong focus on making digital scholarly material, e.g., textual content or research data, accessible for the scientific community, the concept of a data space provides the technical foundations for data democratization, regardless of the context. In the enterprise, these principles find an echo in data curation, which is 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). In an enterprise context, such purposes may include, reporting, self-service business intelligence, governance and data monetization support, e.g., data-driven decision making. Due to the distributed nature of enterprise systems and the multiplicity of storage locations and processing applications (Roszkiewicz 2010), enterprise users typically face similar difficulties as researchers when attempting to gather data for data-driven purposes. These obstacles become even more concerning in the context of data democratization – as employees with little data-related background, training and expertise are incentivized to work with data, data search and gathering activities should be as streamlined as possible.

Rooted in existing IS architecture conceptualizations (Chang et al. 2007; Scheer 1992; Scheer and Schneider 2006) and building on DL and DS literature, the resulting reference model for EDCs synthesizes their key constituents and distinguishes 3 different views: the organization view, the data view, and the function view.

Figure 4 provides an overview of the relevant views on EDCs that make up in the reference model:

- The organization view role mode draws from data democratization and data governance literature to define typical users of EDCs. It also outlines their individual needs and usages. Essay 1.2 introduces the role model in detail.
- The data view consists of a metadata model outlining essential data objects and relationships that need to be documented as a foundation for EDCs. It is based on an extensive review of existing metadata models and is introduced in detail in essay 1.3.
- The function view[7] introduces key functionalities of EDCs and groups them into function trees.

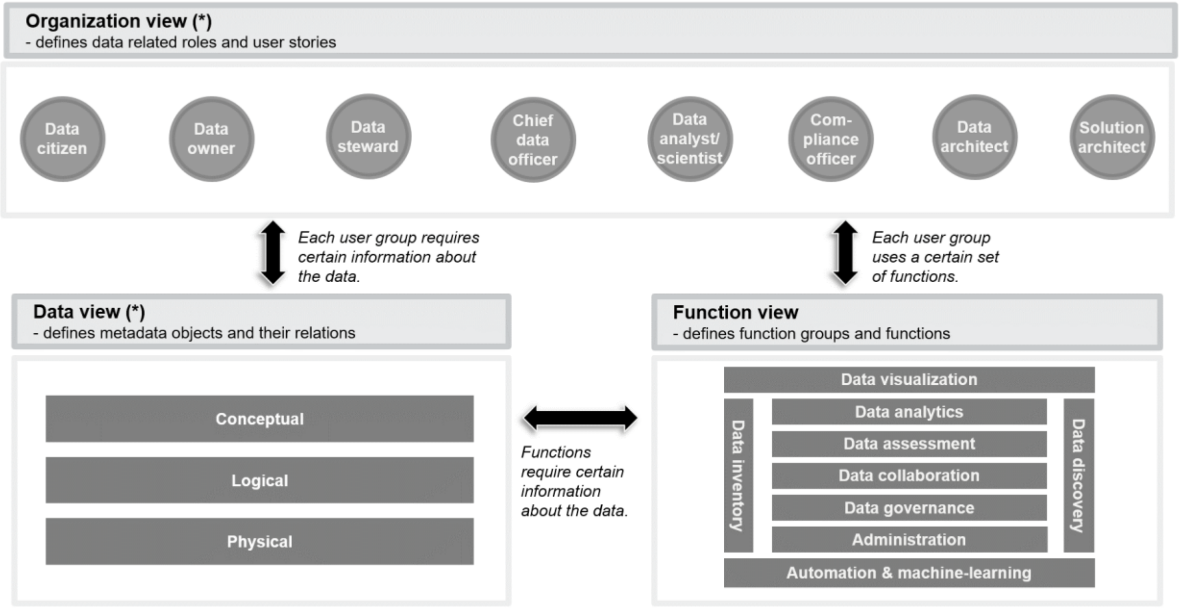

Organization view (*)
- defines data related roles and user stories
Data citizen
Data owner
Data steward
Chief data officer
Data analyst/ scientist
Com-pliance officer
Data architect
Solution architect
Each user group requires certain information about the data.
Each user group uses a certain set of functions.
Data view (*)
- defines metadata objects and their relations
Conceptual
Logical
Physical
Functions require certain information about the data.
Function view
- defines function groups and functions
Data visualization
Data analytics
Data assessment
Data collaboration
Data governance
Administration
Data inventory
Data discovery
Automation & machine-learning

Figure 4. Reference model for enterprise data catalogs ((*) denotes views that are contributions from this thesis)

In designing the EDC reference model, as well as its constituent views, we followed the design science research method outlined by Peffers et al. (2007), with an *Objective-Centered Solution*

[7] The function view was developed by another researcher from the consortium research program.

entry point, as we EDCs are existing, yet loosely defined solutions that address a clearly defined problem (data democratization).

4.3.2 Organization view: identifying user roles and their requirements for data democratization

The idea of data democratization denotes that more users are involved with data, beyond traditional data experts. However, we lack a precise understanding of these new data users and their data-related needs. Hence, essay 1.2 addresses the following research questions:

- Which data-related roles emerge in the context of data democratization?
- How do data catalogs support these roles and their typical data needs and usages?

To answer these questions, in the organization view, we outline a set of roles that reflects EDC requirements within an organization from a user perspective. User roles revolve around the general purposes of data democratization to support data supply, demand, and curation (Borgmann 2003, Lord et al. 2004, p.1). These three perspectives are mirrored by three data-related role categories found in research: data collectors, data consumers, and data custodians (Lee and Strong 2004). On the supply side, data collectors designate people responsible for collecting and inventorying data resources into an EDC. Hence, they model, maintain and create data to be referenced and documented within the EDC. They consist of data architects, and solution architects. From a curation perspective, data custodians work with data that has been integrated into the system thanks to data collectors and make sure that it is fit-for-use. For instance, data owners oversee a specific data domain and maintain related definitions, while data stewards use the EDC to assess and document various aspects of datasets (e.g., quality, maturity, usability), supporting data demand by generating relevant information. On the demand side, data consumers use data to support business purposes and several stakeholders interact with the EDC with various data requirements. For instance, data citizens need to find data and understand data practices, and data analysts require precise data documentation to derive relevant data insights. 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 in the enterprise.

The organization view aims at defining the user roles and purposes that EDCs are meant to serve. Based on literature, expert feedback, and insights from EDC projects, we have identified eight user roles that might benefit from an EDC. We present these user roles together with exemplary user stories (see Table 6). They help understand user needs and provide a link with EDC functions and function groups. We also consolidate these user stories through the lens of data-related collaborative use case scenarios, and present them as vignettes. Taken together, the roles, user stories and use case scenarios show that data catalogs go beyond data documentation and provisioning – through user collaborations and workflows, they improve access to and understanding of data assets, facilitating business value generation. Our contributions thus shed light on existing and emerging data-related roles in data democratization, their requirements, as well as data-related collaborative scenarios.

Table 6. Roles, user stories and related EDC functions

Roles	User stories	Functionality needs
Data Citizen	Understand how to correctly enter data in 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 to data)*
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*
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 to data)*
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*

Roles	**User stories**	**Functionality needs**
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*
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 & ML: *automated classification / tagging* Data Inventory: *metadata management* Data Discovery: *search*
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*
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*

4.3.3 Data view: documenting data for discovery and usage

Enterprise system landscape are often distributed and complex. Because data is stored in multiple databases, there is a need to establish a single version of the truth (Franklin et al. 2005; Van Alstyne et al. 1995). In this context, metadata can be used to provide an abstraction between objects, their storage instances, and their conceptual meaning. While data dictionaries and business glossaries provide definitions to foster the understanding of data, a broader data documentation perspective is needed to support data discovery and usage. Essay 1.3 addresses the following research question: How to organize data documentation to support data discovery and data use in the enterprise context?

In the data documentation view, we answer this question with a metadata model that identifies 22 metadata objects for EDC data documentation, catering to the need of user roles (s. Table 7).

Table 7. Information model overview

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

It was developed based on a review of four existing, domain-agnostic metadata standards: 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). Dublin Core presents a flat list of terms (or attributes), comprising, e.g., creator, description, date, identifier, relation and rights. DCAT and MDR go a step further by grouping attributes and specifying relationships between groups in a metamodel, which focuses on the internal logic between the concepts they introduce, but do not integrate external concepts. Finally, CWM provides a metadata interchange standard that enables XML-based exchange of business intelligence and data warehouse metadata between different tools, platforms and repositories.

In our research, we develop a metadata model (Kerhervé and Gerbé 1997), comprising relevant metadata objects that are to be documented as well as their relationships and reconcile both business- and system-oriented perspectives on data. Hence, following common data modeling principles (Batini et al. 1986; Tsichritzis and Klug 1978), the metadata model covers three modelling layers: conceptual, logical and physical. The conceptual layer provides high-level and non-technical concepts describing enterprise-wide data, the logical layer represents the business view on data and makes the link between the conceptual layer and the physical layer, which represents the lower-level implementation perspective and makes the link with the way data is implemented in systems. Furthermore, as it is meant to provide data documentation for the user roles (according to the role model), we have divided the conceptual layers into several views that further specify needs and requirements of both technical and non-technical data users, in line with the objective of data democratization.

4.3.4 Role of enterprise data catalog and implementation approaches

In addition to outlining key constituents, we set out to understand the role of EDCs in operationalizing the FAIR principles in enterprises, as well as related implementation perspectives. With, essay 1.4 we ask following research question: How can data catalogs support the FAIR principles in the enterprise context?

To assess implementation statuses and identify patterns, we developed a taxonomy that enabled the identification of three distinctive implementation profiles, which were documented as case studies. The taxonomy was developed iteratively using the guidelines for taxonomy development (Nickerson et al. 2013), and integrates components from the EDC reference model. Drawing on the reference model and additional empirical input, it comprises 19 dimensions, clustered along five meta-characteristics:

- scoping and goals,
- user groups,
- functionalities,
- data documentation,
- tools.

Based on the taxonomy, we asked 11 companies to perform a self-assessment of their own implementation approach. which enabled us to obtain a consolidated view highlighting common patterns. We found that a majority of organizations viewed enterprise data catalogs as a way to achieve enterprise-wide data transparency. We confirmed the broad user scope that enterprise data catalogs should address, which was confirmed by the conceptual data layer being cited as the most critical one. Data inventory, discovery, governance and collaboration were highlighted as most desired functionalities, further positioning enterprise data catalogs as key abstraction platforms for enterprise-wide data. Building on these patterns, three cases illustrating prototypical data catalog implementation approaches were identified and documented. They provide additional empirical insights and shed light on the difficulties in involving a broad audience of enterprise users with data. Combining findings from the taxonomy of implementation initiatives and the specific cases contributed to further defining the implications of the FAIR principles in the enterprise context, as per essay 1.4's research question, which Table 8 summarizes.

Table 8. Implications of the FAIR principles in academic and enterprise contexts.

FAIR element	Research context	Enterprise context
Findable	Repositories with search functions	Abstraction platform for enterprise-wide data
Accessible	Repositories store data resources or provide updated links to them. Repositories support the identification of users if required for use of sensitive data	Enterprise systems that store data are interfaced with the platform, linking the physical, logical and conceptual data layers Approval processes and/or access rights are implemented for sensitive data
Interoperable	Use of standardized formats References to other data	
Reusable	Rich documentation, relevant to intended users	

4.4 Implications, limitations, and outlook

The objective of research stream 1 was twofold, as it aimed at defining data democratization and its ties with the FAIR principles in the enterprise context on the one hand, and positioning enterprise data catalogs as emerging platforms for data democratization, on the other hand. By combining these two perspectives, we contribute to the ongoing academic discourse around data openness (Schlagwein et al. 2017), by elaborating on the data democratization process (i.e., towards increased data usage by an extended scope of employees), and on the resources to trigger these effects (i.e., rich data documentation through enterprise data catalogs).

By synthesizing our findings in a reference model, we provide a conceptualization of enterprise data catalogs in the form of an architecture model, anchored in existing contributions on digital libraries, data spaces, and metadata management. In doing so, we clarify the concept of enterprise data catalogs as key components of future enterprise data management system landscapes. We also outline implementation approaches from an analysis of real-world cases, thus outlining the central role and significance of enterprise data catalogs for data democratization in enterprises.

However, we have to acknowledge limitations. Even though we address the user perspective through roles and related usage needs as a goal-oriented view, we do not consider a process-oriented view, i.e., making sure that target users of enterprise data catalogs actually make use of the tool. This issue was specifically highlighted by the Albaco case. Future research could, on the one hand, analyze adoption patterns to outline key aspects of user onboarding, training and engagement. On the other, even though our usage needs were derived in collaboration with participants who were implementing enterprise data catalogs, it could be argued that their point of view is a managerial one, and that more empirical evidence should be collected from end users themselves. In that regard, future studies could rely on body of knowledge on technology acceptance (e.g. the technology acceptance model (Davis 1989), the unified theory of acceptance and use of technology (Venkatesh et al. 2003)) to further analyze user requirements, and investigate enterprise data catalogs perceived usefulness, ease of use and acceptance. We also acknowledge that this research centers on the enablers of data democratization – our perspective is limited to establishing an understanding of the wider use of data in organization but does not substantiate the relationship between data documentation and wider data access on the one side, and business value creation on the other side. Additionally, from a methodological perspective, although our design science research process features a sample population of 13 large organizations from various industries, the understanding of data democratization could be extended to other types of organizations.

In that regard, avenues for future research include, from the one hand, further analysis of the role of data democratization in the value creation process, including qualitative and quantitative studies investigating whether achieving a high-level of data democratization translates into higher data and business value. On the other hand, the findings of this thesis could be applied to smaller enterprises, non-profit organizations and governments as well, to gain an understanding of the enablers and benefits or data democratization beyond multi-national organizations.

5 Research stream 2: Establishing data protection-aware data management practices

5.1 Background

5.1.1 Privacy and compliance literature

In extending data usage throughout the enterprise, organizations should ensure that both existing and emerging data-related activities are compliant with data-related regulations. Following high-profile corporate scandals in the early 2000s, lawmakers drafted governance-related regulations to control organization's practices (Abdullah et al. 2009; Cleven and Winter 2009). Some regulatory compliance requirements apply to virtually all large organization (e.g., the Sarbannes-Oxley Act (SOX) for corporate governance), while others are industry specific (e.g., Basel Accords for banking, Identification of Medicinal Products (IDMP) for pharmaceuticals, the Health Insurance Portability and Accountability Act (HIPAA) for healthcare).

Although data protection regulations have not been a main focus in IS research, compliance with data protection regulations is an emerging topic, it can be linked to the broader Regulatory Compliance Management (RCM) research domain. RCM is defined as "the problem of ensuring that enterprises (data, processes, organization, etc.) are structured and behave in accordance with the regulations that apply, i.e., with the guidelines specified in the regulations" (El Kharbili 2012). It features abundant contributions that have been summarized in two review studies (Abdullah et al. 2009; Cleven and Winter 2009). While related contributions provide valuable insights on the conceptualization of regulations for the IS domain, they usually do not investigate ways to operationalize specific regulations.

In the data protection domain, this gap could be explained by the fact that, prior to EU-GDPR, given the previous EU Data Protection Directive's failure to harmonize Europe's data protection legal landscape (Poullet 2006), this topic was unattractive to researchers outside of the domain of comparative law. IS researchers would have needed to gather regulations from a various countries, gain an understanding of each country's legal framework, and translate each regulation, as a prerequisite to any meaningful analysis – in most cases, such an endeavor is not conceivable. As EU-GDPR is the first regulation that provides a single data protection legal framework for all EU countries, it has seized the interest of IS researchers. In 2018, a query with the keyword "GDPR" on the AIS Electronic Library only returned 27 matches. This number has been multiplied by 10 within the last two years – the same query returns 262 matches as of November 2020, which we have reviewed and classified (s. Essay 2.1). This ongoing influx of IS contributions revolving around a legal instrument answers the call by Bélanger and Crossler (2011) and Smith et al. (2011) to conduct privacy research at a level other than the individual one. As a product of social interactions (Aubert 1998, p. 21), the legal aspect of these contributions places them at the group or at the societal level.

While the majority of these new studies has a narrow scope on one of EU-GDPR's requirements, those that consider the overall regulation tend design dedicated solutions to tackle compliance. They investigate, for instance blockchain-based personal data management solutions, evaluate existing practices, or analyze EU-GDPR's impacts on a specific domain (e.g., social media discourse, innovation, Big Data). There are two shortcomings in these approaches: first, most papers take the compliance requirements for granted and directly look into specific practices or solutions. Second, these studies do not provide insights into the entire regulation's implications on data-related practices from an enterprise-wide perspective. Hence, we are still lacking a broader and solution-agnostic understanding of data-related practices and the required changes with EU-GDPR and similar regulations – from this perspective, the gaps identified by Abdullah et al. (2009) and Cleven and Winter (2009) remain, within the data protection regulatory context, and within the scope of IS research.

Legal research on data protection constitutes another significant academic avenue but adopts a vantage point that differs from these identified gaps. Legal research on data protection sits above regulations, in the sense that it analyzes and questions their underlying mechanisms (e.g., Lazaro and Le Métayer 2015; Rallet et al. 2015) and the way they implement established privacy principles (e.g., Puyraimond 2019). IS research on data protection, on the other hand, sits below regulations, in that they are considered pre-existing frameworks that are meant to be operationalized through IS artefacts, or serve as benchmark for their evaluation. In other words, legal research studies the theoretical and conceptual underpinnings of regulations, whereas IS research focuses on implementation and operational aspects. For instance, around the topic of consent, legal studies challenge the way consent was implemented in EU-GDPR (Armingaud and Ligot 2019) or the relevance of the concept itself (Zanfir 2014), while IS studies suggest technical solutions for the collection of consent (Bergram et al. 2020; Huth et al. 2020; Maunula 2020) or evaluate whether existing implementation actually comply with the regulation (Kurtz et al. 2020).

Legal research on data protection can be linked to privacy research in IS in that, in evaluating the appropriateness of the regulatory implementation established privacy concepts, they tend to evaluate them from the point of view of the impacts on individuals (Fellous-Sigrist 2018; Solove 2013). From this point of view, they are similar to the traditional privacy studies in IS, which analyze privacy from the angle of individual preferences and behaviors. An overview study from Li (2012) synthesizes the main theories around privacy behaviors that have been developed by the IS community, and suggests measures that digital enterprises can implement to mitigate individual privacy concerns. While this approach is orthogonal to the goal of this thesis, which studies existing legal privacy-related frameworks instead of user perceptions, such studies could nurture or complement legal discussions around EU-GDPR and similar regulations.

When it comes to IS artefacts implementing data protection regulatory requirements, legal textbooks that synthesize court rulings and legal doctrine, as well as official guidelines from supervisory authorities may be used as design and evaluation instruments.

5.1.2 Personal data

From a regulatory perspective, personal data can be defined as "data enabling direct or indirect identification of a single physical person, data that is specific to a single physical person without enabling identification, data that can be linked to a physical person, data regarding which anonymization techniques cannot completely mitigate the risk of re-identification" (Debet et al. 2015). In practice, most companies collect personal data about their customers, and it is often referred to as consumer or customer data. In that regard, it can be defined as "a set of data that represents and is associated with the identity, activities and service offering associated with a unique individual" (Tapsell et al. 2018). The aspect of service offering is prevalent in the consumer/customer data literature and has been emphasized in the broader customer relationship management (CRM) field. In CRM, customer data is considered as an opportunity to understand the customer and co-create customer value (Payne and Frow 2005). The related contributions focus on collecting, organizing, and using customer data in order to build long-term relationships with customers.

5.1.3 New data protection regulations

In January 2012, the European Commission published a proposal for an overhaul of data protection law within the European Union, which formally marked the launching of negotiations towards what would become EU-GDPR, four years later. In this document (European Commission 2012), the Commission acknowledged that increase information sharing (on the consumer-side) and processing (on the enterprise-side) posed new challenges that the guidelines and principles from the existing Data Protection Directive (95/46/EC) were unable to correctly address (Mitrou 2017; Nicolaidou and Georgiades 2017). The Commission also acknowledged that the harmonization objective of the directive had not been successfully met. This was mainly due to the very nature of this legislative instrument and led to a fragmentation of data protection rules among member states, which placed organizations and individuals alike in a situation of legal uncertainty. At a glance, EU-GDPR requires that organizations continuously document:

- the personal data they hold (scope – e.g., list of recorded attributes),
- how it was acquired (origin – e.g., online form, e-mail),
- how it is processed (modalities – e.g., advanced analytics),
- to what end (purpose – e.g., targeted marketing),
- as well as who it is shared with (transmission – including third parties such as cloud services providers).

This information should be available for disclosure to authorities (describing an organization's overall data processing practices) and individuals alike (e.g., when exercising the right of access) at any time. Organizations must also overhaul the way they expose their data processing activities (European Data Protection Board 2018a).

As a prerequisite to storing and processing any personal data, EU-GDPR requires organizations to have a valid legal basis for that processing activity. The law provides six legal bases for processing (EU-GDPR, art. 6, paragraph 1, letters a through f):

- a vital interest (e.g., medical emergency services would be allowed to collect data on a person's blood type to safeguard their physical integrity),
- a legal requirement (e.g., baking regulations require that financial institutions collect extensive data about their customers' tax situation),
- a public interest (e.g., public business registers make personal information about corporate representatives publicly available for transparency and trust in business),
- performance of a contract (e.g., an e-commerce merchant cannot fulfil their contractual obligations if they do not collect shipping and payment information from their customers),
- consent (i.e., additional processing purposes for which organizations must collect explicit and case-by-case authorization from individuals),
- a legitimate interest (e.g., purposes related to data administration constitute legitimate interests according to EU-GDPR, recital 48).

Consent management has received considerable publicity in the context of EU-GDPR, as it has numerous implications on organizations' informational duties, and requires them to collect additional data points to be able to prove individual consent (Karjoth and Langheinrich 2019). However, legitimate interest is being increasingly debated among legal scholars, as the indeterminacy inherent to such legal concepts is triggering wrongful corporate behaviors, calling for a more detailed understanding of the concept's contours (Armingaud and Ligot 2019; Puyraimond 2019), which has yet to be established by jurisprudence, given EU-GDPR's young age. When consent is applicable, organizations must explicitly seek authorization from individuals for each processing purpose and make sure that it can be updated and withdrawn (European Data Protection Board 2018b). Under the former legal framework, organizations could obtain extended consent by embedding broad data processing agreements in their general terms and conditions.

One of the differences between the current and future European legal frameworks can be summarized as follows: the former focused on establishing principles that organization should implement in their privacy policies, and the latter goes a step further by requiring that individuals have the ability to control the data they share and the way it is used with a high level of granularity (De Hert and Papakonstantinou 2012, 2016). All of these evolutions constitute a paradigm shift in data protection, towards greater choice and sovereignty for individuals, and more accountability for organizations (De Hert and Papakonstantinou 2012, 2016; Mitrou 2017; Nicolaidou and Georgiades 2017).

EU-GDPR constitutes a landmark regulation for data protection in the EU and has ushered similar regulatory pushes in other parts of the world. In Europe, Switzerland has finalized an overhaul of its data protection legal framework – after several delays, it is set to be enforced in the beginning of 2022 and is expected to incorporate the majority of EU-GDPR's requirements (Métille and Raedler 2017). In 2017, China introduced its cyber security legislation, which covers data protection aspects such as personal information protection and rules for transnational data transmission. In 2018, following a supreme court judgment that declared privacy a fundamental right, India introduced a draft for a Personal Data Protection Bill (Parliament of the Republic of India 2018), with the objective of acting as a reference template for developing countries to introduce similar regulations (Palanisamy and Nandle 2018). The United States of America still does not have a single, general data protection regulation. Instead, several sector-specific laws co-exist, such as the Children's Online Privacy Protection Rule, the Federal Privacy Act (which only applies to federal agencies), and HIPAA (introduced in 1996, it contains requirements similar to EU-GDPR's, but is restricted to health-related data). Since the Facebook-Cambridge

Analytica data scandal of 2018, there have been calls for a federal EU-GDPR-inspired data protection regulation (Rubio 2019).

Although these regulations originate from different legislative bodies, they all address the same issues, and some are directly inspired by EU-GDPR. Therefore, even if their requirements are positioned at differing levels of severity, the underlying concepts (such as personal data, data processing, consent, organizational and technical measures, and processes) remain the same, allowing for comparisons.

The most prominent example is US-CCPA (California State Senate 2018), which became effective on January 1st, 2020 in California. While not fully identical, US-CCPA carries over many of the requirements and principles from EU-GDPR - Table 9 provides an overview of these similarities.

Table 9. Mapping of data management-relevant requirements in EU-GDPR and US-CCPA

Requirement	**EU-GDPR**	**US-CCPA**
Right of information	Art. 7, 13, 14	§1798.100
Right of access	Art. 15, 18, 20	§1798.110 §1798.115
Right of deletion	Art. 15, 17	§1798.105
Right of rectification	Art. 7, 16, 21	N/A
Right of restriction	Art. 18	N/A
Right of consent	Art. 7, 8, 22	§1798.120
Documentation accountability requirement	Art. 19, 24-30	§1798.130
Authorization accountability requirement	Art. 5, 6, 9	§1798.130

The introduction of EU-GDPR has generated interest in the IS community, but related contributions generally explore specific aspects of the regulation in detail – to the best of our knowledge, there are only few contributions investigating ways to operationalize EU-GDPR as a whole. One of them (Russell et al. 2018), proposes an assessment of organizations' propensity for change, but does not investigate the implication of compliance requirements on data management practices, and a study by (Addis and Kutar 2018), provides a country-specific, overall readiness assessment. We notice that none of these contributions propose artifacts aimed at assessing an organization's overall compliance state and at providing guidance for corrective measures.

5.2 Research objectives

According to legal positivism, law is a product of social interactions (Aubert 1998, p. 21). Specifically, data protection law sets privacy-related regulatory constraints applicable to both individuals and organizations. In other words, legal privacy requirements could be studied on a multi-level basis. Smith et al. (2011) suggest that privacy concerns are a proxy aimed at measuring privacy. Following this logic, we argue that data protection is a proxy aimed at enforcing privacy and as such, falls within the information privacy practices topic area identified by Bélanger & Crossler (2011).

This potential lead for interdisciplinary research becomes relevant in the case of EU-GDPR, which constitutes an opportunity for impactful research in information systems. Prior to the introduction of EU-GDPR, researchers interested in data protection law would have had to consider a large number of regulations (e.g., one per country), most of which might not have been available in a language they understood. On the other hand, restricting studies to a single country would have greatly undermined their impact and relevance. As EU-GDPR alleviates the barriers of geographical fragmentation and language specificity, and has served as inspiration for further legal instruments, it provides an adequate basis to reach a general understanding of modern data protection regulation.

Research stream 2 proposes two complementary reference models that aim at developing such an understanding from a data management perspective. Anchored in the resource-based view theory (RBV), the first model utilizes the capability concept as an interface between abstract compliance requirements and their concretization. The second model is based on the data life cycle concept – it synthesizes existing data life cycle models and extends them according to data protection regulatory requirements, outlining necessary data management steps and data objects to be considered. In order to provide system support, the model is complemented by execution semantics in the form of business rules.

As this research stream adopts an operational perspective, it is predominantly grounded in IS literature on data protection regulations. However, legal sources informed the development of both models, ensuring a proper fit with legal requirements. For this purpose, we gathered and analyzed material from authoritative data protection sources, such as textbooks originating from multiple legal traditons, e.g., pan-European (European Union Agency for Fundamental Rights et al. 2018; Synodinou et al. 2017, 2021, 2020; Voigt and Von Dem Bussche 2017), French (Bensoussan et al. 2018; Debet et al. 2015), Belgian (Docquir 2018) and Swiss (Meier 2011), as well as two recent doctoral thesis monographs (Staiger 2017; Thélisson 2020) . We complemented this understanding with insights from official guidelines and interpretations from supervisory authorities (e.g., Chatellier et al. 2019; Commission Nationale de l'Informatique et des Libertés n.d.; European Data Protection Board 2017, 2018a, 2018b; European Data Protection Supervisor 2018, 2019; Information Commissioner's Office 2017), as well as academic papers and doctrinal opinions (e.g., Armingaud and Ligot 2019; Castets-Renard 2019; Cheffert 2018; De Hert and Malgieri 2018; De Hert and Papakonstantinou 2012, 2016; Debet 2018; Fellous-Sigrist 2018; Groos and Veen 2020; Hoeren and Kolany-Raiser 2018; Karjoth and Langheinrich 2019; Lazaro and Le Métayer 2015; Naftalski 2018; Puyraimond 2019; Rallet et al. 2015; Solove 2013; Wiese Schartum 2018; Zanfir 2014).

5.3 Contributions

5.3.1 Reference capability model for data protection

Essay 2.1 addresses the following research question: What data management capabilities need to be built in order to address EU-GDPR's requirements?

Our approach builds on the concept of organizational capabilities stems that stems from the resource-based view, which aims at explaining the sources of a company's sustainable competitive advantage (Barney, 1991). The capability research pushes the resource-based view rationale further and considers that companies should not only possess resources that are valuable, rare, hard to imitate, and non-substitutable (Barney, 1991, Mata et al., 1995) but also combine, develop and utilize these re-sources in a meaningful way as measured by company goals. These "capabilities" offer the advantage of being less imitable and transferable than most physical resources because they are embedded in organizational practices and individual skills (Bharadwaj et al., 1999). The conceptualization of important organizational competencies as capabilities is well-established in various research domains, including information systems research.

We retained Zhang et al.'s (2013) definition of an IT capability, which is "a firm's ability to acquire, deploy, and leverage its IT-enabled resources in combination with other resources and capabilities in order to achieve business objectives". Whereas Zhang et al. (2013) consider business objectives or business strategies as the goal of capabilities (why), we argue that capabilities for data protection are built with a regulatory compliance objective. Therefore, we define data management capabilities for regulatory compliance as a firm's ability to acquire, deploy, and leverage its data resources in combination with other resources and capabilities in order to achieve an organization's compliance objectives (Sadiq et al. 2007).

Table 10. Positioning of capabilities relative to key RCM concepts

RCM concept	Definition (based on (El Kharbili 2012))	Illustration
Regulatory guideline	Stipulates a set of obligation to comply to.	Art. 6 – "Lawfulness of processing": enumerates conditions in which data processing is legal.

RCM concept	Definition (based on (El Kharbili 2012))	Illustration
Compliance requirement (CR)	Pieces of text extracted from the regulatory guideline specifying an expected behavior / a specific condition to fulfill.	Extraction of requirements bearing data management relevance. E.g., art. 6 § 1 a and art. 7 § 1 require that data be processed according to individuals expressed consent.
Capability	Result of the interpretation of CRs in terms of capabilities that are to be implemented or improved.	Manage consent and sub-capabilities: implement consent items, collect consent instances, distribute consent, enforce consent-based processing.
Concretized compliance requirement (CCR)	Implementation of a CR in an enterprise model, fulfilling its legal specification.	A concrete measure implemented in a specific organization to operationalize CRs. E.g., "In company X, consent data should be first recorded in system 1 and pushed to other systems every 12 hours".

The capability model comprises main capability groups, i.e., system and organizational capabilities, reflecting their predominant aspect, and building on capability research. In the RBV, capabilities "involve complex patterns of coordination between people and between people and other resources" (Grant 1991). Authors relying on the RBV in the IS literature usually demarcate technological and organizational aspects that underpin IS capabilities (Baiyere and Salmela 2014; Bharadwaj 2000). Correspondingly, system capabilities are mainly enabled by data-processing systems, while organizational capabilities rely on data protection processes and responsibilities. Capabilities were derived from the EU-GDPR's underlying principles, as described by legal literature, and reflect the "pillars" of the regulation. Sub-capabilities are the result of the analysis and express compliance requirements.

Table 11. Reference capability model for data protection

System capabilities				
Define protected data scope	Identify data objects	Classify data attributes	Locate data records	
Manage consent	Implement consent items	Collect consent instances	Distribute consent	Enforce consent-based processing
Enable data processing rights	Delete data	Pseudonymize data	Transmit data in standardized form	
Organizational capabilities				
Orchestrate data protection activities	Assume data protection responsibilities	Oversee data protection activities	Control compliance of external processors	
Demonstrate compliant data processing	Maintain records of processing activities	Maintain documentation of system landscape	Supervise sensitive processing activities	
Disclose information	To individuals	To authorities		

In essay 2.1, we also provide insights from two assessment cases with insurance companies active on the swiss market and complement our findings with a study of tools advertising EU-GDPR compliance functionality against the capability model.

5.3.2 Reference personal data management life cycle model

Essay 2.2 addresses the following research questions:

- What is the impact of data protection regulations on the personal data life cycle?
- How could data life cycle models be amended in order to address regulatory requirements for data protection?

To address the first research question, we analyze two recent data protection regulation frameworks (EU-GDPR and the US-CCPA). We find that these requirements directly impact the way data objects are created, processed, and maintained. From our analysis, we propose a classification of legal requirements from data protection legislation and show how they impact the data life cycle stages. In order to reflect the changes in data management practices induced by data protection regulatory frameworks, essay 2.2 uses the data life cycle concept as a frame of reference. On a high level of abstraction, "the life cycle of something [...] is the series of developments that take place in it from its beginning until the end of its usefulness" (Collins English Dictionary 2019). The life cycle concept has been applied to various data-related domains (e.g., product data, scientific/research data) and has enjoyed a renewed interest in the context of big and open data landscapes (Möller 2013).

As an answer to the second research question, we propose a reference personal data life cycle model for data protection, which comprises a data life cycle notation for data protection, outlining how general data management activities and steps are impacted by the aforementioned regulations.

Figure 5 presents an overview of our reference personal data life cycle model, along the three major conceptual steps exhibited by existing data life cycle models: onboarding, usage and end-of-life (Möller 2013).

The notation is complemented by data model extensions to capture compliance-relevant data objects and attributes, that should be recorded, updated or deleted at various stages of the personal data life cycle. By suggesting a semi-formal notation, we translate the regulatory requirements expressed in the capability model into a set of rules. In doing so, links up with related studies from the business process management domain. It seems that such a perspective does not currently exist in our research domain.

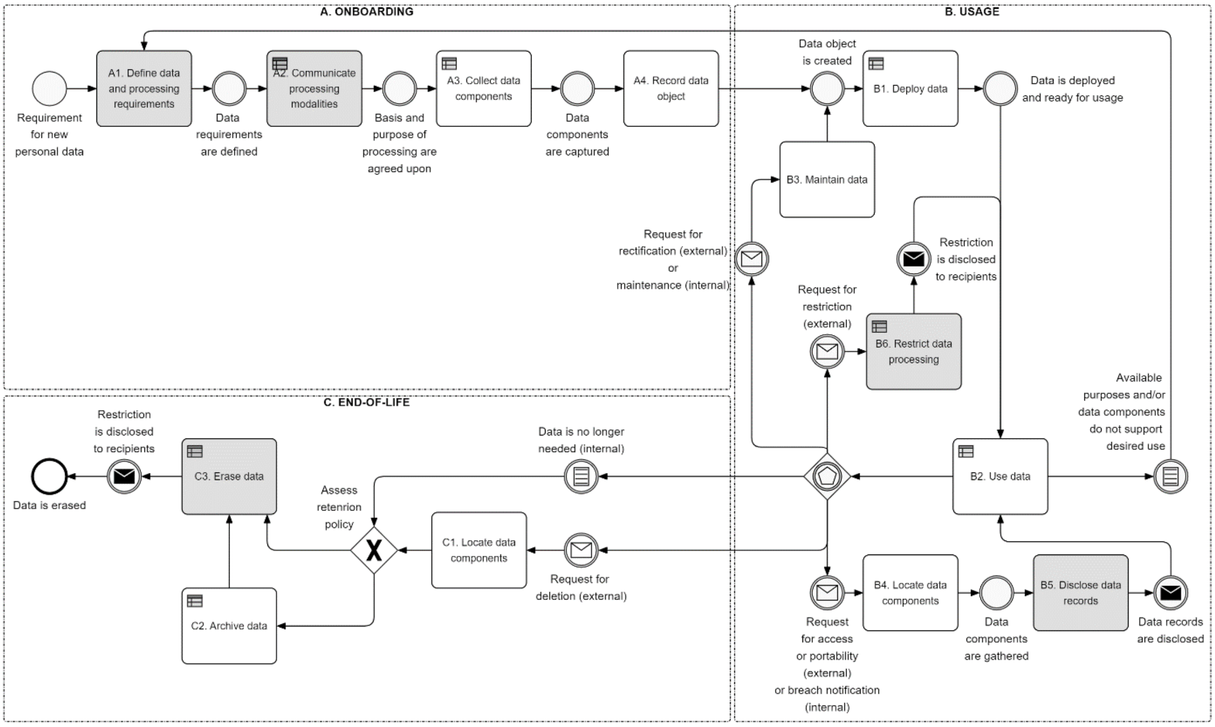

A. ONBOARDING
B. USAGE
C. END-OF-LIFE
Requirement for new personal data
A1. Define data and processing requirements
Data requirements are defined
A2. Communicate processing modalities
Basis and purpose of processing are agreed upon
A3. Collect data components
Data components are captured
A4. Record data object
Data object is created
B1. Deploy data
Data is deployed and ready for usage
B3. Maintain data
Request for rectification (external) or maintenance (internal)
Restriction is disclosed to recipients
Request for restriction (external)
B6. Restrict data processing
Available purposes and/or data components do not support desired use
B2. Use data
Request for access or portability (external) or breach notification (internal)
B4. Locate data components
Data components are gathered
B5. Disclose data records
Data records are disclosed
Restriction is disclosed to recipients
C3. Erase data
Data is erased
Data is no longer needed (internal)
Assess retenrion policy
C1. Locate data components
Request for deletion (external)
C2. Archive data

Figure 5. Reference personal data life cycle model

5.4 Implications, limitations & outlook

From academic perspective, the proposed research adopts an interdisciplinary lens, by combining the legal and information systems research domains. In doing so, it addresses existing gaps in the information privacy research in information systems. Namely, it contributes to studies on the organizational level, and provides a design-oriented approach on privacy concerns and practices (in reference to Bélanger et al., 2011). The suggested capability approach builds on new and forward-thinking regulations to provide a data management-oriented understanding of data protection. We also add to existing studies on the data life cycle (Möller 2013) by bringing in a regulatory perspective to personal data management.

When it comes to information technologies, the lawmaking process is usually slower than the development of the technologies it attempts to regulate. As a result, large organizations typically undertake significant efforts to achieve compliance *ex-post*, which usually translates into specific, large-scale projects. This occurs whenever new regulations are enforced, and the introduction of new data protection regulations is no exception. In this context, from a practitioner's perspective, the proposed research introduces a tandem of reference models meant to ease this compliance effort. The capability model can help organizations – and specifically, stakeholders within such organizations with little or no legal expertise – understand data protection regulatory requirements and provide guidance in assessing (re)design needs and in implementing these requirements. It could also be used as a communication device between practitioners from different educational backgrounds (e.g., data management and legal professionals), fostering collaboration and instilling a sense of common ground that is currently lacking. On the other hand, the amended life cycle model, along with its data model extensions and business rules provide a data-oriented view on regulatory requirements that can help data management organizations in bringing their existing practices to compliance.

In research stream 2, we have left out two neighboring perspectives on modern personal data processing in enterprises. First, our capability and life cycle models do not consider information security requirements, although their incorporation in new data protection regulations was perceived as a standout addition. As information security constitutes an entire research domain and is also handled outside of data management in enterprise, we chose to exclude it form our analysis. Second, we bound our analysis to the legal requirements on personal data processing and have not included ethical considerations or the user perspective on regulatory requirements.

Lastly, even though our data models extensions and business rules can provide system support to some extent, we have focused our efforts on translating regulatory requirements for the IS community at the conceptual level and provide little guidance on specific ways through which organizations can operationalize these requirements, e.g., in terms of tools.

Future research could adopt a solution-oriented lens to data protection implementation, similar to the approach we have taken in research stream 1 for data democratization or could use our models as reference frameworks to evaluate data protection implementation examples. It could also analyze the alignment of regulatory requirements and their implementations on the one side, with ethical data processing principles and consumer perceptions and expectations on the other side.

6 Discussion

6.1 Summary and contributions

This thesis aims at answering the question: what are capabilities and (re)design areas that organizations need to actualize to increase usage of data resources enterprise-wide, while complying with data protection regulatory requirements?

We answer it by exploring two complementary research streams focusing, on the one hand, on realizing data democratization in the enterprise through enterprise data catalogs, and, on the other hand, on translating legal requirements for data protection into compliant data management practices. In these two research streams, we adopt a design-oriented approach, and our outcomes provide a basis to operationalize key capabilities for democratizing data and complying with data protection in data-driven enterprises.

In the first research stream, we position enterprise data catalog as enabler for data democratization and key component of future enterprise systems landscapes. Our outcomes anchor this emerging type of platforms to the known research topics of data libraries and dataspaces (overall concept), metadata (data documentation) and data governance (user roles). In doing so, we enrich the ongoing scientific discourse on data democratization and openness and provide a grounded definition of the enterprise data catalog concept. In addition, our findings are informed by insights from real-world enterprise data catalog implementation projects from 12 multinational companies, ensuring their relevance and testifying to the strategic potential of these new data platforms.

In the second research stream, we analyze modern data protection regulations from a data management perspective, by selecting legal requirements and expressing them using the existing and well-known concepts of capability and data life cycle. In doing so, we enrich the privacy research domain in information systems by analyzing the oft neglected legal component of information privacy, thus answering a call for privacy research to be conducted at the organizational level, rather than the individual one (Bélanger and Crossler 2011; Smith et al. 2011). Furthermore, we address a gap in regulatory compliance literature, as our findings derive corrective solutions (i.e. providing guidance to concretize strategic compliance objectives (Cleven and Winter 2009)) from legal analysis, as opposed to preventive or detective solutions (Abdullah et al. 2009).

In both research stream, we provide an overall conceptual understanding of the topics of interest, through the reference model for enterprise data catalog and the reference capability model for data protection. These models provide a framework that enable sensemaking in the context of new and emerging topics. In both cases, we also complement these frameworks with artefacts that make the link with tangible enterprise concepts, such as data and governance models (i.e., role model, metadata models), as well as enterprise systems (i.e., data catalog functionalities, data protection business rules). In this way, our findings can also benefit practitioners in scoping data democratization and data protection initiatives and translating these concepts into their existing practices.

6.2 Implications

Through our exploration of data democratization and data protection, we find that transparency is the one aspect that establishes a clear relationship between our research streams. Whether they are aiming at increasing data usage or at making sure that they comply with data protection regulatory requirements, organizations must implement tools and measures that enable them to gain knowledge on the data they process, where they are stored, and how they are used.

Diligent data documentation is one of the drivers for data transparency, and a key requirement for data protection compliance – in developing the capability model for EU-GDPR, we found that documentation capabilities (i.e., "maintain records of processing activities" and "maintain documentation of system landscape"). As shown through the EDC RM, EDCs also rely on basic data documentation, and can also be used to document additional, usage-specific information about inventoried data.

As evidenced by the cases of Versuisse and Svizzance, linking documented purposes and systems to actual data resources, systems and processes is a significant challenge, and documentation is not enough on its own. Beyond knowing about their data resources, organizations must be able to locate them and actually govern their use. When it comes to locating the data, we have shown that data inventory functionalities are a building block of both enterprise data catalogs, enabling them to act as single frontend that link various data sources – in parallel, we have seen that all tool categories for EU-GDPR compliance offer functionalities related to data inventory.

When it comes to managing the use of data, we have shown that enterprise data catalogs support usage needs of data collectors and custodians, who make data fir-for-purpose or oversee data-related activities, but also those of data consumers, who perform said activities. The FAIR principles introduce usage licenses and authorization as concepts to control the way data is used.

Both can be implemented through data catalogs which support the documentation of processing purposes and enable authorization workflows when dealing with sensitive data. This becomes especially useful in the case of data protection compliance –when it comes to using data, enterprise data catalogs come into play in the "usage" phase of the reference personal data life cycle model, providing means to implement documented processing purposes and required authorizations. Collaboration functionalities also provide a way to ensure access control to personal data – in that sense, enterprise data catalogs can provide a controlled environment to make sure they are respected. This strengthens the case for data democratization through enterprise data catalogs, as such access and authorizations measures should ideally apply to all data users. These collaboration possibilities can also be seen as an opportunity to address the lack of common ground between business and legal stakeholders, by providing an integrated platform to streamline their exchanges.

In light of these considerations, data transparency emerges as a prerequisite to access, participation and democratization (Schlagwein et al. 2017) – consequently, initiatives towards increased data usage or data protection are both hindered by a lack thereof. Thoroughly documenting data and implementing enterprise data catalogs can help organizations in achieving what legal authors Bensoussan et al. refer to as "digital omniscience" – in their textbook on EU-GDPR, they state that "organizations must have perfect knowledge of personal data and their uses in order to comply with their obligations" (Bensoussan et al. 2018, p. 23). In the data-driven enterprise, this omniscience should apply to all data resources in order to enable data democratization.

6.3 Endnote: managing transparency for a responsible data-driven enterprise

However, we must also reckon with the limits of such omniscience, which should not be granted to every stakeholder in organizations. In that regard, the EDC RM demarcates a precise scope of users and delimits duties and prerogatives associated with their roles. In fact, making all data resources fully accessible to every single employee is neither feasible, nor desirable, and transparency should not be understood as an absolute elimination of all barriers. With regards to those barriers, a difference must be made between involuntary obstacles and intentional safeguards. Obstacles are the result of organizational and technical limitations – they are often the byproduct of the historically distributed nature of corporate IT initiatives (Hirschheim and Klein 2012), resulting in fragmented and organizational unit-centric system landscapes, eventually causing so-called "data silos". Intentional safeguards, on the other side, are barriers that need to be built to mitigate the various risks that come with data openness and transparency. While compliance with regulations certainly counts among the most critical, these risks are multifold.

First, data transparency needs to be managed to avoid replacing data silos with data chaos. The main objective of data transparency is to open up data silos and bring more visibility on data that is relevant for the enterprise. It is not, however, to collect, process and expose the largest possible quantity of data. Rising data volumes, while integral to the concept of "Big Data", do not necessarily create benefits. In fact, the term "data swamp" has appeared as a mirror image of the "data lake" concept, to describe large quantities of data remaining undocumented and unused in enterprise systems, creating additional maintenance overhead, decreasing the quality and trustworthiness of data and eventually becoming a liability rather than an enterprise asset (Beaton 2018; Brackenbury et al. 2018; Koch 2018). Second, an over-abundance of available data could lead enterprise users to a "paradox of choice", a well-described psychological phenomenon (Schwartz 2004; Schwartz and Ward 2012) which, applied to data search (Oulasvirta et al. 2009), suggests that larger quantities of available data would likely result in decreased usage. This leads to the concept of data minimization, as first pillar of the management of transparency and openness in the enterprise. It is mentioned as guiding principle in EU-GDPR and comes with economic benefits as well. Namely, it enables organizations to reduce data storage and processing costs, which also has implications on other company performance metrics, such as corporate social responsibility, in an era of increased awareness and scrutiny of the environmental impact of computing.

As mentioned with regards to the roles of data collectors and custodians, to be able to successfully encourage data use and increase the likelihood of translating such use into business value, organizations need to ensure that data is fit-for-purpose. While "fit-for-purpose" has become a common catchphrase, in practice, the emphasis tends to be placed on the fitness aspect (i.e., data quality and availability) at the detriment of the purpose. We argue that the concept of purpose of use should be positioned as the second pillar of the management of transparency and openness in the enterprise. On the one hand, it is a critical component of data protection compliance, and EU-GDPR requires that organizations thoroughly document such purposes. Beyond strict compliance, and as outlined by the reference personal data management life cycle model, determining the purpose of processing entails defining the data components meant to be collected to serve this purpose. In this sense, it may also be used to apply data minimization and data documentation from the ground up. In a broader perspective, the purpose of use is poised to act as guiding instrument towards sharpening data collection, shaping user roles and implementing access control and safeguards around transparency. It thereby constitutes an opportunity to integrate additional defining criteria, such as data ethics and fairness, which are becoming integral parts of corporate social responsibility (Harkins 2016; Parikka and Härkönnen 2020). In fact, the term "corporate digital responsibility" has recently emerged in academic research (Herden et al. 2021; Lobschat et al. 2021; Orbik and Zozuľaková 2019), with authors positioning this concept as an extension of corporate social responsibility to account for disruptions induced by the digital transformation. In this context, regulatory compliance, data protection, data minimization, openness, transparency, ethics and fairness should be seen as multiple components of a strategic mandate to become a data-driven enterprise, spanning beyond monetary aspect (George et al. 2020). In other words, data value should be understood in terms that exceed pure financial considerations.

As a response to this multiplicity of imperatives, organizations should strive to define a strong data strategy accompanied by internal data standards that consider both risk and responsibility factors[8]. We argue that the reference capability model for data protection provides a basis to abstract requirements from individual regulations so they can be incorporated within such internal (or possibly, cross-industry) standards. In addition, the EDC RM may be used to inform the purposes of use and implement them throughout the organization (i.e., in defining the roles, data documentation, scenarios and functionalities to support these purposes). Our studies show that enterprise data catalogs are uniquely positioned to provide an all-encompassing view on enterprise data, which is a pre-requisite to implementing data minimization and defining clear purposes of use. By delivering transparency on the data itself and enabling its use in a single place, EDCs offers the necessary toolset to implement intentional safeguard, enabling a purposeful management of data transparency.

6.4 Outlook

As avenue for future research, we suggest further analysis of a joint perspective on increased data usage and data protection, by investigating, on the one hand, how data ethics can be integrated into data-to-value business scenarios. As mentioned in the previous section, studies around the concept of corporal digital responsibility are starting to emerge, and researchers should analyze how these altruistic mandates come into play in the data-driven enterprise. The link could also be made with research on individual privacy concerns (Li 2012) to study how such

[8] Related initiatives from MasterCard (MasterCard 2019) and Zurich (Zurich Insurance Group 2019) are prominent examples of such corporate commitments.

efforts would impact individual choices, e.g., whether they may themselves become generators of perceived added value for customers of responsible data-driven enterprises.

On the other hand, researchers could also investigate the concept of data-driven compliance and transparency, where using platforms such as enterprise data catalogs not only support data protection compliance and transparency, but also enable organizations to develop stronger compliance capabilities through a thorough understanding of enterprise-wide data. Here, the thesis' data protection reference models may serve as example of abstracting individual legal requirements, and the data model extensions, as basis to develop new documentation schemes that are not contained in actual documents, but dynamically leverage and relate to available enterprise-wide data resources.

Finally, additional research is needed to characterize and quantify the link between the wider use of enterprise-wide data and the generation of additional business value for the firm. Here, researchers should leverage an interdisciplinary perspective and link the stream of business value of data with insights from behavioral psychology, technology acceptance, data management and actuarial science to put tangible figures on the inward- and outward-facing benefit of becoming a data-driven enterprise.

References

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).

Abdullah, N. S., Indulska, M., and Shazia, S. 2009. "A Study of Compliance Management in Information Systems Research," in *Proceedings of the 17th European Conference on Information Systems (ECIS)*, Verona, Italy, June 8, p. 5.

Addis, M. C., and Kutar, M. 2018. "The General Data Protection Regulation (GDPR), Emerging Technologies and UK Organizations: Awareness, Implementation and Readiness," in *Proceedings of the 23rd UK Academy for Information Systems Conference*, Oxford, United Kingdom, March, pp. 420–440.

Armingaud, C.-E., and Ligot, A. 2019. "Le Consentement : Le Faux-Ami Des Bases Légales ?," *Revue Lamy Droit de l'Immatériel* (160), pp. 44–46.

Arpaci, I. 2017. "Antecedents and Consequences of Cloud Computing Adoption in Education to Achieve Knowledge Management," *Computers in Human Behavior* (70), pp. 382–390. (https://doi.org/10.1016/j.chb.2017.01.024).

Aubert, J.-L. 1998. *Introduction Au Droit et Thèmes Fondamentaux Du Droit Civil*, (7^{th} ed.), Paris, France: Armand Colin.

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.

Baiyere, A., and Salmela, H. 2014. "Towards a Unified View of Information System (IS) Capability," in *Proceedings of the 18th Pacific Asia Conference on Information Systems (PACIS)*, Chengdu, China, June 24, p. 329.

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).

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)

Beaton, B. 2018. "The Moat Effects of Data Swamps," in *Proceedings of the 1st International Conference on Artificial Intelligence for Industries (AI4I)*, Laguna Hills, California, USA, September 26, pp. 85–88. (https://doi.org/10.1109/AI4I.2018.8665705).

Bélanger, F., and Crossler, R. E. 2011. "Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems," *MIS Quarterly* (35:4), pp. 1017–1042.

Beniger, J. R. 1986. *The Control Revolution*, Harvard University Press.

Bensoussan, A., Avignon, C., Bensoussan-Brulé, V., Forster, F., and Torres, C. 2018. *Règlement Européen Sur La Protection Des Données: Textes, Commentaires et Orientations Pratiques*, (2nd ed.), Brussels: Bruylant.

Bergram, K., Bezençon, V., Maingot, P., Gjerlufsen, T., and Holzer, A. 2020. "Digital Nudges for Privacy Awareness: From Consent to Informed Consent?," in *Proceedings of the 28th European Conference on Information Systems (ECIS)*, Marrakesh, Morocco, June 15, p. 64. (https://aisel.aisnet.org/ecis2020_rp/64).

Bertino, E. 2012. "Data Protection from Insider Threats," *Synthesis Lectures on Data Management* (4:4), Morgan & Claypool Publishers, pp. 1–91. (https://doi.org/10.2200/S00431ED1V01Y201207DTM028).

Bharadwaj, A. 2000. "A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Investigation," *MIS Quarterly* (24:1), pp. 169–196.

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., and Venkatraman, N. 2013. "Digital Business Strategy: Toward a Next Generation of Insights," *MIS Quarterly* (37:2), pp. 471–482.

Bohling, T., Bowman, D., LaValle, S., Mittal, V., Narayandas, D., Ramani, G., and Varadarajan, R. 2006. "CRM Implementation: Effectiveness Issues and Insights," *Journal of Service Research* (9:2), pp. 184–194.

Bose, R. 2009. "Advanced Analytics: Opportunities and Challenges," *Industrial Management & Data Systems* (109:2), Emerald Group Publishing Limited, pp. 155–172. (https://doi.org/10.1108/02635570910930073).

Bowie, N. E., and Jamal, K. 2006. "Privacy Rights on the Internet: Self-Regulation or Government Regulation?," *Business Ethics Quarterly* (16:3), Cambridge University Press, pp. 323–342.

Brackenbury, W., Liu, R., Mondal, M., Elmore, A. J., Ur, B., Chard, K., and Franklin, M. J. 2018. "Draining the Data Swamp: A Similarity-Based Approach," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA@SIGMOD)*, New York, New York, USA, June 10, pp. 1–7. (https://doi.org/10.1145/3209900.3209911).

Brownlow, J., Zaki, M., Neely, A., and Urmetzer, F. 2015. "Data-Driven Business Models: A Blueprint for Innovation," University of Cambridge, Cambridge Service Alliance, May 14. (https://doi.org/10.13140/RG.2.1.2233.2320).

Bryman, A., and Bell, E. 2007. *Business Research Methods*, (2nd ed.), Oxford, United Kingdom: Oxford University Press.

Burt, A. 2019. "Privacy and Cybersecurity Are Converging. Here's Why That Matters for People and for Companies.," *Harvard Business Review*. (https://hbr.org/2019/01/privacy-and-cybersecurity-are-converging-heres-why-that-matters-for-people-and-for-companies).

California State Senate. 2018. *California Consumer Privacy Act*.

Castets-Renard, C. 2019. "Accountability of Algorithms in the GDPR and Beyond: A European Legal Framework on Automated Decision-Making," *Fordham Intellectual Property, Media and Entertainment Law Journal* (30:1), p. 91.

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).

Chatellier, R., Delcroix, G., and Hary, E. 2019. "La Forme Des Choix - Données Personnelles, Design et Frictions Désirables," Research Report No. 06, Cahiers IP Innovation & Prospective, Research Report, Commission Nationale de l'Informatique et des Libertés.

Cheffert, J.-M. 2018. "Respect de La Vie Privée : Quand Les Approches Économiques et Juridiques Se Rejoignent," in *Law, Norms and Frefoms in Cyberspace - Droit, Normes et Libertés Dans Le Cybermonde - Liber Amicorum Yves Poullet*, Collection Du CRIDS, E. Degrave, C. de Terwangne, S. Dusollier, and R. Queck (eds.), Brussels: Larcier, pp. 505–524.

Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. “Business Intelligence and Analytics: From Big Data to Big Impact,” *MIS Quarterly* (36:4), Management Information Systems Research Center, University of Minnesota, pp. 1165–1188. (https://doi.org/10.2307/41703503).

Chen, I. J., and Popovich, K. 2003. “Understanding Customer Relationship Management (CRM): People, Process and Technology,” *Business Process Management Journal* (9:5), pp. 672–688. (https://doi.org/10.1108/14637150310496758).

Chen, Y.-J. 2010. “Knowledge Integration and Sharing for Collaborative Molding Product Design and Process Development,” *Computers in Industry* (61:7), pp. 659–675. (https://doi.org/10.1016/j.compind.2010.03.013).

Cleven, A., and Winter, R. 2009. “Regulatory Compliance in Information Systems Research - Literature Analysis and Research Agenda,” in *Enterprise, Business Process and Information Systems Modeling*, Lecture Notes in Business Information Processing, Berlin, Heidelberg: Springer-Verlag, pp. 174–186.

Cleven, A., and Wortmann, F. 2010. “Uncovering Four Strategies to Approach Master Data Management,” in *Proceedings of the 43rd Hawaii International Conference on System Sciences*, , January, pp. 1–10. (https://doi.org/10.1109/HICSS.2010.488).

Commission Nationale de l’Informatique et des Libertés. (n.d.). “Lignes Directrices et Recommandations de La CNIL.” (https://www.cnil.fr/fr/decisions/lignes-directrices-recommandations-CNIL, accessed May 26, 2021).

Creswell, J. W. 2009. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, (3rd ed.), Thousand Oaks, CA: Sage Publications.

Dallemulle, L., and Davenport, T. 2017. “What’s Your Data Strategy?,” *Harvard Business Review* (May-June), pp. 112–121.

Davis, F. D. 1989. “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” *MIS Quarterly* (13:3), pp. 319–340. (https://doi.org/10.2307/249008).

De Hert, P., and Malgieri, G. 2018. “Making the Most of New Laws: Reconciling Big Data Innovation and Personal Data Protection within and beyond the GDPR,” in *Law, Norms and Frefoms in Cyberspace - Droit, Normes et Libertés Dans Le Cybermonde - Liber Amicorum Yves Poullet*, Collection Du CRIDS, E. Degrave, C. de Terwangne, S. Dusollier, and R. Queck (eds.), Brussels: Larcier, pp. 525–554.

De Hert, P., and Papakonstantinou, V. 2012. “The Proposed Data Protection Regulation Replacing Directive 95/46/EC: A Sound System for the Protection of Individuals,” *Computer Law & Security Review* (28:2), pp. 130–142. (https://doi.org/10.1016/j.clsr.2012.01.011).

De Hert, P., and Papakonstantinou, V. 2016. “The New General Data Protection Regulation: Still a Sound System for the Protection of Individuals?,” *Computer Law & Security Review* (32:2), pp. 179–194. (https://doi.org/10.1016/j.clsr.2016.02.006).

Debet, A. 2018. “Les Nouveaux Instruments de Conformité,” in *Le RGPD*, Grand Angle, Paris: Dalloz, pp. 67–72.

Debet, A., Massot, J., and Métallinos, N. 2015. *Informatique et Libertés: La Protection Des Données à Caractère Personnel En Droit Français et Européen*, Les Intégrales, Issy-les-Moulineaux: Lextenso.

Docquir, B. 2018. *Droit Du Numérique*, Répertoire Pratique Du Droit Belge, (R. Andersen, J. du Jardin, P. A. Foriers, and L. Simont, eds.), Brussels: Larcier.

Duan, Y., Cao, G., and Edwards, J. S. 2020. "Understanding the Impact of Business Analytics on Innovation," *European Journal of Operational Research* (281:3), Featured Cluster: Business Analytics: Defining the Field and Identifying a Research Agenda, pp. 673–686. (https://doi.org/10.1016/j.ejor.2018.06.021).

Dubey, R., Gunasekaran, A., Childe, S. J., Blome, C., and Papadopoulos, T. 2019. "Big Data and Predictive Analytics and Manufacturing Performance: Integrating Institutional Theory, Resource-Based View and Big Data Culture," *British Journal of Management* (30:2), pp. 341–361. (https://doi.org/10.1111/1467-8551.12355).

Dublin Core Metadata Initiative. (n.d.). "DCMI: DCMI Metadata Terms." (https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, accessed August 25, 2019).

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).

European Commission. 2012. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*.

European Commission. 2020. "A European Strategy for Data," Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committe of the Regions No. COM(2020) 66 final, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committe of the Regions, Brussels: European Commission. (https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593073685620&uri=CELEX%3A52020DC0066).

European Data Protection Board. 2017. "Guidelines on Data Protection Officers (WP243 Rev.01)," Regulatory Guidelines, Regulatory Guidelines, European Union, May 4.

European Data Protection Board. 2018a. "Guidelines on Transparency under Regulation 2016/679 (WP260 Rev.01)," Regulatory Guidelines, Regulatory Guidelines, European Union, November 4.

European Data Protection Board. 2018b. "Guidelines On Consent Under Regulation 2016/679 (WP259, Rev.01)," Regulatory Guidelines, Regulatory Guidelines, European Union, October 4.

European Data Protection Supervisor. 2018. "Avis Préliminaire Sur Le Respect de La Vue Privée Dès La Conception," Government Report No. Avis 5/2018, Government Report, European Union.

European Data Protection Supervisor. 2019. "Annual Report 2019," Government Report, Government Report, European Union.

European Union Agency for Fundamental Rights, European Court of Human Rights, Council of Europe, and European Data Protection Supervisor. 2018. *Manuel de Droit Européen En Matière de Protection Des Données*, (2018th ed.), Luxemburg: Office des Publications de l'Union Européenne.

Even, A., Shankaranarayanan, G., and Watts, S. 2006. "Enhancing Decision Making with Process Metadata: Theoretical Framework, Research Tool, and Exploratory Examination," in *Proceedings of the 39th Hawaii International Conference on System Sciences*, Kauai, Hawaii, USA, January 4, p. 152. (https://doi.org/10.1109/HICSS.2006.152).

Fadler, M., and Legner, C. 2020a. “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).

Fadler, M., and Legner, C. 2020b. “Who Owns Data in the Enterprise? Rethinking Data Ownership in Times of Big Data and Analytics,” in *Proceedings of the 28th European Conference on Information Systems (ECIS)*, Marrakesh, Morocco, June 15, p. 207.

Federal Ministry for Economic Affairs and Energy (BMWi). 2020. “GAIA-X: The European Project Kicks off the next Phase,” Berlin: Federal Ministry for Economic Affairs and Energy (BMWi). (https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/gaia-x-the-european-project-kicks-of-the-next-phase.html).

Feldman, M. S., and Orlikowski, W. J. 2011. “Theorizing Practice and Practicing Theory,” *Organization Science* (22:5), pp. 1240–1253. (https://doi.org/10.1287/orsc.1100.0612).

Fellous-Sigrist, M. 2018. “Consent in the Digital Context: The Example of Oral History Interviews in the United Kingdom,” in *La Diffusion Numérique Des Données En SHS - Guide de Bonnes Pratiques Éthiques et Juridiques*, Presses universitaires de Provence. (https://hal-amu.archives-ouvertes.fr/hal-02058184).

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).

Fisher, C. W., Chengalur-Smith, I., and Ballou, D. P. 2003. “The Impact of Experience and Time on the Use of Data Quality Information in Decision Making,” *Information Systems Research* (14:2), INFORMS, pp. 170–188.

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., 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).

George, J. J., Yan, J. (Kevin), and Leidner, D. E. 2020. “Data Philanthropy: Corporate Responsibility with Strategic Value?,” *Information Systems Management* (37:3), pp. 186–197. (https://doi.org/10.1080/10580530.2020.1696587).

Ghasemaghaei, M., Ebrahimi, S., and Hassanein, K. 2018. “Data Analytics Competency for Improving Firm Decision Making Performance,” *The Journal of Strategic Information Systems* (27:1), pp. 101–113. (https://doi.org/10.1016/j.jsis.2017.10.001).

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).

Goldkuhl, G. 2012. "From Action Research to Practice Research," *Australasian Journal of Information Systems* (17:2), pp. 57-78. (https://doi.org/10.3127/ajis.v17i2.688).

Grant, R. M. 1991. "The Resource-Based Theory of Competitive Advantage: Implications for Strategy Formulation," *California Management Review* (33:3), p. 114.

Groos, D., and Veen, E.-B. van. 2020. "Anonymised Data and the Rule of Law," *European Data Protection Law Review* (6:4), Lexxion Publisher, pp. 498–508. (https://doi.org/10.21552/edpl/2020/4/6).

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.

Gualtieri, M., Yuhanna, N., Kisker, H., Curran, R., Purcell, B., Christakis, S., Warrier, S., and Izzi, M. 2016. "The Forrester Wave^{TM}: Big Data Hadoop Cloud Solutions, Q1 2016," Consortium Report, Consortium Report, Forrester Research.

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).

Harkins, M. W. 2016. "Corporate Social Responsibility: The Ethics of Managing Information Risk," in *Managing Risk and Information Security: Protect to Enable*, M. W. Harkins (ed.), Berkeley, CA: Apress, pp. 129–137. (https://doi.org/10.1007/978-1-4842-1455-8_9).

Herden, C. J., Alliu, E., Cakici, A., Cormier, T., Deguelle, C., Gambhir, S., Griffiths, C., Gupta, S., Kamani, S. R., Kiratli, Y.-S., Kispataki, M., Lange, G., Moles de Matos, L., Tripero Moreno, L., Betancourt Nunez, H. A., Pilla, V., Raj, B., Roe, J., Skoda, M., Song, Y., Ummadi, P. K., and Edinger-Schons, L. M. 2021. "'Corporate Digital Responsibility,'" *Sustainability Management Forum*. (https://doi.org/10.1007/s00550-020-00509-x).

Hirschheim, R., and Klein, H. K. 2012. "A Glorious and Not-So-Short History of the Information Systems Field," *Journal of the Association for Information Systems* (13:4), pp. 188–235.

Hoeren, T., and Kolany-Raiser, B. 2018. "The Importance of Data Quality for Big Data," in *Law, Norms and Frefoms in Cyberspace - Droit, Normes et Libertés Dans Le Cybermonde - Liber Amicorum Yves Poullet*, Collection Du CRIDS, E. Degrave, C. de Terwangne, S. Dusollier, and R. Queck (eds.), Brussels: Larcier, pp. 619–636.

Huth, D., Both, A., Ahmad, J., Sauer, G., Yilmaz, F., and Matthes, F. 2020. "Process and Tool Support for Integration of Privacy Aspects in Agile Software Engineering," in *Proceedings of the 26th Americas Conference on Information Systems (AMCIS)*, Salt Lake City, Utah, USA, August 15, p. 6. (https://aisel.aisnet.org/amcis2020/systems_analysis_design/systems_analysis_design/6).

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://www.journal.ecrc.nsysu.edu.tw/index.php/pajais/article/view/526).

IBM. 2017. "10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations," IBM. (ftp://ftp.www.ibm.com/software/in/pdf/10_Key_Marketing_Trends_for_2017.pdf).

IDC. 2020. "The Seagate Rethink Data Survey," Consultancy Report, Consultancy Report, IDC.

Information Commissioner's Office. 2017. "Consultation: GDPR Consent Guidance," Government Report, Government Report.

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).

Işık, Ö., Jones, M. C., and Sidorova, A. 2013. "Business Intelligence Success: The Roles of BI Capabilities and Decision Environments," *Information & Management* (50:1), pp. 13–23. (https://doi.org/10.1016/j.im.2012.12.001).

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., Kaliyaperumal, R., Kersloot, M. G., Kirkpatrick, C. R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.-A., da Silva Santos, L. O. B., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M. D., Willighagen, E. L., Wittenburg, P., Roos, M., Mons, B., and Schultes, E. 2019. "FAIR Principles: Interpretations and Implementation Considerations," *Data Intelligence* (2:1–2), MIT Press, pp. 10–29. (https://doi.org/10.1162/dint_r_00024).

Kaplan, B., and Duchon, D. 1988. "Combining Qualitative and Quantitative Methods in Information Systems Research: A Case Study," *MIS Quarterly* (12:4), pp. 571–586. (https://doi.org/10.2307/249133).

Karjoth, G., and Langheinrich, M. 2019. "Technische Gestaltung von Informed Consent," *Digma: Zeitschrift Für Datenrecht Und Informationssicherheit*, pp. 72–79.

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.

Kitchens, B., Dobolyi, D., Li, J., and Abbasi, A. 2018. "Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data," *Journal of Management Information Systems* (35:2), Routledge, pp. 540–574. (https://doi.org/10.1080/07421222.2018.1451957).

Koch, R. 2018. "Draining the Data Swamp: An Organization's Data Lake Is Only as Good as the Preparation and Maintenance Planning That Go into Creating It," *Strategic Finance* (99:12), pp. 62–64.

Kumpati, M. 1988. *Database Management System with Active Data Dictionary*.

Kurtz, C., Wittner, F., Vogel, P., Semmann, M., and Böhmann, T. 2020. "Design Goals for Consent at Scale in Digital Service Ecosystems," in *Proceedings of the 28th European Conference on Information Systems (ECIS)*, Marrakesh, Morocco, June 15, p. 69. (https://aisel.aisnet.org/ecis2020_rp/69).

Lazaro, C., and Le Métayer, D. 2015. "Control over Personal Data: True Remedy or Fairy Tale ?," *SCRIPT-Ed* (12:1), p. 32.

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, Y. 2012. "Theories in Online Information Privacy Research: A Critical Review and an Integrated Framework," *Decision Support Systems* (54:1), pp. 471–481. (https://doi.org/10.1016/j.dss.2012.06.010).

Lobschat, L., Mueller, B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., and Wirtz, J. 2021. "Corporate Digital Responsibility," *Journal of Business Research* (122), pp. 875–888. (https://doi.org/10.1016/j.jbusres.2019.10.006).

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H. 2009. "Overview and Framework for Data and Information Quality Research," *Journal of Data and Information Quality* (1:1), pp. 1-22. (https://doi.org/10.1145/1515693.1516680)

Marsh, R. 2005. "Drowning in Dirty Data? It's Time to Sink or Swim: A Four-Stage Methodology for Total Data Quality Management," *Journal of Database Marketing & Customer Strategy Management* (12:2), pp. 105–112. (https://doi.org/10.1057/palgrave.dbm.3240247).

MasterCard. 2019. "Mastercard Introduces Consumer-Centric Model for Digital Identity." (https://www.mastercard.com/news/press/2019/march/mastercard-introduces-consumer-centric-model-for-digital-identity/, accessed May 21, 2021).

Mathiassen, L. 2002. "Collaborative Practice Research," *Information Technology & People* (15:4), pp. 321–345. (https://doi.org/10.1108/09593840210453115)

Matt, C., Hess, T., and Benlian, A. 2015. "Digital Transformation Strategies," *Business & Information Systems Engineering* (57:5), pp. 339–343. (https://doi.org/10.1007/s12599-015-0401-5).

Maunula, G. 2020. "Advancing Technological State-of-the-Art for GDPR Compliance: Considering Technology Solutions for Data Protection Issues in the Sharing Economy," *Journal of the Midwest Association for Information Systems (JMWAIS)* (2020:2). (https://doi.org/10.17705/3jmwa.000060).

Meier, P. 2011. *Protection Des Données: Fondements, Principes Généraux et Droit Privé*, Précis de droit Stämpfli, Bern: Stämpfli.

Métille, S., and Raedler, D. 2017. "Swiss Data Protection Act Reform Set in Motion," *Data Protection Leader* (14:2), pp. 14–16.

Mitrou, L. 2017. "The General Data Protection Regulation: A Law for the Digital Age?," in *EU Internet Law: Regulation and Enforcement*, T.-E. Synodinou, P. Jougleux, C. Markou, and T. Prastitou (eds.), Cham: Springer International Publishing, pp. 19–57. (https://doi.org/10.1007/978-3-319-64955-9_2).

Möller, K. 2013. "Lifecycle Models of Data-Centric Systems and Domains: The Abstract Data Lifecycle Model," *Semantic Web* (4:1), pp. 67–88.

Naftalski, F. 2018. "Feuille de Route : Les Incontournables," in *Le RGPD*, Grand Angle, Paris: Dalloz, pp. 253–259.

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).

Nicolaidou, I. L., and Georgiades, C. 2017. "The GDPR: New Horizons," in *EU Internet Law: Regulation and Enforcement*, T.-E. Synodinou, P. Jougleux, C. Markou, and T. Prastitou (eds.), Cham: Springer International Publishing, pp. 3–18. (https://doi.org/10.1007/978-3-319-64955-9_1).

Orbik, Z., and Zozuľaková, V. 2019. "Corporate Social and Digital Responsibility," *Management Systems in Production Engineering* (27), pp. 79–83. (https://doi.org/10.1515/mspe-2019-0013).

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).

Oulasvirta, A., Hukkinen, J. P., and Schwartz, B. 2009. "When More Is Less: The Paradox of Choice in Search Engine Use," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, New York, New York, USA, July 19, pp. 516–523. (https://doi.org/10.1145/1571941.1572030).

Palanisamy, M., and Nandle, R. 2018. "Understanding India's Draft Data Protection Bill," , September 13. (https://iapp.org/news/a/understanding-indias-draft-data-protection-bill/, accessed January 31, 2019).

Parikka, H., and Härkönnen, T. 2020. "Corporate Social Responsibility Encompasses Data - Perspectives and Proposals for Promoting the Responsible Use of Data," Consortium Report, Consortium Report, Helsinki, Finland: Sitra.

Park, Y., Sawy, O. E., and Fiss, P. 2017. "The Role of Business Intelligence and Communication Technologies in Organizational Agility: A Configurational Approach," *Journal of the Association for Information Systems* (18:9). (https://doi.org/10.17705/1jais.00467).

Parliament of the Republic of India. 2018. *The Personal Data Protection Bill*.

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).

Pflaum, A. A., and Gölzer, P. 2018. "The IoT and Digital Transformation: Toward the Data-Driven Enterprise," *IEEE Pervasive Computing* (17:1), pp. 87–91. (https://doi.org/10.1109/MPRV.2018.011591066).

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.

Porter, M. E., and Heppelmann, J. E. 2015. "How Smart, Connected Products Are Transforming Companies," *Harvard Business Review* (October 2015). (https://hbr.org/2015/10/how-smart-connected-products-are-transforming-companies).

Poullet, Y. 2006. "EU Data Protection Policy. The Directive 95/46/EC: Ten Years After," *Computer Law & Security Review* (22:3), pp. 206–217. (https://doi.org/10.1016/j.clsr.2006.03.004).

Puyraimond, J.-F. 2019. "L'intérêt Légitime Du Responsable Du Traitement Dans Le RGPD : In Cauda Venenum ?," *Droit de La Consommation* (1:122), pp. 39–77.

Rallet, A., Rochelandet, F., and Zolynski, C. 2015. "De la Privacy by Design à la Privacy by Using," *Reseaux* (189:1), La Découverte, pp. 15–46.

van Reisen, M., Stokmans, M., Basajja, M., Ong'ayo, A. O., Kirkpatrick, C., and Mons, B. 2019. "Towards the Tipping Point for FAIR Implementation," *Data Intelligence* (2:1–2), MIT Press, pp. 264–275. (https://doi.org/10.1162/dint_a_00049).

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).

Rubio, M. 2019. "To Impose Privacy Requirements on Providers of Internet Services Similar to the Requirements Imposed on Federal Agencies under the Privacy Act of 1974," No. S.142, Congress of the United States, January 16.

Russell, K. D., O'Raghallaigh, P., O'Reilly, P., and Hayes, J. 2018. "Digital Privacy GDPR: A Proposed Digital Transformation Framework," in *Proceedings of the 24th Americas Conference on Information Systems (AMCIS)*, New Orleans, Louisiana, USA, August 16, p. 36.

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).

Sadiq, S., Governatori, G., and Namiri, K. 2007. "Modeling Control Objectives for Business Process Compliance," in *Proceedings of the 5th International Conference on Business Process Management (BPM)*, Brisbane, Australia, September 24, pp. 149–164.

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).

Sauron, J.-L. 2018. "Le RGPD: Outil Ou Entrave à La Société de l'information?," in *Le RGPD*, Grand Angle, Paris: Dalloz, pp. 75–81.

Scheer, A.-W. 1992. "Embedding Data Modelling in a General Architecture for Integrated Information Systems," in *Entity-Relationship Approach — ER '92*, Lecture Notes in Computer Science, G. Pernul and AM. Tjoa (eds.), Berlin, Heidelberg: Springer, pp. 139–161. (https://doi.org/10.1007/3-540-56023-8_10).

Scheer, A.-W., and Hars, A. 1992. "Extending Data Modeling to Cover the Whole Enterprise," *Communications of the ACM* (35:9), pp. 166-172. (https://doi.org/10.1145/130994.131007).

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).

Schlagwein, D., Conboy, K., Feller, J., Leimeister, J. M., and Morgan, L. 2017. "'Openness' with and without Information Technology: A Framework and a Brief History," *Journal of Information Technology* (32:4), SAGE Publications Ltd, pp. 297–305. (https://doi.org/10.1057/s41265-017-0049-3).

Schüritz, R., Seebacher, S., and Dorner, R. 2017. "Capturing Value from Data: Revenue Models for Data-Driven Services," in *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*, Waikoloa Village, Hawaii, USA, January 4, pp. 5348–5357. (https://doi.org/10.24251/HICSS.2017.648).

Schwartz, B. 2004. *The Paradox of Choice: Why More Is Less*, The Paradox of Choice: Why More Is Less, New York, New York, USA: HarperCollins Publishers, pp. xi, 265.

Schwartz, B., and Ward, A. 2012. "Doing Better but Feeling Worse: The Paradox of Choice," in *Positive Psychology in Practice*, John Wiley & Sons, Ltd, pp. 86–104. (https://doi.org/10.1002/9780470939338.ch6).

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).

Shankaranarayanan, G., and Even, A. 2006. “The Metadata Enigma,” *Communications of the ACM* (49:2), pp. 88–94. (https://doi.org/10.1145/1113034.1113035).

Smith, H. J., Dinev, T., and Xu, H. 2011. “Information Privacy Research: An Interdisciplinary Review,” *MIS Quarterly* (35:4), pp. 989–1016.

Solove, D. J. 2013. “Introduction: Privacy Self-Management and the Consent Dilemma,” *Harvard Law Review* (126:7), Harvard Law Review Association, pp. 1880–1903.

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.

Staiger, D. N. 2017. “Data Protection Compliance in the Cloud,” Doctoral thesis, Zürich, Switzerland: University of Zürich.

Stock, D., and Winter, R. 2011. “The Value of Business Metadata: Structuring the Benefits in a Business Intelligence Context,” in *Information Technology and Innovation Trends in Organizations: ItAIS: The Italian Association for Information Systems*, A. D’Atri, M. Ferrara, J. F. George, and P. Spagnoletti (eds.), Heidelberg: Physica-Verlag HD, pp. 133–141. (https://doi.org/10.1007/978-3-7908-2632-6_16).

Synodinou, T., Jougleux, P., Markou, C., and Prastitou-Merdi, T. (eds.). 2017. *EU Internet Law: Regulaton and Enforcement*, Cham: Springer International Publishing. (https://doi.org/10.1007/978-3-030-69583-5).

Synodinou, T., Jougleux, P., Markou, C., and Prastitou-Merdi, T. (eds.). 2021. *EU Internet Law in the Digital Single Market*, Cham: Springer International Publishing. (https://doi.org/10.1007/978-3-030-69583-5).

Synodinou, T.-E., Jougleux, P., Markou, C., and Prastitou, T. (eds.). 2020. *EU Internet Law in the Digital Era*, Cham: Springer International Publishing.

Szczepański, M. 2020. “Is Data The New Oil? Competition Issues In The Digital Economy,” Briefing, Briefing, European Parliamentary Research Service. (https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/646117/EPRS_BRI(2020)646117_EN.pdf).

Thélisson, E. 2020. “Le Règlement Général Sur La Protection Des Données et Son Impact En Suisse,” Doctoral thesis, Lausanne, Switzerland: University of Fribourg.

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).

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).

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. (https://doi.org/10.1016/0167-9236(94)00042-4).

Venkatesh, V., Brown, S., and Bala, H. 2013. "Bridging the Qualitative–Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems," *MIS Quarterly* (37:1), pp. 21–54.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), pp. 425–478.

Vilminko-Heikkinen, R., and Pekkola, S. 2017. "Master Data Management and Its Organizational Implementation: An Ethnographical Study within the Public Sector," *Journal of Enterprise Information Management* (30:3), Emerald Publishing Limited, pp. 454–475. (https://doi.org/10.1108/JEIM-07-2015-0070).

Vnuk, L., Koronios, A., and Gao, J. 2012. "Enterprise Metadata Management: Identifying Success Factors For Implementing Managed Metadata Environments," in *Proceedings of the 16th Pacific Asia Conference on Information Systems (PACIS)*, Ho Chi Minh City, Vietnam, July 11, p. 42. (https://aisel.aisnet.org/pacis2012/42).

Voigt, P., and Von Dem Bussche, A. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*, Cham: Springer International Publishing.

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.

Vom Brocke, J., and Buddendick, C. 2006. "Reusable Conceptual Models–Requirements Based on the Design Science Research Paradigm," in *Proceedings of the 1st International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, Claremont, California, USA, February 24, pp. 576–604.

Warren, S., and Brandeis, L. 1890. "The Right to Privacy," *Harvard Law Review*, pp. 193–220.

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.

Westin, A. F. 2003. "Social and Political Dimensions of Privacy," *Journal of Social Issues* (59:2), pp. 431–453. (https://doi.org/10.1111/1540-4560.00072).

Wiese Schartum, D. 2018. "Intelligible and Effective Data Protection Legislation," in *Law, Norms and Frefoms in Cyberspace - Droit, Normes et Libertés Dans Le Cybermonde - Liber Amicorum Yves Poullet*, Collection Du CRIDS, E. Degrave, C. de Terwangne, S. Dusollier, and R. Queck (eds.), Brussels: Larcier, pp. 765–782.

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), p. 160018. (https://doi.org/10.1038/sdata.2016.18).

Winter, R. 2008. "Design Science Research in Europe," *European Journal of Information Systems* (17:5), pp. 470–475. (https://doi.org/10.1057/ejis.2008.44).

Wixom, B. H., and Watson, H. J. 2001. "An Empirical Investigation of the Factors Affecting Data Warehousing Success," *MIS Quarterly* (25:1), pp. 17–41. (https://doi.org/10.2307/3250957).

Wixom, B. H., Yen, B., and Relich, M. 2013. "Maximizing Value from Business Analytics," *MIS Quarterly Executive* (12:2), p. 13.

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).

Wortmann, F., and Flüchter, K. 2015. *Internet of Things*, Springer. (http://dl.gi.de/handle/20.500.12116/10631).

Zaidi, E., De Simoni, G., Edjlali, R., and Duncan, A. D. 2017a. "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).

Zaidi, E., De Simoni, G., Edjlali, R., and Duncan, A. D. 2017b. "Data Catalogs Are the New Black in Data Management and Analytics," Consortium Report, Consortium Report, Gartner.

Zanfir, G. 2014. "Forgetting About Consent. Why The Focus Should Be On 'Suitable Safeguards' in Data Protection Law," in *Reloading Data Protection: Multidisciplinary Insights and Contemporary Challenges*, S. Gutwirth, R. Leenes, and P. De Hert (eds.), Dordrecht: Springer Netherlands, pp. 237–257. (https://doi.org/10.1007/978-94-007-7540-4_12).

Zheng, W. 2005. "A Conceptualisation of the Relationship Between Organisational Culture and Knowledge Management," *Journal of Information & Knowledge Management* (04:02), World Scientific Publishing Co., pp. 113–124. (https://doi.org/10.1142/S0219649205001110).

Zurich Insurance Group. 2019. "Zurich Announces Industry-Leading Data Commitment." (https://www.zurich.com/en/media/news-releases/2019/2019-0903-01, accessed May 21, 2021).

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, Switzerland

Submission for Business Information Systems & Engineering

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 are at the core of emerging business models and have become a cornerstone of decision-making and business processes (Dallemulle and Davenport 2017; George et al. 2014; Wixom and Ross 2017). Therefore, enterprises need to efficiently allocate data supply activities and bring them in alignment with the increasing demand for data. One of the key challenges continues to be that interrelated enterprise data is distributed over multiple databases and remains in operational silos (Hai et al. 2016; Halevy et al. 2016; Roszkiewicz 2010). Current research has started exploring data democratization as a concept of making data more broadly available for 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). However, it has not yet addressed the means, and specifically, the platforms that support data democratization.

Among these emerging platforms are Enterprise Data Catalogs (EDCs) that serve as unified data inventory and support technical as well as business professionals in finding, accessing, and using data. Industry experts emphasize that EDCs are an integral component of future enterprise IT landscapes (Belissent et al. 2019), and companies increasingly turn to data catalogs to make their data FAIR (findable, accessible, interoperable and reusable – Labadie et al. 2020). But 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 yet, and has neither been conceptualized, nor related to prior concepts and enterprise applications. From a practical perspective, companies have varying scope 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, thus complexifying the solution selection process (Goetz et al. 2020; Sallam et al. 2020; Zaidi et al. 2017). Hence, sense making 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:

What are the main constituents of an Enterprise Data Catalog as emerging platforms for data democratization?

The main contribution of our study is a multi-layer reference model (Frank 2014) that synthesizes the key constituents of an EDC and thereby lays the foundation for understanding its role as 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). To integrate and accumulate knowledge from academic and practitioner communities, we built the EDC reference model over a time period of 18 months in a close industry-research collaboration, following the guidelines of design science research (Peffers et al. 2007). The resulting EDC 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 as well as an ongoing analysis of current EDC solutions and implementations. To reflect the different perspectives on EDCs, the reference model synthesizes key constituents in three views rooted in existing IS architecture conceptualizations (Chang et al. 2007; Scheer 2001; Scheer and Schneider 2006): organization, function, and data. The organization view consists of eight data-related roles that reflect the increasing number of business users and technical experts that work with data within an organization. The function view defines nine function groups with corresponding sub-functions which support data demand and supply. The data view identifies 22 metadata objects which guide the documentation of data for technical and non-technical user roles.

With this reference model, we position the EDC as an evolutionary metadata management concept (Roszkiewicz 2010; Sen 2004) which 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 constituents organized in three architectural views. Our findings thereby inform research in the field of data management (Legner et al. 2020) and complement studies on big data and analytics infrastructures (Hyun et al. 2020; Fadler and Legner et al. 2020). Practitioners can use the EDC RM for understanding the scope and characteristics of EDCs, as well as for assessing and selecting a suitable solution and guiding the implementation.

The remainder of this article is structured as follows. In the background section, we elaborate on prior concepts which address similar, yet complementary ideas to EDCs. We then present our research design and process in detail. Afterwards, we elaborate on the considerations underlying the reference model development and its main constituents: an organization view, a function view, and a data view. To demonstrate its applicability, we used the reference model to classify 15 vendor solutions and derived two archetypes based on this assessment. We pursue with a discussion and an outlook of our research.

2 Background: Platforms for data democratization

An EDC supports enterprises in democratizing data. From prior research, we identify two concepts which facilitate data democratization and thereby pursue similar goals to an EDC, but in different contexts. First, the Digital Library (DL) that has a strong focus on making digital scholarly material, e.g. textual content or research data, accessible for the research communities (Wilcox 2018). Second, the DataSpace (DS) that describes the technical infrastructure for making interrelated data findable and accessible across distributed databases (Franklin et al. 2005). Hence, both approaches create a fundamental understanding of platforms that support data democratization and develop architecture considerations that can be translated to EDCs.

2.1 Digital Library

Libraries have always played an important role in democratizing information to a large audience (Wallace and Van Fleet 2005). Today, the Digital Library (DL) has become a central component of knowledge infrastructures (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 first concerns on the limits of printed materials preserved in physical libraries. With the advent of the web in the beginning of the 1990s and the explosion of scholarly material, the DL proliferated. Borgman (2003)'s influential DL definition 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 or 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 contents. Further components and parts of other digital library architectures are, for instance, user authentication or collaboration support (Calhoun 2014). With steady growing amount of digital content, the world wide web has also been considered as DL. This led to ambitious initiatives like the Stanford Integrated Digital Library project which aimed 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 were Sergey Brin and Larry Page who presented in 1998 their work on the Pagerank algorithm to efficiently crawl and index the web, which ultimately became the starting point for Google. While DLs had a major focus on managing textual content in the beginning, their scope has been extended to manage multi-media 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 by humans and machines (Wilkinson et al. 2016) and thereby help in democratizing 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 four 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*, different services to find relevant data are provided by a DSSP. Here, either data or metadata can be queried and 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 tighten either by the user or semiautomatically. 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 DS systems. For instance, Google proposes a catalog (named GOODS) which manages metadata of datasets distributed over heterogenous systems and provides services to users to find relevant datasets faster (Halevy et al. 2016). Hellerstein et al. (2017) argue that the changing requirements for data management with regards 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 clearly support data democratization in companies, they focus on technical architectures and services, but neither explore their integration into enterprise IT landscapes nor elaborate 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 as well as conceptualization is missing.

Drawing on our literature review on DS and DL concepts, we isolate three essential components that can be translated to EDCs. 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's (2003, p. 42). In a similar way, EDCs are supporting "needs and uses" of different enterprise roles, comprising 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 are functions. Both concepts, DL as well as DS comprise on the one side functions for storing, indexing, and cataloging data and on the other side user functions for creating, searching, browsing, discovering, and using data (Borgman 2003; Franklin et al. 2005).

Table 12. Platforms for data democratization

	Digital Library (DL)	**DataSpace (DS)**	**Enterprise Data Catalog (EDC)**
	(Borgman 2003; Calhoun 2014; Fox and Sornil 2003)	(Franklin et al. 2005; Halevy et al. 2016; Hellerstein et al. 2017)	(Russom 2017; Zaidi et al. 2017)
Purpose	Providing access to large numbers of academic information sources	Find interrelated data across distributed databases	Facilitate data democratization in enterprises
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 stand as evolution of data dictionaries, business glossaries and metadata repositories
Users	Education/ research	Not clearly defined: Organizations on various levels (e.g., enterprises, government agencies, libraries, "smart" homes)	Large scope of employees of the enterprise
Examples	Stanford Integrated Digital Library (Stanford 1999) Fedora (Staples et al. 2003)	Google Dataset Search (GOODS) (Halevy et al. 2016) International Data Space (IDS) (Otto et al. 2019)	Enterprise Data Catalog solutions

In the enterprise context, these topics have been partly addressed 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. On the opposite end, business glossaries document key terms in a way that is understandable for business users. 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 enterprise landscapes (Kumpati 1988). Yet, these concepts are not integrated and only address restricted user and functional scopes compared to DL and DS. Extensions in both areas are essential to data democratization and are addressed by EDCs, as emerging platforms for data democratization.

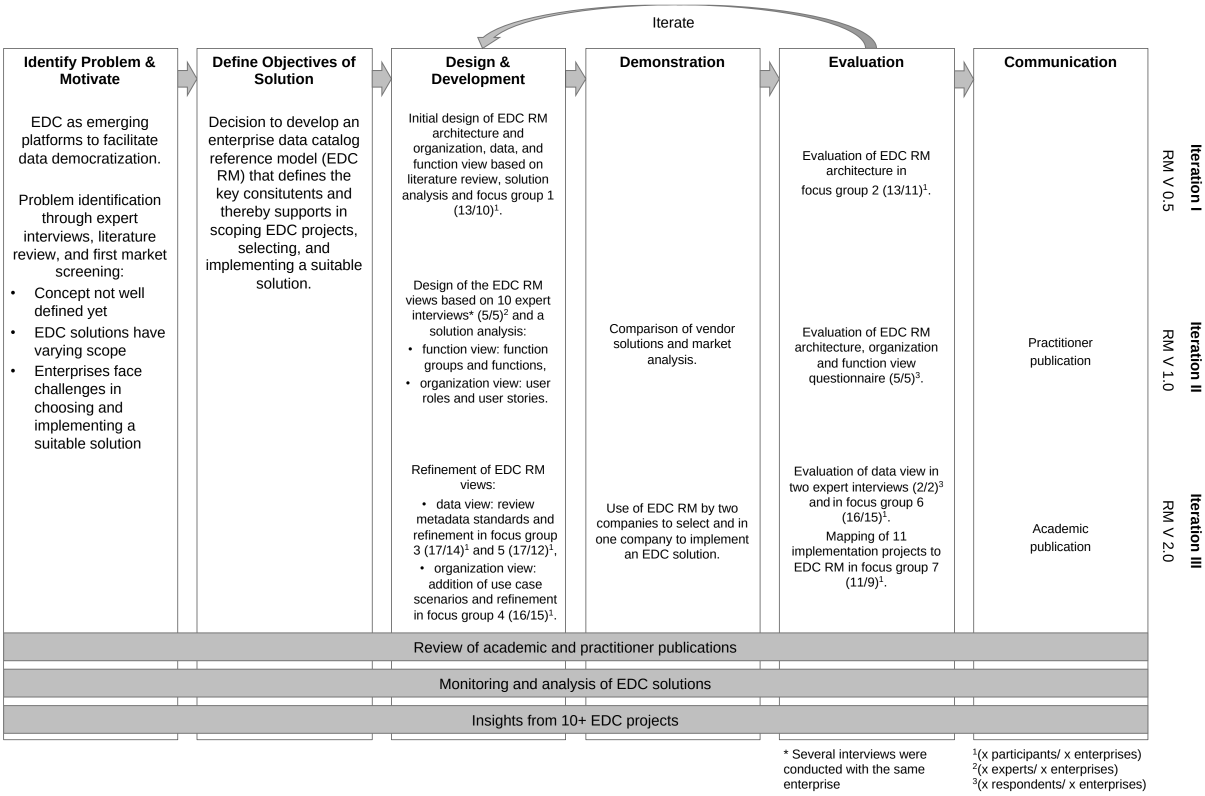
3 Research design

3.1 Research objectives

The goal of our research is to provide an understanding of the EDC concept as emerging platform for data democratization by developing a reference model. Reference models are important artifacts that help in accumulating design knowledge from academic and practitioner communities and have become very popular to provide guidance in 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.

3.2 Reference model development based on design science research

Reference models are usually developed in iterations of design and evaluation following design science principles (Winter and Schelp 2006). To develop the EDC reference model, we followed the design science research method outlined by Peffers et al. (2007). Being confronted with emerging solutions (EDC) that address a contemporary problem (data democratization) yet not well defined in research and practice, we chose the *Objective-Centered Solution* initiation. As a reference model is an artifact for solving practical problems, frequent and early iterations with practitioners are essential (Sonnenberg and vom Brocke 2012) to reach an effective solution (Hevner et al. 2004). We developed the EDC reference model following the steps outlined by Peffers et al. (2007) in three major iterations over a timespan of 18 months (see Figure 6), each comprising a design and evaluation step. As the model converged to a stable state with version 1.0, we also included demonstration steps.



Identify Problem & Motivate	Define Objectives of Solution	Design & Development	Demonstration	Evaluation	Communication	
EDC as emerging platforms to facilitate data democratization. Problem identification through expert interviews, literature review, and first market screening: • Concept not well defined yet • EDC solutions have varying scope • Enterprises face challenges in choosing and implementing a suitable solution	Decision to develop an enterprise data catalog reference model (EDC RM) that defines the key consitutents and thereby supports in scoping EDC projects, selecting, and implementing a suitable solution.	Initial design of EDC RM architecture and organization, data, and function view based on literature review, solution analysis and focus group 1 (13/10)[1].		Evaluation of EDC RM architecture in focus group 2 (13/11)[1].		Iteration I RM V 0.5
		Design of the EDC RM views based on 10 expert interviews* (5/5)[2] and a solution analysis: • function view: function groups and functions, • organization view: user roles and user stories.	Comparison of vendor solutions and market analysis.	Evaluation of EDC RM architecture, organization and function view questionnaire (5/5)[3].	Practitioner publication	Iteration II RM V 1.0
		Refinement of EDC RM views: • data view: review metadata standards and refinement in focus group 3 (17/14)[1] and 5 (17/12)[1], • organization view: addition of use case scenarios and refinement in focus group 4 (16/15)[1].	Use of EDC RM by two companies to select and in one company to implement an EDC solution.	Evaluation of data view in two expert interviews (2/2)[3] and in focus group 6 (16/15)[1]. Mapping of 11 implementation projects to EDC RM in focus group 7 (11/9)[1].	Academic publication	Iteration III RM V 2.0

Review of academic and practitioner publications

Monitoring and analysis of EDC solutions

Insights from 10+ EDC projects

* Several interviews were conducted with the same enterprise

[1] (x participants/ x enterprises)
[2] (x experts/ x enterprises)
[3] (x respondents/ x enterprises)

Figure 6. Research process

Throughout the entire 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 13). The experts that joined the group were overseeing EDC implementation initiatives or were 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, whereas others aimed at lowering 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 13. Enterprise data catalog projects of participating companies

Company	Industry	Revenue range	Purpose	Status
A	Adhesives	1 – 50 B €	Metadata management	Rollout and onboarding
B	Pharmaceuticals	1 – 50 B €	Support of data analytics	Implementation in progress
D	Chemistry	50 – 100 B €	Support of data governance and data analytics	Rollout and onboarding
C	Sportswear	1 – 50 B €	Support of data analytics	Rollout and onboarding
E	Manufacturing	1 – 50 B €	Metadata management	Rollout and onboarding
F	Pharmaceuticals	1 – 50 B €	Support of data governance and data analytics	Rollout and onboarding
G	Manufacturing	50 – 100 B €	Metadata management (register, search & retrieve data)	Pilot
H	Automation	1 – 50 B €	Support of data governance	Tool selection
I	Retail	100+ B €	Support of data governance	Continuous usage and maintenance
J	Tobacco	50 – 100 B €	Support of data governance	Continuous usage and maintenance
K	Information Technology	1 – 50 B €	Support of data governance and data analytics, metadata management	Continuous usage and maintenance
L	Fashion and jewelry	1 – 50 B €	Data glossary	Rollout and onboarding
M	Packaging	1 – 50 B €	Support of data governance, analytics, inventory and automation	Scoping and tool selection

3.3 Iterations

Based on our review of prior literature, as well as existing EDC solutions, we have designed the EDC reference model iteratively, in close industry-research collaboration. As generic and abstract design knowledge, the EDC reference model explicates (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).

Iteration I – Reference model V 0.5 (January 2018 to 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 a first analysis of selected EDC solutions, we gained insights into the functional scope of EDCs, and a focus group 1 helped us identifying typical users. We translated these insights into a multi-layer 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 comprising 13 data management experts from 11 enterprises. The participants assessed the general structure and confirmed its usability for their own EDC projects. Major points of improvements 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 exemplary user stories were yet not representative enough for the companies' own requirements.

Iteration II – Reference model V 1.0 (July 2018 to September 2018): In the design and development step of the second iteration, we enhanced the EDC RM (Version 1.0), foremost the function view, based upon the feedback from the previous iteration. In order to do so, we conducted a series of expert interviews (one or more interviews per exert; ten interviews in total), as well as a detailed analysis of EDC solutions on the market and gained insights on EDC requirements for 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 extended by online searches for further tools and by insights from interviews with practitioners. From about one-hundred identified solutions, we filtered out 15 that are in line with companies' priorities and understanding of an EDC (see Table 15). 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 which we further composed through distinct functions. As demonstration step, we mapped the 15 selected solutions to the right 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 the means of individual interviews with five EDC project managers and through a semi-structured questionnaire based on evaluation criteria proposed by Prat et al. (2015). Respondents were asked to rate the relevance of user roles and their exemplary user stories, and function groups and their individual functions. We captured the answers by using a Likert-scale with five answer options (*strongly disagree*, *disagree*, *uncertain*, *agree*, *strongly agree*). For the user roles and user stories, all respondents answered between agree and strongly agree on average concerning the relevancy for their company. For most of the function groups and functions, the respondents answered between *agree* and *strongly agree* on average concerning the relevancy for their company. Only for the function groups data assessment and data analytics, the respondents responded with *uncertain* and *agree* on average. In addition, we asked the respondents to rate if the reference model is complete, easy to understand, and useful for their company. Overall, the respondents commonly agreed that the reference model is easy to understand and useful. However, a few were uncertain whether the reference model was yet complete, and we included their feedback in the next design iteration.

Iteration III – Reference model V 2.0 (September 2018 to November 2019): Based on the expert feedbacks gathered in the second iteration, a minor change was made in the function view, where we separated one function group in 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 by means of a template with guiding instructions and afterwards 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. Furthermore, 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 stage, the focus group reached a consensus that the organizational and function views had converged to a stable state.

In parallel, we resumed 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 on internally and during focus group 3 and 5, after which we reached a stable version. This version was further refined through expert interviews with representatives from two external organizations, who had experience in developing similar models in the context of EDC implementation projects. We integrated the expert's 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 to the reference model – as a result of focus group 7, we found that the EDC reference model is extensive enough to categorize and support EDC implementation projects. This was confirmed in demonstration steps, where two organizations (company I from out participant sample (s. Table 13), as well as an external organization active in the energy industry) used the reference model (particularly the organization and function views) during request for proposal (RfP) meetings with EDC vendors to compare offerings and select an EDC solution. Furthermore, company B (Table 13) relied on the EDC reference model to guide their 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 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 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 view, function view, data view) and their relation to each other (see Figure 7). In the second level, we deconstruct each view into its key constituents.

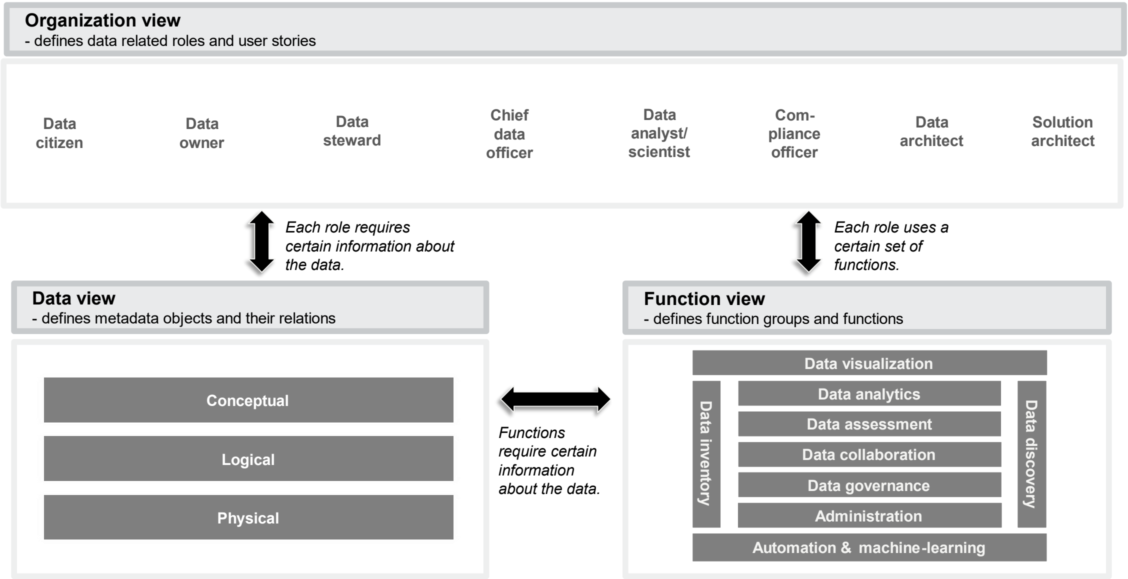


Figure 7. Enterprise data catalog reference model architecture

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 is composed of employees, with varying levels of data-related expertise and expectations. Based on literature, expert input, and insights from implementation projects, we have identified eight user roles for EDCs together with exemplary user stories (see Table 14).

Table 14. 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 in 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 to 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

User roles	User stories	Related function groups and functions	Related metadata objects
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 to data)*	Data domain Business object attribute Data object attribute Value domain Role Actor Board / Council
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 & 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,

User roles	User stories	Related function groups and functions	Related metadata objects
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
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 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 into the system thanks to data collectors and make sure 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 in the enterprise.

Each user role has specific requirements towards data and views data differently; for example, a marketing manager, in his role as data citizen, wants to understand how a certain forecast was calculated, whereas a data analyst requires domain knowledge about the specific data she is working with. We use user stories to describe how their specific tasks could be better achieved using an EDC. Each user story is associated with relevant function groups and metadata objects which we will present in detail in succeeding 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 data assets forward 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 that relate 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 understand data usage patterns and flows between systems in order to properly use it subsequently 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 was especially highlighted by 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; to retrieve and use data; and to 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 with appropriate function groups for visualization, automation & ML, and data collaboration. In the following, we describe each function group individually.

Table 15. Function view: Function groups and functions

Function group	Function
Data inventory	Data registration Metadata management Business glossary Data dictionary Data provenance

Function group	Function
	Data ingestion / crawling Upload / link content
Data discovery	Search Dataset recommendation Data access Data subscription Data delivery
Data analytics	Data story Data application repository Data query Data lake monitoring
Data assessment	Data usage Data quality Data risk Data valuation Data profiling Data lineage
Data collaboration	Tagging Rating Following / Updates Commenting Messaging / User chat Sharing
Data governance	Role & responsibility management Workflow Rule & policy Data access management
Data visualization	Graphs Diagrams In-table visualization Dashboards / Cockpit Data flow / Network visualization
Automation & machine learning	Automated scanning / ingestion Automated classification / tagging Normalization / data similarity Data unification Usage pattern analysis Recommendation
Administration	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. Hereby, an EDC uses a pre-defined metadata model (see section 4.2) which describes data for technical as well as non-technical user roles and allows 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 as well as 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, Cambridge Semantics provide with their Anzo® Smart Data Lake an EDC solution which uses a semantic graph model to document data and their relations from a physical to a conceptual level.

With the **Data discovery & delivery** function group, data can be found and obtained in a guided way by users. DLs as well as 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 changed over time (Halevy et al. 2016). In a similar way, an EDC's most basic functionality is a search function that matches a user request with the related data resources, based on metadata. 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 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 Collibra's solution. Here, users can discover data either by searching or receiving a recommendation. Once a relevant dataset is found, a user requests access to data. This process behaves similar to the checkout in an eCommerce shop: A user adds her/his data of interest to a shopping cart triggering a workflow after checkout in which a corresponding data responsible 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 / scientists. Being connected to code repository solutions such as GitHub, these roles can maintain their analytics application repository in link 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 enhances not only 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 was 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 instance. Data stories also allow to onboard employees faster because they describe analytics applications in a more comprehensive way 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 not only data as input to analytics application, but also 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. Alation's solution allows to write queries and to delve sample datasets within the tool itself. This gives data scientists, for instance, 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 to evaluate and measure data according to specific metrics, e.g., business value or 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. Thereby, "*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). While data valuation is a rather new field of research, a rich body of knowledge exist for 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 which 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 are able to collaborate when maintaining, documenting, or using data. Besides commenting, users can also collaborate by rating or tagging datasets. These are typical functionalities that are used for curating content in modern platform environments like Facebook, for instance, but are considered in DLs as well. 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 the 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, required work on data is identified early in the process and directly assigned to a responsible. Data quality issues can be solved more quickly, for instance.

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 comprises, for instance, 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 research DL with the notion of data curations which is 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, a curator role maintains the DL's content over its lifecycle. While the content in DL is rather homogenous and is usually publicly available, in the enterprise context data is rather heterogenous, involves more complex rules, and is oftentimes also confidential. Hence, EDC solutions on the market provide different functionalities to facilitate data governance, e.g., nominating roles, assigning responsibilities, and establishing workflows for data throughout the enterprise. 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 remitted 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, or metadata about data using dashboards or cockpits and data flow or network visualizations. This function group facilitates the other functions groups and helps user roles in decision making. In Collibra, the lineage of data can be visualized to gain transparency how data flows between systems. All data in Informatica's EDC can be visualized through tableau as a 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 either be done using a rule-based or a learning-based approach. In Zaloni's solution, for instance, complete workflows can be automated using rules. In Alation's EDC solution, a learning-based approach is used to recommend users which tables to join when s/he starts 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 was 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 as well as their relationships. Our proposed model addresses the following requirements that 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, e.g., data citizens, data analysts, data protection officers, data architects, data stewards, data owners). To reconcile both business- and system-oriented perspectives on data, metadata objects follow data modelling guidelines and are organized in 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, addressing governance and analytics considerations, in addition to classical business concepts.

Table 16. 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

Modeling layer	Model view	Metadata object
		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):

- The business process view describes where and how data is used in an organization, through the documentation of business domains, capabilities, and processes. *Business processes* represent how an enterprise performs its activities, and are enabled by *business capabilities*, which consist in a combination of technological, informational, and organizational resources, and representing what a company does (Bharadwaj 2000; Grant 1991). The business domain represents strategic business areas of an enterprise and reflect 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 is understandable for business users.
- the analytics view refers to metrics, key performance indicators, and reports. Metrics quantifiable measure reflecting the state of the enterprise. They are the basis key performance indicators (KPIs). Finally, reports organize and present metrics and / or KPIs in human-readable form, enabling visualization by 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) – it focuses on the detail level of entities and their relationships, and documents core *data domains* as well as related *business object* and their *business object attributes* are documented, along with the *applications* that create and *transform* them. It contains, for instance, single entity definitions (e.g., a “customer” could be mapped to multiple physical instantiations, and have various conceptual meanings depending on specific business contexts).

The physical layer reflects the implementation view on 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 Discussion

The suggested reference model conceptualizes EDCs as emerging platforms for data democratization through defining the key constituents along three different architecture views. It thereby allows not only to define the EDCs' scope, but also to discuss and compare it to prior concepts and to assess different implementation variants of EDCs which we observed in implementation projects and the market analysis.

5.1 EDCs as evolution of metadata concepts

Compared to previous concepts, we see the EDC as an evolutionary concept of metadata management (Sen 2004) because they aggregate existing metadata concepts (i.e. data dictionary, business glossary, and metadata repository) to provide a holistic viewpoint on data and connect technical-focused and business-oriented user roles. By using the reference model to compare these concepts (s. Table 17), we clearly observe that EDCs go beyond these existing concepts and facilitate data democratization for a broad audience within organizations. From a functional perspective, data dictionaries, business glossaries, and metadata repositories serve a purpose of data inventory, as they are meant to provide documentation for all data or business objects. Business glossaries and metadata repositories can also support governance efforts, as they provide additional information (e.g., definition, metrics) on the data. Metadata repositories can also support data discovery functions, 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. This highlights the key differentiator of data catalogs, which are meant to encompass preceding metadata management solutions, and extend them from a functional perspective, by enriching data documentation capabilities with data usage capabilities, thus catering to the needs of a broader variety of users.

Table 17. EDCs compared to 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 & delivery			▒	■	■	■
Data analytics				▒	▒	■
Data assessment				▒	■	▒
Data collaboration				▒	■	▒
Data governance		▒	▒	▒	■	▒
Data visualization				■	■	■
Automation & ML				▒	▒	■
Information layer						
Conceptual		■		■	■	■
Logical			■	■	■	■
Physical	■		▒	■	■	■

Our analysis of EDC solutions, as well as feedback from experts, also show that EDCs provide core functionalities, which 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 data to put it to use. In addition, EDCs also offer functionalities that enable direct use of data resources (i.e., data analytics, data assessment, automation & ML), as well as direct communication 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.

5.2 Analytics-oriented vs. governance-oriented EDCs

While most of the EDC solutions on the market offer basic functionalities to inventory, govern, and discover data, none of these cover all function groups. In fact, the inventory function is the common denominator. Most of the analyzed EDCs are complete stand-alone solutions, while certain solutions (for instance Ab Initio, Informatica, Talend, and SAP)[9] require the combination of several components and tools from the respective product portfolios.

The analysis and comparison of EDC solutions and implementation projects provides interesting insights as it allows identifying 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 in order 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 are, for example, Anzo Smart Data Lake 4.0 (Cambridge Semantics), Enterprise Data Catalog (Informatica), Smart Data Catalog (Waterline), or Zaloni Data Management Platform.
- **Governance-oriented EDC:** Other EDC solutions focus on Data Collaboration and Data Governance. These tools primarily aim at supporting 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 are, for example, Adaptive Metadata Manager, Collibra Data Governance Center, Information Value Management (Datum), IBM InfoSphere Information Governance Catalog, Axon Data Governance (Informatica) or SAP Information Steward.

6 Conclusion and outlook

This study proposes a multi-level reference model that analyzes novel data catalog platforms through a triptych of architecture views - the organization view, the function view, and the data view - and sets these in relation to each other. Our results showcase the wide reach of data catalogs – their ability to act as frontend for enterprise-wide data, to serve the purposes of a variety of technical and business users, and to facilitate collaborations make them a solid foundation for data democratization in organizations.

Being multi-leveled, the proposed reference model allows to gain an overview of an EDC's constituents but also understand each constituent in greater detail. We derive the reference

[9] 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".

model architecture as common basis from architecture considerations that have been developed for platforms supporting data democratization, specifically Digital Libraries (DL) and DataSpaces (DS). The second and succeeding levels give detailed information on the three architecture views. The organization view outlines user requirements of eight EDC user roles in form of user stories and links each role to required function groups and metadata objects. This perspective shows that EDCs act as integrated platforms connecting different user roles (e.g., data scientist and data owner) while coordinating data management activities (e.g., managing data access) enterprise wide in an efficient way. The function view defines nine function groups to support data supply and demand. This part extends the general functions derived from the DS concept (Franklin et al. 2005) that 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 blueprint for enterprises seeking to design their own, company-specific metadata model in support of providing data documentation for data democratization platforms. It goes beyond existing metadata standards that contain flat lists of attributes, by proposing enterprise-specific metadata objects, featuring views dedicated to usage and governance contexts, and grouped in conceptual, logical and physical layers.

As generic and abstract design knowledge, the EDC reference model explicates (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 recommended practice, it is intended to form the basis for assessing vendor solutions and creating company-specific situational designs (instantiation). The EDC reference model anchors these emerging platforms for data democratization in enterprises to the known research topics of the Digital Library and the DataSpace (overall concept), metadata management (data view with metadata objects), and data governance (organization view with user roles). This conceptualization enriches the ongoing scientific discourse on data democratization and provides a grounded definition of the enterprise data catalog concept. It also contributes to research on data management (Legner et al. 2020) as well as the emerging big data and analytics platforms (Fadler and Legner 2020; Hyun et al. 2020). Since our findings are informed by real-world insights from organizations that implement EDCs, they also inform practitioners and can serve as a starting point to select an EDC solution but also give guidance for the implementation.

As with any empirical work, this study has its limitations. As EDCs are a novel concept, most of the enterprises were still in the early phase of their EDC implementation journey. Moreover, we observe that 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.

6.1 Acknowledgements

This research was supported by the Competence Center Corporate Data Quality (CC CDQ). The authors would like to thank the experts that contributed to this research.

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).

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$^{\text{TM}}$: 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. 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.

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).

Democratizing Data in Enterprises: Towards Enhanced Data Usage and Collaboration with Data Catalogs

Clément Labadie, Markus Eurich, and Christine Legner

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

Presented as Communication of the 20th Symposium of the Association Information et Management (AIM), 2020 – Best Paper Award

Extended journal manuscript

Abstract. Data democratization denotes a paradigm according to which all employees have access to data and are empowered to use it. It can be viewed as a cornerstone of digitalization, and enterprises striving to truly benefit from the value of their data resources need to reshape system landscapes and define new data-driven use cases. However, they typically face two challenges: First, they lack a precise understanding of the user communities in the enterprise that would engage in such use cases, as well as of their specific needs of accessing, using, and managing data. Second, enterprise data is distributed across silos, that reflect system and organizational boundaries, and are hardly documented. Prior research has mostly dealt with these two aspects separately, through data governance (to establish clear roles and responsibilities) and metadata management (as an attempt to overcome data silos), usually considering narrow user groups. Although the topic of data democratization has recently been picked up by academics and practitioners, current contributions focus on conceptual definitions or platform design. However, a user perspective is missing, and there is a need to understand relevant user roles and their requirements to ensure the success of data democratization initiatives. Against this background, this study aims to clarify, on the one hand, which data-related roles emerge in the context of data democratization. On the other hand, we identify enterprise data catalog as suitable platform candidate to address data democratization challenges and specify their role in supporting enterprise users to work with data. Based on a design science research approach that involved experts from twelve multi-national organizations, we identify eight emerging data-related roles and capture their requirements of accessing, using, and managing data. These findings contribute to the academic debate on data democratization and the FAIR principles (i.e., making data findable, accessible, interoperable and reusable) in the enterprise context. Practitioners can use the insights to manage organizational change within the context of the digital transformation.

Keywords: data management, data catalog, data democratization, data governance

Table of contents

List of figures

List of tables

1 Introduction

In the data economy, enterprises aim at monetizing their data resources (Grover et al. 2018; Legner et al. 2020; Wixom et al. 2013; Wixom and Ross 2017), thus integrating data into their value creation process (Bharadwaj et al. 2013). One concrete way of doing so is by introducing data-driven business models (Brownlow et al. 2015; George et al. 2014; Schüritz et al. 2017), which are "*designed to create additional business value by extracting, refining and ultimately capitalizing on data*" (Brownlow et al. 2015). Recently, enterprises have come to the realization that the value potential of data resources can only be maximized when data usage is not the purview of a limited group of data experts, and that a wider audience of enterprise users need to be empowered to integrate data into their activities. However, data are managed in silos that reflect system and organizational boundaries, and are not yet available to the increasing number of employees that aspire to access and use it.

To overcome these issues, enterprises strive for democratizing data and perform a paradigm shift according to which enterprises "*open[…] organizational data to as many employees as possible […]*" (Awasthi and George 2020). Data-driven companies like AirBnB (Feng 2017; Williams 2017) and Google (Halevy et al. 2016) emphasize that the meaningful documentation of and widespread access to data is an important enabler of data democratization. In the same wave, enterprise data catalogs are considered an important instrument of making enterprise data findable, accessible, interoperable and reusable (so-called FAIR principles). Prior research has mostly dealt with these aspects through data governance (to establish clear roles and responsibilities) and metadata management (as an attempt to overcome data silos), usually considering narrow user groups. Although the topic of data democratization has recently been picked up by academics and practitioners, current contributions focus on conceptual definitions or platform design.

Two interesting gaps can be identified: first, in order to democratize data, we need more insights about the data-related roles within an enterprise and their specific needs of finding, accessing, using, and managing data. Second, the role of data catalogs as enabler of data democratization has only started to be discussed in academic literature (Labadie et al. 2020). While data catalogs stand in a tradition of data dictionaries, metadata repositories, and business glossaries, they are mostly discussed in the practitioner community with focus on platform design (Goetz et al. 2018; McKendrick 2018; Zaidi et al. 2017).

To address these gaps, we investigate the following research questions: *Which data-related roles emerge in the context of data democratization? How do data catalogs support data-related roles and their typical data needs and usages?*

Our study is based on a design science research approach (following the methodology suggested by Peffers et al., 2007) and involved experts from 12 multi-national organizations. By analyzing their existing initiatives together with literature on digital libraries (DL) and data governance, we developed a set of typical roles along with their requirements for finding, accessing, using data and managing data. To understand the significance of enterprise data catalogs in data democratization, we outline the way they empower user to work with data, by analyzing the purposes for which each user role can benefit from enterprise data catalogs, the functionalities that support these purposes and the data-related collaborations that arise from the use of an integrated data platform. Our research sheds light on existing as well as emerging data-related roles, their requirements (formulated as user stories), as well as data-related collaborative use case scenarios (summarized as vignettes). It contributes to the academic debate on data democratization and the FAIR principles in enterprise contexts, by highlighting the processual aspect of data democratization and the need to consider user requirements beyond data governance mandates. Practitioners can use the insights to identify emerging data-related roles in their organizations and implement data catalogs in support of their data-driven and digital transformation.

The remainder of the paper is structured as follows: in the next section, we introduce the concept of data democratization in detail and introduce the concept of data catalogs as well as the FAIR principles. Afterwards, we explain the design science research approach used in this study. Subsequently, we synthesize typical user needs for finding, accessing, using, and managing data with data catalogs. Then, these results are consolidated through the lens of data-related collaborations and presented as vignettes. The paper concludes with a discussion, summary and an outlook on future research.

2 Background and related work

2.1 Data democratization

Data democratization is a novel concept that has been coined by the business community (Díaz et al. 2018; Marr 2017; Yuhanna et al. 2019), but still lacks a solid definition. An academic definition of data democratization has recently been proposed by (Awasthi and George 2020) –

however, it relies mostly on practitioner insights and is behaviouristic, in the sense that it focuses on the inputs and outputs of data democratization, without elaborating on the democratization process itself. In the following, we make an attempt at providing a theoretical foundation for data democratization drawing upon social and political sciences and the conceptualization of democratization as a transition activity.

Democratization is an extensively researched social and political sciences. It is often defined as a process of transitioning from an authoritarian political regime to a representative one (Kauffman, 2018). In the related literature, a democratization process entails the following components (Grugel and Bishop 2013):

- *representation* through clean elections,
- introduction of *individual rights*,
- *inclusion*, i.e., a decrease or elimination of structural obstacles to participation in the exercise of power (Dryzek 1996).

The last component, inclusion, is especially interesting when applied to data and information. In the political context, democratization is the transition from a state at which power, as a resource, is only exercised by a selected group of people, to a state where it becomes accessible to a broader group of people. When applied to informational resources, this process has been described as the *democratization of knowledge*, which had a significant impulse following the invention of the printing press, which contributed to remove obstacles to the dissemination of information (Wallace and Van Fleet 2005).

In the same spirit, enterprises engage in data democratization activities, in the sense of empowering employees extending beyond traditional data experts to work with data, by removing obstacles to find, access and make use of data resources (as seen, for instance, in "digital native" organizations such as AirBnB (Feng 2017; Williams 2017)). This entails, on the one hand, breaking up exclusionary data silos and (Schlagwein et al. 2017), on the other hand, providing employees with a concrete way to access data (Awasthi and George 2020; Hyun et al. 2020).

Based on these considerations, in order to integrate the concepts from social and political sciences and reinforce the processual aspect of data democratization, we propose the following definition of data democratization: as *the process of empowering a group of users spanning beyond established data experts to find, access and use data, by removing obstacles to data exploration and sharing in enterprises*.

2.2 Data catalogs as digital libraries

The definitions of democratization proposed by the literature express conceptualizations of the processes and related guiding principles towards removing obstacles to a resource. However, they do not specify how this democratization process should be realized and which specific tools could support it.

In the context of consuming information, it has been recognized that the introduction of the printing press and, subsequently, the development of physical libraries, have been driving forces of the democratization of knowledge (Wallace and Van Fleet 2005). In research, the concept of digital libraries has been developed, in anticipation of the limitations of keeping physical, printed material in libraries (Licklider 1965), and has seen numerous contributions following the development of personal computing and the internet (Calhoun 2014). According to Borgman's (2003, p.42) definition of digital libraries, they consist of two main conceptual aspects:

- Storage and retrieval: this aspect positions digital libraries as "*an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (...)*" (Borgman, 2003, p. 42). It encompasses data resources, their documentation and the technical capabilities required to create, search and use them.
- User and community needs: this aspects expresses the idea that "*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). It puts the emphasis on the interaction between individuals, groups and data, and thus focuses on the use of data.

Enterprise data catalogs can be seen as an enterprise-driven equivalent to digital libraries, in a way that they are meant to organize and present enterprise data, for a specific community of users. From a technical perspective, enterprise data catalogs can be associated with metadata, which is generally defined as "*data about data*" (Sen, 2004) and "*aim at facilitating access, management and sharing of large sets of structured and/or unstructured data*" (Kerhervé & Gerbé, 1997). Sarting with the *description of field names* (1960s) and *table definitions* (1970s), system-specific *data dictionaries* were created in the 1980s, *metadata repositories* (1990s), *business data dictionaries* (2000s) and *business glossaries* (2010s) have been established to support business users to work with data. These concepts are usually loosely integrated and do not constitute, in and of themselves, an enterprise counterpart of digital libraires – enterprise data catalogs integrate these existing approaches, thus providing a platform for a diversity of enterprise users to find, access and use data.

2.3 Enterprise Data catalogs as implementation of the FAIR principles

Enterprise data catalog initiatives can be viewed as an implementation of the FAIR principles for the enterprise context (Díaz et al. 2018). In academia, the FAIR principles address the obstacles commonly encountered by scientists when collecting data for their research activities (Wilkinson et al. 2016). They "*describe distinct considerations for contemporary data publishing environments with respect to supporting both manual and automated deposition, exploration, sharing, and reuse*". More specifically, the FAIR principles address obstacles related to:

- finding data: this obstacle relates to data exploration and search. According to the *findable* principle, users must be able to discover data through embedded browsing and searching capabilities.
- accessing data: this obstacle relates to data availability. According to the *accessible* principle, after having identified suitable data, users must be able to retrieve them (e.g., by exposing them through APIs and standard interfaces).
- using data: this obstacle relates to the user's ability to use the data, and has two main implications. First, according to the *interoperable* principle, data must be distributed using standardized formats, so that users don't run into incompatibility issues preventing them to use the data. Second, according to the *reusable* principle, data must be properly documented and feature a description of their contents, enabling users to understand them, as a prerequisite to adequately using them.

The FAIR principles are meant to be applied to the design and implementation of platforms supporting data exploration, sharing and re-use. In doing so, such platforms would address those obstacles and reduce the amount of time and effort that researchers need to invest in gathering data, thus enabling them to focus on their own contributions. Thus far, the FAIR principles have played an important role in the academic world. Since their introduction, researchers have suggested implementation considerations to guide the design of solutions (Jacobsen et al. 2019), and related scientific initiatives are emerging, such as the Internet of FAIR Data and Services (van Reisen et al. 2019). Despite their popularity in research, the FAIR principles have experienced few thorough applications in other settings, including corporate settings (van Reisen et al. 2019). In this study, we consider enterprise data catalogs as a sophisticated emerging solution to realize the FAIR principles in the enterprise context. From a business perspective, "*a data catalog maintains an inventory of data assets through the discovery, description and organization of datasets. The catalog provides context to enable data analysts, data scientists, data stewards and other data consumers to find and understand a relevant dataset for the purpose* *of extracting business value*" (Zaidi et al. 2017). Data catalogs have also been characterized as data management platforms that "*take metadata management from its backwater silos to a centralized cross-platform facility that is feature-rich and comprehensive*" (Russom, 2017, p. 3). Therefore, they aim at exposing enterprise-wide data resources to a wide group of users, and provide them with functionalities to leverage them, thus supporting data democratization.

2.4 Research gap

Data democratization leads to all employees working with data, and data catalogs can serve as facilitator for making data FAIR. However, being a fairly recent concept, data catalogs have not yet been conceptualized from a theoretical perspective. However, this concept is based on two prerequisites: from a technical point of view, data must be available in a meaningful way, i.e., they must be findable, accessible, interoperable, and reusable (FAIR). From an organizational perspective, data catalogs must focus on *user and community needs* as outlined by Borgman (2003). While there is increasing interest in data catalogs, it mostly stems from the practitioner community and focuses on platform design. To make enterprise data catalogs successful as platforms that provide access and context on enterprise-wide data resources, reconcile data supply and demand, and enable data democratization, there is a need to consider the point of view of intended users, which is recognized by research as a critical criterion for ensuring user acceptance (Jarke et al. 2011). This is not yet covered in literature, and there are two gaps to be addressed, consistent with Borgman's (2003) findings:

- Understanding who are potential **users** of enterprise data catalogs, user involvement being commonly considered as a decisive factor for IS success (Bano and Zowghi 2015),
- Understanding how these users would benefit from enterprise data catalogs with regards to data democratization objectives, i.e. user **needs**, as inability to meet user requirements has been identified as the main contributive factor to IS failures in prior research (Dwivedi et al. 2015).

The very objective of a democratization process is to empower a wider group of people (here: enterprise users) to gain usage of a given resource (here: enterprise-wide data). Therefore, it is paramount to understand their requirements for using this resource, in order to successfully orchestrate this process. In this view, this study sets out to outline the relevant enterprise roles for data democratization and clarify the contribution of enterprise data catalog as platforms that support data-related user requirements of finding, accessing and using enterprise data.

3 Research context and approach

This research context is a research program on data management (Österle and Otto 2010). It convenes data management expert from around 15 multinational organizations, active in diverse industrial areas, and a team of researchers, in order to develop artefacts around specific topic areas. Consortium research have proven to be fruitful when tackling wicked topics such as ours, as they provide a favorable environment for the accumulation of knowledge. Specifically, the interactions between academic researchers and participant organizations allows for synthesizing both scientific and operational knowledge, thus ensuring the academic rigor, practical relevance and generalizability of the outcomes (Legner et al. 2020).

Research activities were performed by a team of 4 researchers in close collaboration with 17 senior data management experts from around 15 multi-national organizations, over a period of 12 months. These experts were overseeing implementation initiatives or were closely involved with key implementation aspects. Thus, they had a distinct knowledge of data catalog implementation and could provide relevant insights regarding the intended audience and related usage of enterprise data catalogs. As our study adopts a user-oriented IS design approach, this environment enables frequent and in-depth interactions with representatives of target enterprise data catalog users, through focus groups and individual interviews, which are requirement elicitation techniques suitable for our approach (Nuseibeh and Easterbrook 2000).

The roles and usage needs were developed following an iterative Design Science Research (DSR) process (depicted in Figure 8), as suggested by Peffers et al. (2007), based on literature on data governance and concepts related to enterprise data catalogs, namely digital libraries and dataspaces, as well as practitioner input. Frequent interactions between the experts supported the DSR design cycles, as experts could provide feedback on the artefact along various steps of the data catalog implementation process in their organizations.

Table 18. Overview of participating organizations

Company	Industry	Revenue range	Expert (title)	Catalog purpose	Status
A	Adhesives	1 – 50 B €	Lead Data Architect	Metadata management	Rollout and onboarding
B	Automation	1 – 50 B €	Head of Corporate Master Data Management	Support for data governance	Tool selection
C	Chemistry	50 – 100 B €	Data Catalog Product Owner and Enterprise Architect	Support for data governance and analytics	Rollout and onboarding

Company	Industry	Revenue range	Expert (title)	Catalog purpose	Status
D	Fashion and jewelry	1 – 50 B €	Data glossary manager	Data glossary	Rollout and onboarding
E	Information technology	1 – 50 B €	Solution Advisor Expert	Support for data governance, metadata management and data analytics	Continuous usage and maintenance
F	Manufacturing	1 – 50 B €	Corporate Data Management	Metadata management	Rollout and onboarding
G	Manufacturing	50 – 100 B €	Enterprise Architect for IoT and Digitalization	Metadata management	Pilot
H	Packaging	1 – 50 B €	Global Master Data Driver	Support of data governance, analytics, inventory and automation	Scoping and tool selection
I	Pharmaceuticals	1 – 50 B €	Associate Director Information Management	Support for data analytics	Implementation in progress
J	Pharmaceuticals	1 – 50 B €	Business Data Analyst, Global Data Team, Ops IT	Support for data governance and data analytics	Rollout and onboarding
K	Retail	100+ B €	Team Lead Master Data Management	Support for data governance	Continuous usage and maintenance
L	Tobacco	50 – 100 B €	Manager Enterprise Data Governance System	Support for data governance	Continuous usage and maintenance
M	Sportswear	1 – 50 B €	Director of Tech Consultancy and Lead Solution Architect	Support for data analytics	Rollout and onboarding

Initial discussions around the concepts of the FAIR principles, data democratization and enterprise data catalogs brought the problem to the attention of the research team and helped to *identify the problem and motivate* the research effort. The group discussed enterprise data catalogs, which are emerging platforms meant to provide a single access point for enterprise data and offering embedded functionalities to work with it. The motivation can be summarized as follows: first, enterprise data catalogs were perceived as important enabler for making enterprise-wide data available to a broader audience of employees, including non-data experts. Second, the participant companies that joined were considering implementing enterprise data catalog solutions but were struggling to clearly motivate associated benefits and specify a plan forward.

As a result of these discussions, in the objective definition phase, the group set out to further examine the role and capabilities of enterprise data catalogs in supporting data democratization in the enterprise context to develop a conceptual understanding of these emerging platforms according to three design areas: data documentation, implementation and usage aspects. In this study, we focus on the latter.

Participants also expressed purposes for which employees would use and benefit from enterprise data catalogs. The decision was thus made to further substantiate typical enterprise roles and outline data-related usages supported by enterprise data catalogs, enabling employees spanning beyond the traditional scope of data experts to increase their usage of data.

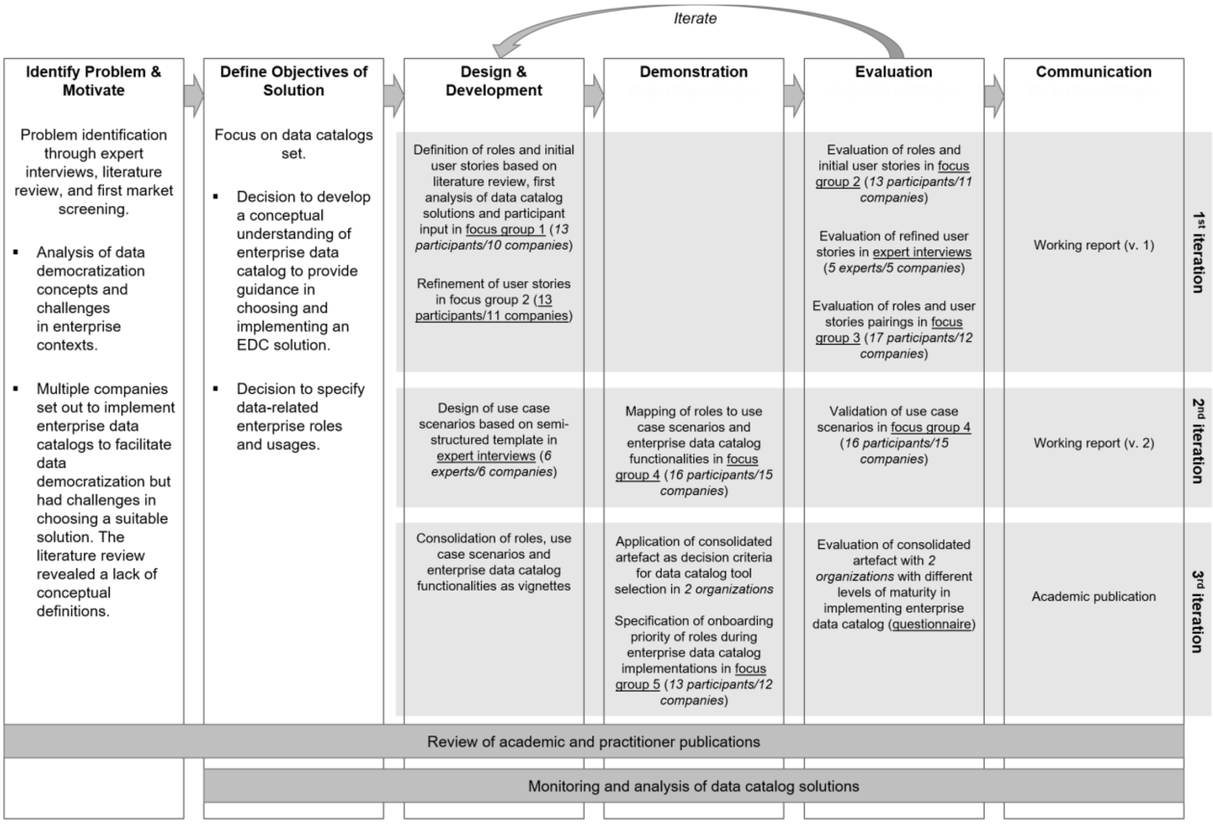

Iterate
Identify Problem & Motivate
Define Objectives of Solution
Design & Development
Demonstration
Evaluation
Communication
Problem identification through expert interviews, literature review, and first market screening.
▪ Analysis of data democratization concepts and challenges in enterprise contexts.
▪ Multiple companies set out to implement enterprise data catalogs to facilitate data democratization but had challenges in choosing a suitable solution. The literature review revealed a lack of conceptual definitions.
Focus on data catalogs set.
▪ Decision to develop a conceptual understanding of enterprise data catalog to provide guidance in choosing and implementing an EDC solution.
▪ Decision to specify data-related enterprise roles and usages.
Definition of roles and initial user stories based on literature review, first analysis of data catalog solutions and participant input in focus group 1 (13 participants/10 companies)
Refinement of user stories in focus group 2 (13 participants/11 companies)
Evaluation of roles and initial user stories in focus group 2 (13 participants/11 companies)
Evaluation of refined user stories in expert interviews (5 experts/5 companies)
Evaluation of roles and user stories pairings in focus group 3 (17 participants/12 companies)
Working report (v. 1)
1st iteration
Design of use case scenarios based on semi-structured template in expert interviews (6 experts/6 companies)
Mapping of roles to use case scenarios and enterprise data catalog functionalities in focus group 4 (16 participants/15 companies)
Validation of use case scenarios in focus group 4 (16 participants/15 companies)
Working report (v. 2)
2nd iteration
Consolidation of roles, use case scenarios and enterprise data catalog functionalities as vignettes
Application of consolidated artefact as decision criteria for data catalog tool selection in 2 organizations
Specification of onboarding priority of roles during enterprise data catalog implementations in focus group 5 (13 participants/12 companies)
Evaluation of consolidated artefact with 2 organizations with different levels of maturity in implementing enterprise data catalog (questionnaire)
Academic publication
3rd iteration
Review of academic and practitioner publications
Monitoring and analysis of data catalog solutions

Figure 8. Research approach based on Design Science Research (Peffers et al. 2007)

The subsequent phases consisted of iterative design cycles, featuring back-and-forth interactions between the research team and the practitioners, either in group or on an individual basis. Following a user-oriented approach, we applied common requirements engineering techniques to capture user requirements for EDCs, namely user stories followed by use case scenarios, as they have been shown to build on one another (Gilson et al. 2020; Wautelet et al. 2016).

In the first iteration, set out to build an initial understanding of data-related roles and usage needs in the context of enterprise data catalogs. We described these usage needs in terms of user stories, which are a common requirements engineering method for capturing user requirements based on roles, goals and benefits – they have also proved useful to decompose user requirements for software into "comprehensible chunks" (Lucassen et al. 2016, p. 211). These first outcomes were based on three inputs: the literature review on related concepts (e.g., DL), a first analysis of selected EDC solutions, and focus group 1 – these activities helped us identifying typical EDC users. We translated these insights into eight user roles and user stories. They were then evaluated by 13 data management experts from 11 enterprises who assessed them and confirmed their usability for their own EDC projects (focus group 2) – while the eight user roles received general agreement, the exemplary user stories were yet not representative enough for the companies' own requirements. Conversely, the user stories were refined based on participant feedback, and evaluated again in five interviews with EDC project managers. We captured the answers by using a likert scale with five answer options (*strongly disagree*, *disagree*, *uncertain*, *agree*, *strongly agree*) and all respondents answered between agree and strongly agree on average concerning the relevancy for their company. We conducted a broader additional evaluation of the pairings of roles and user stories during focus group 3 – participants were asked to rate each pairing in terms of importance (using a liker scale with five answer options) and implementation status (likert scale with three answer options). Most pairings received an average importance rating of "high" to "very high", with only 3 (out of 26) receiving an average importance rating of "moderate". This confirmed the relevance of the outlined roles and user stories.

The second iteration entailed the extension of user stories, by establishing links between the roles and enterprise data catalog functions and outline collaborations between several roles. For this purpose, we used use case scenarios to describe these activities, as they are meant to depict "*a list of actions or event steps, typically defining the interaction between roles and a system in order to achieve a goal*" (Wautelet et al. 2016, p. 128). Hence, we first conducted expert interviews to derive use case scenarios, using a semi-structured template comprising the following dimensions: *trigger* (goal)*, users* (among the 8 *roles* defined and validated in iteration 1), *data,* and *functions*. The research team also derived additional use case scenarios following the same template, and they were validated during focus group 4 – participants nominated the most relevant ones, which were subsequently mapped to typical functions exhibited by available enterprise data catalog solutions.

In the third iteration, as a final development step, the research team clustered and consolidated roles, user stories and usage scenarios as illustrative vignettes. Vignettes are simulation of real-life events meant to capture how individuals would behave in a given, concrete, yet hypothetical situation (Gould 1996). In that regard, they convey narratives that can either be used as input to collect responses from individuals (Hughes and Huby 2004), or to present the results of data analysis following interactions between researchers and participants, ensuring that they have reached a shared understanding of the situation (Urquhart 2001; Urquhart et al. 2003). In the information systems discipline, vignettes have been used in the latter way, to synthesize the results of qualitive and interpretive fieldwork (Marabelli and Galliers 2017; Ofner et al. 2013). For our purposes, we used vignettes to consolidate individual and group feedback gathered throughout interviews and focus group meetings, and present envisioned interactions between user roles, beyond individual user stories, ensuring that they were in line with participants' views and experiences. Subsequently, as demonstration step, company K (s. Table 18) as well as an external organization (energy company, 1 – 50 B € revenue range) applied the roles and vignettes to their tool selection process. They screened tool vendor offerings according to their ability to support the user stories and use case scenarios, determining the selection of candidates for requests for proposals, as well as contributing to the final purchasing decision. This demonstrates the utility of the roles, user stories and use case scenarios. Finally, the consolidated roles, user stories and use case scenarios were evaluated by two companies from the initial sample, one of them having been among the most advanced of the group in terms of implementation, and the other having progressed from initial concept to implementation over the course of the research effort. Through a questionnaire, both companies were asked to assess the understandability, consistency, completeness, validity, usefulness and adaptability on a likert scale with five answer options, ranging from "fully disagree" to "fully agree". All of the evaluated criteria received ratings between 4 and 5, except for the adaptability aspect – as the most mature company was already implementing their data catalog when the artefact was finalized, their approach was only partly informed by it. Finally, in focus group 5, we asked participant companies to classify the roles according to the order in which they would onboard enterprise users, in order to gain additional insights on implementation considerations. During these discussions, participants once again confirmed the alignment of the eight roles with their implementation experiences.

4 Findings

4.1 Roles

In the digital libraries research stream, Borgman emphasizes that catalogs should be built "*by [and for] a community of users*" (Borgman 2003, p. 42). Conversely, in the case of data catalogs, we aim at defining which roles constitute such a user community in the enterprise context, considering both data specialists and non-technical user groups. Furthermore, as one of our goals was to understand how these roles would interact with each other with regards to data usage, we also clustered them in three groups, reflecting data catalogs' objectives of reconciling data supply, curation and demand (s. section 2.2). Drawing upon roles established in data governance literature (DAMA International 2017; Khatri and Brown 2010; Otto 2011a; Weber et al. 2009), we clustered roles along the purposes of data catalogs to support data supply, curation and demand (Borgman 2003; Lord et al. 2004). We identified 8 roles that can be categorized in 3 groups: data collectors, data custodians and data consumers (Lee and Strong 2004).

Data collectors designate employees responsible for onboarding data resources into an enterprise's data infrastructure (*data supply*). As IT-oriented roles (DAMA International, 2017, p. 569), they maintain the platforms and structures (e.g., data models) needed to store and distribute data. Their main objective is to prepare data resources. Data providers consist of the following roles:

- *Solution architects* implement business requirements as software solutions or platforms. They need to understand how the data are organized and used in order to build appropriate solutions, thus making data available wihin the enterprise.
- *Data architects* define how data is stored in enterprise systems, by providing data models and metadata (Korhonen et al. 2013; Otto 2011a). They need to understand how data is organized and used, and align their models accordingly.

Data custodians are usually considered as business-oriented roles (DAMA International 2017; Lee and Strong 2004; Otto 2011b). They maintain the data itself, making sure they match quality standards and are well documented. Their main objective is to make data resources fit-for-use. Data custodians consist of the following roles:

- *Data owners* are responsible for data pertaining to a specific domain (e.g., customer, supplier, material), and "define the related requirements of intended use" (Khatri and Brown 2010). They continuously maintain the data and their definitions, and control access to them over their entire lifecycle. The data owner role is well established and enjoys renewed interest and relevance in the context of increased data exploitation (Fadler and Legner 2020).
- *Data stewards* are subject matter experts responsible for metadata (DAMA International, 2017, p. 569), who perform data assessment tasks (e.g., in terms of quality, maturity, consistency), define appropriate improvement measures. They also support business users with data-related matters (Otto 2011a), which makes them a key role for data democratization.

While data collectors and data custodians work on the data itself, ensuring that it is available and usable, data consumers use data to support business purposes. So far, these roles have not been covered in prior literature, which mostly elaborates on control and governance aspects. As the objective of data democratization initiatives is to facilitate access to data and bring down the barriers to data consumption (Hyun et al. 2020), they are now coming into focus. Creating data transparency and providing access will also enable data consumers to not only use data when it is necessary to their activities, but also leverage it as a way to generate additional value (Grover et al. 2018). Data consumers consist of the following roles:

- *Data citizens* represents employees who rely on data for their day-to-day work (Korhonen et al. 2013), but are not data specialists. They have been referred to as data consumers or data beneficiaries (Khatri and Brown 2010), and their tasks often include data entry and use of reporting functionalities to generate task-specific insights.
- *Data analysts* are responsible for identifying new use cases for data usage and analytics. They make an extensive use of available data resources to design proof-of-concepts for providing added business value.
- *Compliance officers* are responsible for implementing regulations in an enterprise. As such, they need to identify, flag, review and control the use of all data resources that fall within the scope of regulations (e.g., personal data in the case of the General Data Protection Regulation (GDPR), or pharmaceutical product data in the case of the Identification of Medicinal Products norms (IDMP)).
- *Chief data officers* are responsible for orchestrating an enterprise's data practices, with an emphasis on business value. This, they require an overview of their enterprise's complete data resources to make strategic decisions and track progress of ongoing data-related initiatives.

Out of these 8 roles, data citizens and compliance officers can be considered non data experts. Data analysts and chief data officers, while data-proficient by definition, are new types of roles that was not a part of the traditional data-related roles. In terms of implementation, participant feedback indicated that data collectors and data custodians would be onboarded to enterprise data catalogs in priority[10], for two main reasons. First, from a user acceptance perspective, participants identified them as roles that would be easily convinced of the usefulness of an enterprise data catalogs, and that would require the least amount of training. Second, they are also necessary roles to build the underlying architecture for an enterprise data catalog and

[10] Data steward, data architect and data owner were rated first, second and third respectively.

populate it with data resources. The role of data citizen was nominated in fourth place, further testifying to the data democratization objectives of enterprise data catalogs. The chief data officer was nominated in seventh place – while participants highlighted its strategic importance in order to get support from upper management and securing resources for implementing an enterprise data catalog, the CDO themselves would not count among the first users of such a platform. The compliance officer was rated last, and participants indicated that the enterprise data catalog would first need to reach a high level of maturity and cover most of an organization's data resources in order to reliably support compliance activities.

4.2 User stories

In this section, we explore in further detail the data-related needs of enterprise users outlined in the previous section and specify how existing and emerging data-related activities can be facilitated by enterprise data catalogs. In the digital libraries research stream, Borgman emphasizes that "*digital libraries […] and their functional capabilities support the information needs and uses of the [user] community*" (Borgman 2003, p. 42). Thus, we present these usage needs on a per-role basis, through exemplification, and link them with functional capabilities of data catalogs (s. Table 19).

Table 19. Overview of roles, usage and functionality needs for data catalogs

Roles	User stories	Functionality
Data Citizen	Understand how to correctly enter data in 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 to data)*
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*

Roles	User stories	Functionality
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 to data)*
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*
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*
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 & ML: *automated classification / tagging* Data Inventory: *metadata management* Data Discovery: *search*
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*
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*

4.2.1 Data citizen

Data citizen represent non-data expert, business users in organization who do not only consume data, but often also enter data. Data entry is among the most common source of errors and data quality issues, as data referencing specific, existing master data objects must be entered in a specific way in order to be recognized in automated processes. In order to avoid time-consuming remediation actions, it is important that data citizen understand *how to correctly enter data in a system*. For instance, when registering a product in inventory, a data citizen may need to refer to a specific product category (due to, for instance, licensing constraints). Although they may have an intuitive understanding of the various existing product categories, they can use the data catalog to *search* for naming *rules and policies*, as well as browse a *business glossary* to clarify the meaning of terms. The same functions help *understand how to interpret data in a report* and can be complemented by *documented data stories* made available through the data catalog.

Data discovery and visualization functions also support data citizen in *finding data for a specific task*, e.g., through *drilling down* from the business context (for instance, a specific business process), through *search* functionalities, and social media-like *subscription* functionalities providing information on new and/or updated datasets in a specific domain. Data discovery is further augmented by recommendation as well as by data collaboration functionalities, which not only expose data resources to data citizens, but also enable them to identify trusted data sources, i.e., data sources that have been successfully used by others. Finally, data collaboration functionalities such as *rating* and *commenting* also enable the signaling of potential issues with data, by enabling data citizen to *provide feedback on datasets*. They can also *contact* the responsible stewards or owners directly, through user communication within the data catalog.

4.2.2 Data owner

Data owners are responsible for key data elements or data domains (Otto 2011a) and play a key role in populating the data catalog with datasets, as well as in providing contextual information (e.g., metadata on data quality) on these datasets. Data catalogs can be interfaced with most data processing systems through connectors, APIs or semi-automated importation of data. Once data sources are linked with the data catalog, data owners can use *data inventory* functionalities to *register* the datasets that fall within their domain of responsibility. From thereon, the data owner will primarily maintain definitions regarding technical (*data dictionary*) or business (*business dictionary*) data-related terms. As data catalogs acts as a frontend for data resources stored in various systems, they can compare default and real-life values for data elements in those systems.

If data owners notice undesirable variations, they may take corrective measures from the data catalog:

- By *maintaining value domains*, they can restrict the values that can be assigned to specific data elements to pre-defined ones.
- Using *workflow* functionalities, they can *set up approval processes* so that they can check and approve data changes before they are saved.

Workflows are made possible by data catalogs' data access functionalities, thanks to which data owners can also grant access to data in their ownership to selected stakeholders, and *share guidelines and definitions* with them, through *collaboration* functionalities.

4.2.3 Data steward

Data stewards need to make sure that enterprise data corresponds to agreed levels of data quality (Wang and Strong 1996). Thus, one of their main tasks is to assess the data quality continuously and implement improvement measures. They also maintain data definitions over time.

Once data has been onboarded by data owners, data stewards are in charge of continuously *assessing data in their area of responsibility* sure they meet requirements over time and maintain a corresponding data documentation. For this purpose, they will use data assessment functionalities, e.g., to identify which data sources are used the most (*data usage*), those suffering from quality issues (*data quality*), in combination with other dimensions (data profiling) in order to prioritize needs and improvements.

4.2.4 Chief data officer

The chief data officer is a new data-related role, with a strategic perspective on the management of data in the organization. As business increasingly relies on data, they need an overview of all data assets enterprise-wide, in order to assess their value, to track progress made and make decisions.

As data resources are increasingly linked to the data catalog and documented by data owners and stewards, chief data officers can *gain an overview on data assets* through the platform. From thereon, a chief data officer can use data governance functionalities to document roles in the data catalog and attribute these roles and responsibilities to the user accounts that will access the data catalog platforms. Based on these role definitions, workflows can be created, assigning each role with rights on the platform and triggering, for instance, authorization and review processes when needed.

Finally, chief data officers can leverage *data assessment* functionalities within the data catalogs. These functionalities aggregate the metadata that is provided by data owners and data stewards on data quality, valuation and risk, as well as data usage statistics provided by the data catalog platform. In this situation, the data catalog acts as a *benchmarking* dashboard enabling chief data officers to *assess enterprise-wide data assets along desired criteria*.

4.2.5 Data analyst

Data analysts are data specialist who work with data to generate business insights.

In order to *understand a problem domain*, the data analyst can benefit from the integration of the data catalog with enterprise-wide data resources to search for and discover datasets related to a specific topic (i.e., problem domain).

From there on, it is possible to *obtain relevant data for a given problem (starting from business meaning or technical field)*. Thanks to the integration between the catalog platform and data storage, exploration is possible on the data-field level.

In order to benefit from previous work or enable other users to reuse the results of the analytics work, these activities can be documented within the data catalog to *provide or retrieve documentation on analytics work with data* and *publish datasets, possibly with a data story of a successfully implemented analytics application.*

At any point during data exploration and work, the analyst can leverage collaboration functionalities of data catalogs to *provide feedback on datasets (e.g., usability, quality)* directly to the responsible data owners and stewards.

4.2.6 Compliance officer

In the *discover compliance-sensitive data and locate systems / attributes* scenario, the compliance officer relies on structures provided by the data owner and documentation (e.g., flags for personal data, references to storage systems) maintained by data owners, stewards, as well as data and solution architects. By accessing this information as well as the data sources within the data catalog, the officer could establish a map of personal data storage in the enterprise.

When it comes to *understanding compliance issues in a specific dataset*, the officer could use the data catalog to explore databases down to the field level directly from the data catalog. As the subject-matter expert for data protection, she could enrich the documentation of the dataset with additional information regarding compliance, e.g., linking it to impact assessment or internal guidelines regarding the use of personal data. If data exploration reveals data resources containing personal data that were not flagged, the officer can also *label data (attributes) that need(s) to be protected*.

The compliance officer can then review requests, grand or refuse access, and keep records of authorization history and purposes within the catalog. This information can then be used to *check who uses and has access to which data*.

As a whole, the ability to obtain an overview of the storage and usage of personal data enterprise-wide enables compliance officers to *prove the compliance of data usage*.

4.2.7 Data architect

Data architects define how data should be stored and consumed across enterprise systems (Weber et al. 2009), and work closely with solution architects. They provide and *manage data models* and metadata for different stakeholders, based on an assessment of how data is used across systems. By considering data at different layers (i.e., logical, conceptual and physical), and based on inputs from solution architects, they ensure that data models are correctly mapped, and *link business definitions to the physical layer*.

4.2.8 Solution architect

Solution architects (or IT architects) implement business requirements as software solutions (Weber et al. 2009). In doing so, they align software-specific concepts with enterprise and data architecture, as well as with security policies.

As input for their word, they can use the data catalog to *retrieve available documentation* on data (provided by data architects) and *discover the technical schemas for specific systems*. In doing so, they can uncover the types of data that reside in specific systems, e.g., personal data, and mark them so that other relevant roles (e.g., compliance officer) can handle them accordingly.

In order for an enterprise data catalog to provide a single, unified view on enterprise data, solution architects can leverage visualization and lineage functionalities to *map data schemas between systems*, and conceptualize the way these data resources are used by providing an *understanding of cross-system data life cycles*.

5 Data catalogs as platforms for data collaborations

As a single point of entry for all data-related activities, enterprise data catalogs act as an integrating platform, and promote the convergence of activities. In that sense, to support the increased usage of data, they should not only provide access to enterprise-wide data resources (i.e., remediating technical data silos), but also enable cross-functional communication and sharing (i.e., remediating organizational data silos). To better understand how enterprise data catalogs facilitate data democratization, we investigate data-related collaborations between various user roles and building upon their individual user stories. These collaborations are hereby conceptualized as use case scenarios, since they involve several roles, follow a specific sequence, and depict user interactions with specific enterprise data catalog functionalities (Wautelet et al. 2016). In the following, we will illustrate four concrete use case scenarios – they are described as vignettes, to condense and present our empirical data analysis in a meaningful way (Urquhart 2001).

5.1 Creating, sharing and re-use of data stories

This vignette is set in the context of business intelligence and analytics, in an enterprise that manufactures connected, portable appliances. It illustrates how data analysts can use data catalogs throughout the entire process of creating data stories, i.e., selecting appropriate data, building the story, documenting it, sharing it, and interacting with other enterprise users.

In this situation, the data analyst may want to explore all data resources related to a device's battery, from specification and manufacturing to real-time data from live devices. Here, the data analyst could investigate potential causes of battery drain in live devices using battery-related data points, as well as data from the device sensors (e.g., location, temperature), also accessible in the catalog. In the organization, said insights are typically shared with other stakeholders (e.g., chief data officer, data citizen), but the knowledge around the analysis process itself usually is not.

Through data analytics functionalities, data catalogs enable data analysts to *document* this analysis process in the form of *data stories*. They would contain a summary of data sources that were selected, of transformations that were performed on them, as well as comments on the insights that were derived (e.g., what they mean, how they should be interpreted).

These data stories can be *recommended* to relevant users through the data catalog, as they *search* for datasets, through *data subscriptions* that they have signed up to, or through *sharing*. Users could also comment on these data stories, initiating a conversation with the analyst. By exposing knowledge and encouraging communication between user groups, data catalog support both the democratization of data itself, and of data knowledge.

5.2 Data quality root cause

This vignette relates to a common, data quality-related situation, where business users encounter issues caused by defective data, and need to escalate it to the data management organization.

It starts with a data citizen, who identifies a data quality defects in a dataset that she/he is working with, for instance as a result of a failure in an automated business process. Using process-to-data drill-down functionalities, she/he can isolate the impacted data element. Through collaboration features in the data catalog, he/she is then able to identify the person responsible for the impacted dataset (data stewards) and initiates a request for clarification.

The data steward then makes use of metadata documentation and data assessment functionalities (such as lineage) and visualization (such as data flows) functionalities to identify which system and data source have contributed to the failure.

5.3 Impact analysis

In this vignette, we illustrate how enterprise data catalogs can facilitate communication in preparation to upstream changes to critical or widely used datasets.

Data stewards are responsible for the continuous monitoring, maintenance of data. As data resources are harmonized and made accessible through the single access point of the data catalog, they may need to implement improvement measures. Here, a data steward intends on making changes to a specific dataset.

For instance, he/she might want to harmonize reference data across several business functions, in order to ease future maintenance.

To select which datasets should be taken care of in priority, the data steward can leverage *data assessment* functionalities in the data catalog – in doing so, he/she can isolate datasets that are most used (*data usage*) and/or that suffer from defects (*data quality*).

Before making any changes, he/she needs to assess the technical impact of planned changes, e.g. on business processes or metrics and key performance indicators that rely on the reference data that will be changed. For this purpose, he/she can use *data visualization* functionalities, and *drill-down* from data to processes/reports, or the other way around.

From an organizational perspective, the data steward may want to engage with other business stakeholder (e.g., data owners or data citizen) that routinely use the data that will be change, in order to inform them and gather issues and requirements. For this purpose, he/she can use the *data profiling* (*data assessment*) functionality, in conjunction with the documented *roles and responsibilities* (*data governance*) in the data catalog, to identify those stakeholders. Using collaboration functionalities, the data steward could also set up a dedicated page outlining the planned changes, giving other stakeholders the opportunity to rate and comment them.

5.4 Data protection

In this vignette, we take a specific instance of a compliance officer, and specifically consider the situation of a data protection officer who is in charge of implementing requirements from the GDPR.

Although data documentation and review are an important part of compliance (s. Section 4.3.6), data catalogs can also support authorization *workflows* for accessing data resources.

Assuming that all data resources containing personal information are correctly flagged and documented, data citizens and analysts can request access to these resources through the data catalog, e.g., for use in a data story or for integration in a business process. As part of this *workflow*, they can outline the details of the purpose for which they plan to use the data.

Upon receiving the request, the data protection officer can review whether these purposes fall within the boundaries of the authorized personal data processing activities (based on contracts, user consent and/or legitimate for instance). As a result of this assessment, she/he can grant or refuse the access to the personal dataset (*data authorizations*).

He/she can also document *rules and policies* on datasets containing personal data, and include the reasoning behind granting or refusals of pasts requests, as a way to proactively inform data citizen and data analysts of what they are allowed to do with personal data.

6 Discussion and outlook

Data democratization is a means to facilitate the leveraging of data value. It is a transformation process that aims at empowering all employees, including data novices, to work with data.

Based on a design research science approach, we have identified 8 data-related roles, specified several concrete user stories for each role, and outlined collaborative use case scenarios, supported by enterprise data catalogs. In doing so, we shed light on the necessary alignment of a suitable underlying technical solution (here: enterprise data catalogs) with the modes of working and related requirements of intended users (through user stories and collaborative use case scenarios).

Our study makes two main contributions: First, the identified roles extend beyond usual data experts, reconciling data supply, curation, and demand. The emergence of these roles and the intention to democratize data illustrate the impact of the increasing data abundance and its governance on professions and skills. They complement existing research, that typically focusses on roles in the context of data governance and extends them to embrace non-experts and data citizens. They help understand user needs and provide a link with enterprise data catalog functions and function groups, which clarifies their involvement in supporting data democratization, and realize the FAIR principles in the enterprise context.

Second, we show that enterprise data catalogs go beyond documenting and publishing data, i.e., make data FAIR. They also promote collaboration between users, either by assigning ownership and responsibilities, or by enabling communication between various stakeholders (e.g., by collaborative features such as tagging and commenting). The user stories and collaborative use case scenarios demonstrate that data catalogs can improve the value generation of data assets and increase data transparency. Transparency should not only apply to datasets, but also to the way they are used. Target-oriented workflows are support by a data catalog. For instance, on the demand side, a data citizen searches for the right data for a specific task. He or she reviews several datasets and finally submits a request to use the selected dataset. On the supply side, the data architect onboards a new data source and categorizes the data. As a kind of data broker that matches supply and demand the data owner documents metadata information of this dataset and grants access and usage rights of this requested dataset to the data citizens. These collaborative use case scenarios are a defining aspect of data democratization in enterprises compared to FAIR principles implementations in academic settings.

Our findings contribute to the ongoing academic debate on data democratization and the FAIR principles in the enterprise context. As we are among the first to conduct research on data-related roles and their specific data needs in the context of data democratization, there are certainly limitations and needs for future research. First, the roles we have identified are tailored to the use of data catalogs. Future research could develop a role model that specifies a generally applicable model for digital transformation, including, for example, roles in enterprise analytics platforms. Second, as enterprises are just beginning to implement and apply these roles and data management principles such as FAIR, further empirical insights are useful to underpin and improve our findings.

With regards to the FAIR principles, we recognize that becoming a data-driven enterprise is rather a strategic, organizational imperative, which needs to be communicated to employees, especially data citizen, so that they are incentivized to incorporate data into their daily activities. In other words, future research should focus on ways to increase data literacy of employees throughout the enterprise, as a way to support data democratization. Furthermore, reaping the benefits requires, professionals, especially former data novices, need to acquire new skills. Even though advanced data documentation tools like enterprise data catalogs feature facilitating functionalities, training courses and on-the-job training are required to empower all employees to efficiently work with data. In any case, excellent data management that enables data monetization calls for data democratization to realize the untapped value potential of data.

References

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.

Bano, M., and Zowghi, D. 2015. “A Systematic Review on the Relationship between User Involvement and System Success,” *Information and Software Technology* (58), pp. 148–169. (https://doi.org/10.1016/j.infsof.2014.06.011).

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., and Venkatraman, N. 2013. “Digital Business Strategy: Toward a Next Generation of Insights,” *MIS Quarterly* (37:2), pp. 471–482.

Borgman, C. L. 2003. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*, MIT Press.

Brownlow, J., Zaki, M., Neely, A., and Urmetzer, F. 2015. “Data-Driven Business Models: A Blueprint for Innovation,” University of Cambridge, Cambridge Service Alliance, May 14. (https://doi.org/10.13140/RG.2.1.2233.2320).

Calhoun, K. 2014. *Exploring Digital Libraries: Foundations, Practice, Prospects*, London, UK: Facet Publishing.

DAMA International. 2017. *Data Management Body of Knowledge (DAMA-DMBOK)*, (2nd Edition.), (D. Henderson, S. Earley, and L. Sebastian-Coleman, eds.), Basking Ridge, NJ: Technics Publications.

Díaz, A., Rowshankish, K., and Saleh, T. 2018. “Why Data Culture Matters,” *The McKinsey Quarterly* (3), p. 37.

Dryzek, J. S. 1996. “Political Inclusion and the Dynamics of Democratization,” *The American Political Science Review* (90:3), pp. 475–487. (https://doi.org/10.2307/2082603).

Dwivedi, Y. K., Wastell, D., Laumer, S., Henriksen, H. Z., Myers, M. D., Bunker, D., Elbanna, A., Ravishankar, M. N., and Srivastava, S. C. 2015. “Research on Information Systems Failures and Successes: Status Update and Future Directions,” *Information Systems Frontiers* (17:1), pp. 143–157. (https://doi.org/10.1007/s10796-014-9500-y).

Fadler, M., and Legner, C. 2020. “Who Owns Data in the Enterprise? Rethinking Data Ownership in Times of Big Data and Analytics,” in *Proceedings of the 28th European Conference on Information Systems (ECIS)*, Marrakesh, Morocco, June 15, p. 207.

Feng, J. 2017. “How Airbnb Democratizes Data Science With Data University,” *Medium*, , June 3. (https://medium.com/airbnb-engineering/how-airbnb-democratizes-data-science-with-data-university-3eccc71e073a, accessed December 17, 2020).

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).

Gilson, F., Galster, M., and Georis, F. 2020. “Generating Use Case Scenarios from User Stories,” in *Proceedings of the International Conference on Software and System Processes (ICSSP)*, Seoul, South Korea, June 26, pp. 31–40. (https://doi.org/10.1145/3379177.3388895).

Goetz, M., Leganza, G., Hoberman, E., and Hartig, K. 2018. “The Forrester Wave^{TM}: Machine Learning Data Catalogs, Q2 2018,” Consortium Report, Consortium Report, Forrester Research.

Gould, D. 1996. “Using Vignettes to Collect Data for Nursing Research Studies: How Valid Are the Findings?,” *Journal of Clinical Nursing* (5:4), pp. 207–212. (https://doi.org/10.1111/j.1365-2702.1996.tb00253.x).

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.

Grugel, J., and Bishop, M. L. 2013. *Democratization: A Critical Introduction*, Macmillan International Higher Education.

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)*, San Francisco, California, USA, July 26, pp. 795–806. (https://doi.org/10.1145/2882903.2903730).

Hughes, R., and Huby, M. 2004. "The Construction and Interpretation of Vignettes in Social Research," *Social Work and Social Sciences Review* (11:1), pp. 36–51. (https://doi.org/10.1921/swssr.v11i1.428).

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://www.journal.ecrc.nsysu.edu.tw/index.php/pajais/article/view/526).

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., Kaliyaperumal, R., Kersloot, M. G., Kirkpatrick, C. R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.-A., da Silva Santos, L. O. B., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M. D., Willighagen, E. L., Wittenburg, P., Roos, M., Mons, B., and Schultes, E. 2019. "FAIR Principles: Interpretations and Implementation Considerations," *Data Intelligence* (2:1–2), MIT Press, pp. 10–29. (https://doi.org/10.1162/dint_r_00024).

Jarke, M., Loucopoulos, P., Lyytinen, K., Mylopoulos, J., and Robinson, W. 2011. "The Brave New World of Design Requirements," *Information Systems* (36:7), Special Issue: Advanced Information Systems Engineering (CAiSE'10), pp. 992–1008. (https://doi.org/10.1016/j.is.2011.04.003).

Kauffman, C. M. 2018. "Democratization," *Britannica Academic*. (https://academic.eb.com/levels/collegiate/article/democratization/602971).

Khatri, V., and Brown, C. V. 2010. "Designing Data Governance," *Communication of the ACM* (53:1), pp. 148–152.

Korhonen, J. J., Melleri, I., Hiekkanen, K., and Helenius, M. 2013. "Designing Data Governance Structure: An Organizational Perspective," *GSTF Journal on Computing (JoC)* (2:4). (http://dl6.globalstf.org/index.php/joc/article/view/576).

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).

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).

Licklider, J. C. R. 1965. *Libraries of the Future*, Cambrige, Massachusetts, USA: The M.I.T. Press.

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.

Lucassen, G., Dalpiaz, F., Werf, J. M. E. M. van der, and Brinkkemper, S. 2016. "The Use and Effectiveness of User Stories in Practice," in *Proceedings of the 22nd İnternational Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ)*, Gothenburg, Sweden, March 14, pp. 205–222. (https://doi.org/10.1007/978-3-319-30282-9_14).

Marabelli, M., and Galliers, R. D. 2017. "A Reflection on Information Systems Strategizing: The Role of Power and Everyday Practices," *Information Systems Journal* (27:3), pp. 347–366. (https://doi.org/10.1111/isj.12110).

Marr, B. 2017. "What Is Data Democratization? A Super Simple Explanation And The Key Pros And Cons," *Forbes*, , July 24. (https://www.forbes.com/sites/bernardmarr/2017/07/24/what-is-data-democratization-a-super-simple-explanation-and-the-key-pros-and-cons/, accessed February 7, 2020).

McKendrick, J. 2018. "Intelligent Data Catalogs: At The Forefront Of Digital Transformation," Practitioner Report, Practitioner Report, Forbes Insights. (http://info.forbes.com/rs/790-SNV-353/images/Informatica_Intelligent%20Data%20Catalogs.pdf).

Nuseibeh, B., and Easterbrook, S. 2000. "Requirements Engineering: A Roadmap," in *Proceedings of the 22nd International Conference on Software Engineering (ICSE)*, Limerick, Ireland: ACM, June 4, pp. 35–46. (https://doi.org/10.1145/336512.336523).

Ofner, M., Otto, B., and Österle, H. 2013. "A Maturity Model for Enterprise Data Quality Management," *Enterprise Modelling and Information Systems Architecture – International Journal of Conceptual Modeling* (8:2), pp. 4–24. (https://doi.org/10.18417/emisa.8.2.1).

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. "A Morphology of the Organisation of Data Governance," in *Proceedings of the 19th European Conference on Information Systems (ECIS)*, V. Tuunainen, J. Nandhakumar, M. Rossi, and W. Soliman (eds.), Helsinki, Finland, June 9. (http://aisel.aisnet.org/ecis2011/272/).

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).

van Reisen, M., Stokmans, M., Basajja, M., Ong'ayo, A. O., Kirkpatrick, C., and Mons, B. 2019. "Towards the Tipping Point for FAIR Implementation," *Data Intelligence* (2:1–2), MIT Press, pp. 264–275. (https://doi.org/10.1162/dint_a_00049).

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).

Schlagwein, D., Conboy, K., Feller, J., Leimeister, J. M., and Morgan, L. 2017. "'Openness' with and without Information Technology: A Framework and a Brief History," *Journal of Information*

Technology (32:4), SAGE Publications Ltd, pp. 297–305. (https://doi.org/10.1057/s41265-017-0049-3).

Schüritz, R., Seebacher, S., and Dorner, R. 2017. "Capturing Value from Data: Revenue Models for Data-Driven Services," in *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*, Waikoloa Village, Hawaii, USA, January 4, pp. 5348–5357. (https://doi.org/10.24251/HICSS.2017.648).

Urquhart, C. 2001. "Bridging Information Requirements and Information Needs Assessment: Do Scenarios and Vignettes Provide a Link?," *Information Research* (6:2).

Urquhart, C., Light, A., Thomas, R., Barker, A., Yeoman, A., Cooper, J., Armstrong, C., Fenton, R., Lonsdale, R., and Spink, S. 2003. "Critical Incident Technique and Explicitation Interviewing in Studies of Information Behavior," *Library & Information Science Research* (25:1), pp. 63–88. (https://doi.org/10.1016/S0740-8188(02)00166-4).

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. 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.

Wautelet, Y., Heng, S., Hintea, D., Kolp, M., and Poelmans, S. 2016. "Bridging User Story Sets with the Use Case Model," in *Advances in Conceptual Modeling Workshop (ER)*, Gifu, Japan, November 14, pp. 127–138. (https://doi.org/10.1007/978-3-319-47717-6_11).

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.

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), p. 160018. (https://doi.org/10.1038/sdata.2016.18).

Williams, C. C. 2017. "Democratizing Data at Airbnb," *Medium*, , May 12. (https://medium.com/airbnb-engineering/democratizing-data-at-airbnb-852d76c51770, accessed December 17, 2020).

Wixom, B. H., Yen, B., and Relich, M. 2013. "Maximizing Value from Business Analytics," *MIS Quarterly Executive* (12:2), p. 13.

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

Yuhanna, N., Leganza, G., and Hoberman, E. 2019. "Big Data Fabric 2.0 Drives Data Democratization," Consortium Report, Consortium Report, Forrester Research.

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).

Empowering Data Consumers to Work with Data: Data Documentation for the Enterprise Context

Clément Labadie, Markus Eurich, and Christine Legner

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

Published in the Proceedings of the 15th International Conference on Wirtschaftsinformatik (WI), 2020

Abstract. Enterprises that are engaging in digital transformation need to empower an increasing number of data consumers (sometimes referred to as "data citizens") to work with data. A prerequisite is data documentation – data assets should be inventoried and well-described to facilitate data selection by non-data experts, who need to both find and understand them. This research paper proposes a reference model for data documentation in the enterprise context. It was developed in collaboration with 25 large enterprises, following a Design Science Research process. Compared to existing metadata standards that contain flat lists of metadata attributes, the reference model organizes metadata objects in logical and physical layers and features views dedicated to usage and governance contexts. It thereby improves maintenance and consistency in data documentation, when dealing with hundreds of interdependent data resources, and allows to express inherent relationships between metadata attributes.

Keywords: metadata, data documentation, reference model, design science research

Table of contents

List of figures

List of tables

1 Introduction

“Data consumers don't speak physical and logical. They speak business (Goetz et al. n.d.).” This statement echoes the larger, well-documented phenomenon of the digital transformation, prompting enterprises to integrate digital resources as part of their value creation process (Bharadwaj et al. 2013) and rethink their strategy accordingly (Matt et al. 2015). With the digital transformation, data has become a critical asset for enterprises (Pentek et al. 2017; Mohr and Hürtgen 2018; Belissent et al. 2019). In research, this trend has found an echo in the growing body of knowledge on Data-Driven Business Models (DDBMs) (Brownlow et al. 2015; Schüritz et al. 2017), which are “designed to create additional business value by extracting, refining and ultimately capitalizing on data (Brownlow et al. 2015)”.

Enterprises that are engaging in digital transformation need to empower an increasing number of data consumers to work with data. These data consumers include traditional data experts (e.g., business intelligence specialists, data managers or data architects), but increasingly comprise employees who start employing self-service business intelligence (sometimes referred to as “data citizens”) or more advanced data science tools. In order to enable their activities, data that is spread in enterprise systems need to be discoverable for humans and machines alike. These challenges, i.e., extracting data from various sources and refining them to make them ready-for-use have also be conceptualized through the FAIR principles, according to which data should be Findable, Accessible, Interoperable and Reusable (Wilkinson et al. 2016).

A prerequisite for making data FAIR is data documentation – data assets should be inventoried and well-described to facilitate data use by non-data experts, who need to both find and understand them. Metadata, i.e., data about data (Inmon et al. 2010; Roszkiewicz 2010), is a natural candidate when it comes to documenting data. It is a long-standing topic that features a large number of standards, such as Dublin Core (DC) (Dublin Core Metadata Initiative n.d.) and the Data Catalog Vocabulary (DCAT) (World Wide Web Consortium (W3C) n.d.). While these standards emphasize data discovery, they are unable to accurately address the specific requirements of the enterprise context, which brings added layers of complexity, both in terms of systems (e.g., highly distributed and siloed applications) and organization (e.g., governance, roles and responsibilities).

Conversely, our study aims at answering the following research question: how to organize data documentation to support data discovery and data use in the enterprise context? To that end, following a Design Science Research (DSR) process, we developed a metadata model in close collaboration with 25 multi-national enterprises with varying levels of experience in setting up enterprise-wide data documentation, using metadata and data catalog tools. As a reference model, this model is intended to serve as blueprint for enterprises seeking to design their own, company-specific metadata model in support of providing data documentation for the increasing number of data consumers (i.e., both data experts and data citizens). Compared to existing metadata standards that contain flat lists of attributes, the suggested metadata model groups objects in logical and physical layers and features views dedicated to usage and governance contexts.

The remainder of this paper is structured as follows. We start with an analysis of metadata-related research and standards, and by further specifying the research gap. We then detail our research methodology and related steps that went into the model's development. We continue by presenting the reference model for data documentation, its components. We conclude by presenting the model's demonstration and discussing its contribution.

2 Background and motivation

2.1 Metadata definition and categories

To this day, data documentation is most often associated with metadata, that are commonly defined as "data about data" (Roszkiewicz 2010), and "aim at facilitating access, management and sharing of large sets of structured and/or unstructured data" (Kerhervé and Gerbé 1997). Various initiatives have attempted to describe them for specific applications, e.g., earth sciences or multimedia systems (Kerhervé and Gerbé 1997). Although these initiatives make context-dependent suggestions for organizing metadata, general sub-categories can be identified (Hillmann et al. 2008; Inmon et al. 2010; Marco 2000):

- **Structural** metadata describes the general data model, e.g., type, attributes of objects and relationships between objects.
- **Administrative** metadata provides information to help manage a resource, e.g., users (with rights) and dates (creation, last update).
- **Terminological** metadata provides an understanding on the data, e.g., definitions, abbreviations, cataloging records and comments from creators and users.
- **Governance** metadata provides an overview of the data landscape from a management point of view, e.g., ownership, roles, responsibilities and level of confidentiality (Roszkiewicz 2010).

- **Context** metadata provides information on the environment in which the data exists, e.g., business processes and business purposes (use cases).
- **Use** metadata provides information on how data are consumed, e.g., search logs, usage statistics, processing systems.

The literature highlights that metadata has found a variety of technical applications over the years (Sen 2004), starting from the bibliography and library domains (Hillmann et al. 2008). Yet, the ones that relate to the enterprise context tend to focus on technical descriptions (Sen 2004). However, in an enterprise context, metadata should exceed technical aspects and describe business aspects of data, such as their use in business processes (Burnett et al. 1999).

2.2 Metadata standards

As mentioned above, metadata is often discussed in specific contexts, leading to a wide variety of metadata standards. In order to acquire a broad overview, we performed a wide-ranging review of existing standards from academic and non-academic sources, resulting in a list of 129 standards. Resources such as the Research Data Alliance's directory of metadata standards (Chen et al. n.d.) were used as a starting point and the list was enriched with standards highlighted in academia (Clobridge 2010; Ferguson and Hebels 2003; Hillmann et al. 2008; Inmon et al. 2010; Vetterli et al. 2000; Xiao et al. 2015). To also embrace non-academic sources, we used search string "metadata standard" on Google Search and Google Scholar. The identified metadata standards cover a vast array of topics ranging from archiving (e.g., Open Archival Information System - OAIS), bibliography (e.g., Library of Congress Classification and Subject Headings – LCC, LCSH), cultural material (e.g., Categories for the Description of Works of Art – CDWA), multimedia content (e.g., Synchronized Multimedia Integration Language – SMIL) and geographical data (e.g., ISO 19115). These standards address different aspects of metadata, and they can be classified in the following types (Baca 2016):

- **Data structure** standards define metadata element sets and schemas, which are categories (or containers) of metadata. They define a predefined set of attributes meant to describe objects pertaining to the domain of interest. *Example: Library of Congress Classification – LCC.*
- **Data value** standards are controlled vocabularies, which define terms, names and other values used to populate metadata elements, i.e., they restrict the value domain to fill metadata structures. *Example: Library of Congress d Subject Headings –LCSH.*

- **Data content** standards define cataloging rules and codes, which provide guidelines for the format and syntax of the values used to populate metadata elements. *Example: Synchronized Multimedia Integration Language – SMIL.*
- **Data format / technical interchange** standards define the encoding in machine-readable form of metadata elements, i.e., a manifestation of a particular data standard, encoded and marked up for machine processing. *Example: Extensible Markup Language – XML.*

Out of the 129 identified standards, we excluded all domain-specific standards (e.g., related to bibliography/libraries, archiving/preservation, finance, cultural material, natural sciences, geography, medicine, climate, museum curation, government, education, agriculture, astronomy), which resulted in 18 remaining standards. We then further refined our selection by excluding standards solely specifying data formats or technical interchange and encoding schemes. As a result, four standards were retained as relevant for the enterprise context and data catalogs (s. Table 20 below): 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).

Table 20. Overview of selected metadata standards

Name	Source	Metadata standard type	Domain
Dublin Core (DC)	Working group (DCMI)	Structure	General
Data Catalog Vocabulary (DCAT)	Standards organization (W3C)	Structure, Format	Data catalogs
Common Warehouse Metamodel (CWM)	Industry consortium (OMG)	Structure, Format	Business intelligence
Metadata Registry Metamodel and Basic Attributes (MDR)	Standards organization (ISO/IEC)	Structure	General

Dublin Core presents a flat list of terms (or attributes), comprising, e.g., creator, description, date, identifier, relation and rights. DCAT and MDR go a step further by grouping attributes and specifying relationships between groups in a metamodel, which focuses on the internal logic between the concepts they introduce, but do not integrate external concepts. Finally, CWM provides a metadata interchange standard that enables XML-based exchange of business intelligence and data warehouse metadata between different tools, platforms and repositories.

In terms of metadata categories (s. Section 2.1 above), all comprise structural, administrative and terminological metadata. DCAT and CWM additionally include context metadata, but none of them specifically cover governance and use metadata, which are critical in an enterprise context.

2.3 Research gap

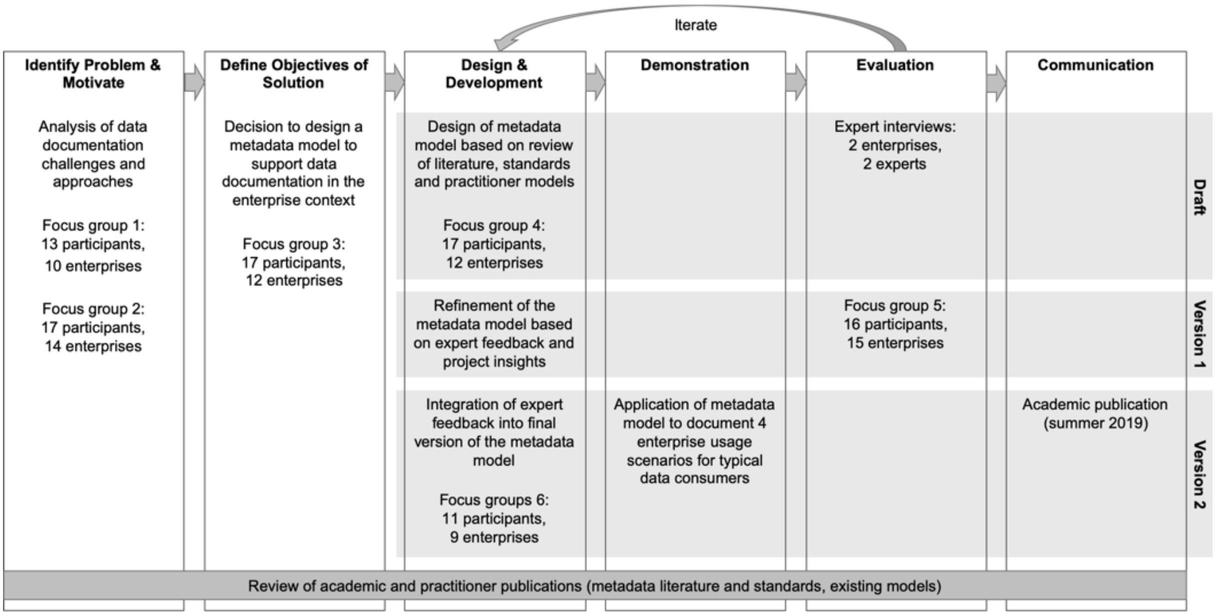
Although available standards may constitute adequate starting points, DC and MDR focus on metadata for describing single data resources or datasets but are not entirely suitable for the enterprise context. In fact, DC and MDR provide fundamental metadata attributes for describing data resources, but they are presented as a flat list. This makes it difficult to maintain them, when dealing with hundreds of data resources, and to express inherent relationships between metadata attributes. In that regard, it is not surprising that DC has been extended in further domain-specific applications[11]. On the other hand, standards such as CWM focus on a specific aspect of metadata management (in CWM's case, business intelligence), but do not cover the overall enterprise context and specifically the business processes that create and consume data. DCAT goes further and introduces metadata objects with relationships but adopts a data publication perspective, as it is meant to link data catalogs available on the web. While all of them provide useful building blocks for metadata management, they lack an integrating, enterprise-driven framework.

3 Research process

To provide a reference framework for data documentation in the enterprise context, we developed a metadata model following the iterative DSR process suggested by (Peffers et al. 2007). Figure 9 depicts all research steps and highlights the interaction between the research team and practitioners, which took place in the form of 6 focus group meetings with representatives from 25 large enterprises, as well as 2 expert interviews. Participants in the focus groups had extensive experience with data documentation and were involved in metadata and data catalog initiatives to support data discovery and use in their enterprises. Focus group meetings were part of a collaborative effort between researchers and a stable core team of practitioners over the course of more than 12 months. This collaboration included reviews of

[11] For instance, the Visual Resource Association Core Categories (VRA Core) and the Australian Government Locator Service (AGLS) both rely on DC, extending it for cultural material and government domains, respectively.

metadata standards and analysis of existing practitioner models and led to the development of the metadata model.


Iterate
Review of academic and practitioner publications (metadata literature and standards, existing models)

	Identify Problem & Motivate	Define Objectives of Solution	Design & Development	Demonstration	Evaluation	Communication
Draft	Analysis of data documentation challenges and approaches Focus group 1: 13 participants, 10 enterprises Focus group 2: 17 participants, 14 enterprises	Decision to design a metadata model to support data documentation in the enterprise context Focus group 3: 17 participants, 12 enterprises	Design of metadata model based on review of literature, standards and practitioner models Focus group 4: 17 participants, 12 enterprises		Expert interviews: 2 enterprises, 2 experts	
Version 1			Refinement of the metadata model based on expert feedback and project insights		Focus group 5: 16 participants, 15 enterprises	
Version 2			Integration of expert feedback into final version of the metadata model Focus groups 6: 11 participants, 9 enterprises	Application of metadata model to document 4 enterprise usage scenarios for typical data consumers		Academic publication (summer 2019)

Figure 9. Research process

During the problem identification phase, we analyzed the issues faced by practitioners in democratizing data within their enterprise and documenting data for various purposes. In focus group 1, we identified and validated typical data consumers, i.e., for which enterprise stakeholders data should be documented. In focus group 2, we outlined data usage of these typical consumers, as well as the type of data documentation to support them. These discussions lead to the conclusion that there is a need for a core understanding of fundamental data objects to be documented, as well as several extensions, describing more specific data objects that may not apply to all business contexts (either company-specific or domain-specific).

In the objective definition phase, the decision to design a metadata model was made (focus group 3). This model is meant to represent the core model for enterprise data documentation, i.e., the recommended minimum approach, that addresses the requirements of different types of data consumers. The core metadata model is meant to serve as a blueprint for enterprise-wide data documentation, for instance in the context of data catalog initiatives.

The subsequent phases consisted of iterative design cycles, each comprising design and evaluation steps, with the last cycle also featuring a demonstration of the finalized model. The first design iteration incorporated insights from our review of metadata standards (specifically Dublin Core and DCAT), as well as existing models from practice at different level of maturity, i.e., a completed model and two models in development, as depicted in **Erreur ! Source du**

renvoi introuvable.. It resulted in a draft model as alpha artifact that was discussed in a broader group during focus group 4. It was also evaluated through expert interviews from Company B and C which were in the process of developing their approach to data documentation. We also reviewed data documentation models from both enterprises and performed a mapping with our draft version.

The second design iteration led to the first stable version of the metadata model (version 1). We consolidated feedback from focus group 4 and both expert interviews, and incorporated insights from two additional practitioner models (s. Table 21). Version 1 was discussed during focus group 5, and it was decided to improve terminological clarity and put an emphasis on establishing shared metadata object definitions in the next iteration.

The third design iteration led to the final version of the metadata model (version 2). In this step, we reworked the model according to participant feedback and refined the definition of metadata objects, which were extensively discussed during focus group 6. Once all metadata objects were agreed upon, we demonstrated the metadata model by applying it to document 4 enterprise usage scenarios.

Table 21. Input models for artefact development

Source	Type	Level of maturity	Context / industry	Version (design iteration)
DCAT	Standard (W3C)	Established	Facilitating interoperability between data catalogs published on the web	Draft (first iteration)
Dublin Core	Standard (DCMI)	Established	Digital resources	Draft (first iteration)
Company A	Practitioner	Finalized	Pharmaceuticals	Draft (first iteration)
Company B	Practitioner	In development	Packaging	Draft (first iteration)
Company C	Practitioner	In development	IT	Draft (first iteration)
Company D	Practitioner	In development	Manufacturing	Version 1 (second iteration)
Company E	Practitioner	In development	Automation	Version 1 (second iteration)

4 A metadata model for enterprise data documentation

4.1 Model objectives

As shown in section 2, existing approaches to data documentation do not fully address the specific requirements and the complexity of the enterprise-wide context. Out of the relevant standards identified, DC and MDR focus on metadata describing single data resources or datasets. On the other hand, DCAT and CWM do address a multiplicity of data sources but focus on the online distribution of datasets and on data interchange between warehouse and business intelligence platforms, respectively. The enterprise context is characterized by complex data landscapes, i.e., large numbers of interdependent data resources stored in enterprise systems, that underlie strong data governance in order to ensure data quality and comply with internal and external standards and regulations. In the course of digitalization, enterprises need to enable an increasing number and variety of data consumers, with different data needs and degrees of data literacy, to access and use these data resources.

Conversely, our proposed model is designed as a reference framework for data documentation in an enterprise context, and addresses the following requirements that were identified in the focus group sessions: First, it should support data democratization and provide data documentation for typical data consumers (both experts and non-experts, e.g., data citizens, data analysts, data protection officers, data architects, data stewards, data owners). Second, it should align the different perspectives on data, specifically the business-oriented and the system-oriented perspectives. Third, it was found that for using data in an enterprise context, data consumers need to also understand responsibilities and relevant regulations for data. Lastly, the reference model should identify and structure metadata attributes in a "core model" for data documentation in an enterprise context, i.e. act as a baseline of minimum requirements for all enterprises, regardless of their area of business. The core model could then be extended to include documentation aspects dedicated to specific enterprise functions, such as metadata specific to research and development, for instance.

The suggested model should be able to support data catalog implementation and use cases, which we demonstrate in section 4.4 **Erreur ! Source du renvoi introuvable.**. As for implementation, we recommend a use-case driven approach, where enterprises would select and prioritize the metadata objects and attributes that need to be implemented for a specific business purpose.

4.2 Model structure

The reference model for enterprise data documentation is expressed in form of a metadata model (Kerhervé and Gerbé 1997), comprising relevant metadata objects that are to be documented as well as their relationships. This allows for distinguishing different perspectives on data and for defining relationships between data and other relevant objects in the enterprise context. Defining data documentation based on a metadata model also eases maintenance and improves consistency of data documentation. The metadata model, which is depicted by Figure 10, is organized in different views, starting with the logical and physical view on the data at the center.

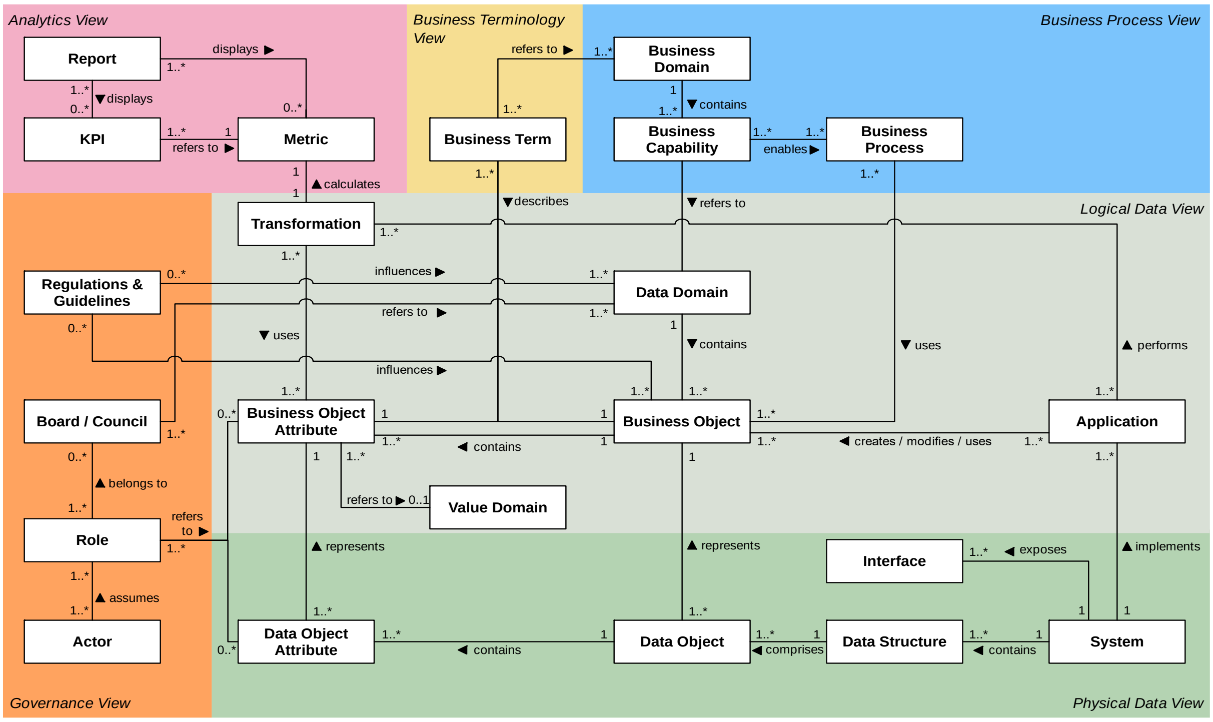


Figure 10. Metadata model overview

The **logical view** represents the conceptual or business view on data, whereas the **physical view** represents the implementation perspective and makes the link with the way data is implemented in systems. The other views 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., business process, analytics, and the related business terminology). The **governance view** depicts the relevant regulations, guidelines and responsibilities for data. Each view comprises several metadata objects, which we will define in the next section.

4.3 Metadata objects

The central element[12] of the model is the *business object*, which describes a reoccurring set of information used in multiple business contexts and minimum one data domain. It can be either created, used, or changed in business processes (Martin 1977) and analyzed in reports that aggregate metrics and KPIs. A business object is specified by business *object attributes* that are characteristics of the *business object* and can contain either free text or a restricted set of values (*value domain*). For instance, in an enterprise context, a *business object* can be a record of a supplier – *attributes* could comprise, e.g., name, street, zip code, country, VAT number, and the country attribute's restricted value domain could consist in a list of ISO country codes. Additionally, *business objects* belong to a *data domain*, which specifies their context – in our example, a supplier is part of the business partner domain. Finally, *business objects* can be created, used or changed by *applications*.

Business object and *business object attribute* are both reflected on the physical view, representing their specific implementation in the respective *systems*, with the *data object* and the *data object attribute*, respectively. They are projected into a *data structure*, which unifies *data objects* and *data objects attributes* into a single, distinct format, stored in a *system*. Depending on the database paradigm, the structure can represent, e.g., a table (relational database), a class (object-oriented database) or a key value store (graph database). *Systems* are the physical counterpart of *applications* and expose interfaces, which are meant to transfer data to other *systems*.

Moving onto the usage context, *business objects* are used within *business processes*, either as input or output. *Processes* represent how an enterprise performs its activities, and are enabled by *business capabilities*, which consist in a combination of technological, informational, and organizational resources, and representing what a company does. The business domain represents strategic business areas of an enterprise and reflect is strategic goals. In our example (i.e., supplier as *business object*), procurement is the business domain. It features business capabilities such as a global sourcing capability consisting, e.g., of an ERP system and/or supplier relationship management system (technology), a purchasing team (organization) – it is realized through e.g., strategic sourcing and procure-to-pay *business processes*, which make use of supplier (business partner *data domain*) and product (material *data domain*) *business objects* (information).

Business domains contain several, related *business objects* and *business object attributes*. In addition, *business terms* specify synonyms or alternative expressions for business objects and

[12] In this section, the term "element" is used in substitute for the term "metadata object".

attributes. Since they are domain-specific, *business terms* are necessarily linked with at least one *business domain*. For instance, the term “debtor” may be used within procurement and financial *business domains* as synonym for "supplier".

On the logical data view, the *transformation* is the gateway to the analytics view. It queries data from one or several *business object attributes* to produce a *metric*, which is a quantifiable measure reflecting the state of the enterprise; in our example, the number of defective parts received from suppliers. *Metrics* are the input for *key performance indicators* (KPIs), which evaluate the success of the enterprise at specific activities, and show the degree of fulfillment of a *metric*, with regards to a stated objective. A *KPI* expresses this number in terms of e.g., percentages, setting a threshold stating that a deficiency rate higher than e.g., 5% is not acceptable. Finally, reports organize and present metrics and / or KPIs in human-readable form, enabling visualization by different dimensions. In our example, a *report* can then display this information by several dimensions, e.g., supplier name, supplier location, date and / or material category.

The governance view features organizational and regulation aspects. On the organizational side, the *actor* assumes certain data-related *roles*. According to their *role* assignment, *actors* refer to designated *business object attributes* and *data object attributes* in the performance of their tasks and responsibilities. For instance, the *actor* Jane Doe bears the *role* of lead strategic buyer – in this context, she is responsible for overseeing negotiations with suppliers, and thus interacts with *attributes* such as product specifications, name and reference price. Several *roles* can be involved in *boards and councils*, which designate working groups bearing advisory and / or decision-making power, with regards to one or more *data domain*. In our example, Jane Doe may be part of a data stewardship council setting guidelines for the procurement *data domain*.

The governance view also contains an element depicting *regulations and guidelines*. It designates any guideline, or set of guidelines, that constrain the structure and / or behavior of an enterprise (El Kharbili 2012). It can refer to legal texts (e.g., the General Data Protection Regulation – GDPR), contracts (which may specify binding service level agreements) and standardization documents (e.g., use of standardized customs codes in shipping documents) (El Kharbili 2012), among others. Requirements can apply to either entire data domains (e.g., GDPR applies to all data domains containing personally identifiable information, such as the business partner data domain) or specific business objects (e.g., product safety regulations apply solely to the materials data domain).

In the following section, we will demonstrate how these metadata object come into play when realizing selected use cases.

4.4 Demonstration

In order to demonstrate the metadata model, we opted for a use-case driven approach. During focus group 5, we gathered usage scenarios from 8 companies representing 8 industries. They depict how typical data consumer groups (e.g., data analysts, data stewards, data owners, data protection officers, data architects and data citizen) may benefit from data documentation that is organized according to the metadata model. The demonstration's purpose is to show how our proposed model may empower data consumers to find and use appropriate data, suited to their needs and issues. Based on practitioner input from the focus groups, we have retained 4 key usage scenarios (s. Table 22): 1) selecting appropriate data for analysis, 2) understanding governance impact of planned data-related changes, 3) investigating data-related business process failure, and 4) assessing data protection requirements.

Table 22. Summary of usage scenarios

Designation	Role	Purpose	Metadata objects
Selecting data for reports	Data analyst, data citizen	- Find metadata definitions - Locate relevant business objects - Generate reports	Business term, Business object, Business object attribute, Application, Transformation, Report
Understanding impact of data changes	Data steward	- Find stakeholders impacted by planned data change - Identify relationship between business objects	Data domain, Business object attribute, Data object attribute, Value domain, Role, Actor, Board / Council
Investigation process failures	Process owner	- Identify faulty data - Business process drill-down - Data quality reporting	Business process, business object, Business object attribute, Data object attribute, Application, Role, Actor, Metric, Key Performance Indicator
Assessing data protection requirements	Data protection officer	- Locate compliance sensitive data - Document policies and business rules - Identify international data transmissions	Regulations & Guidelines, Data domain, Business term, Business process, Business object, Business object attribute, Data object attribute, System,

The first use case is set in the context of business intelligence and analytics, in an enterprise that transforms from traditional manufacturing towards producing connected, portable appliances. A data analyst wishes to understand possible causes of battery drain, and needs to extract battery-related information, as well as data from the appliance's sensors (e.g., location, temperature). Without proper data documentation, the data analyst would have to take guesses as to where (i.e., in which *system*) to find appropriate information (*data object* and *attributes*).

Additionally, the way *data objects* and *attributes* are defined and phrased in the system may not be transparent. Using data documentation relying on our metadata model, the analyst could browse definitions from *business terms* in order to isolate *business (and data) objects* containing the information they are looking for, and trace it back to designated *applications and systems*, eventually enabling them to select the relevant data objects and attributes for the analysis or check existing *reports* and the way *metrics* and *KPIs* are calculated.

In the second use case, the emphasis is placed on the governance context. Here, a data steward plans to make changes to an *attribute*, e.g., by restricting the *value domain* in order to remedy data quality issues. This will be reflected in an update of the internal *guidelines*. Before implementing any change, the data steward needs to understand who will be impacted by planned changes, in order to involve affected stakeholders. This is made possible by the metadata model, as it connects *roles* with the *business object attributes* and *data object attributes* they refer to. Additionally, the data steward could reach out to any *board / council* in charge, by tracing the *attribute* back to its parent *data domain*.

The third use case, centered around investigating business process failures, follows a similar rationale of impact-analysis, this time with regards to the business process context. Here, a process owner witnesses errors in business process, and wishes to investigate its root cause. Starting from the *business process* in question, the metadata model enables him or her to discover *business objects* and related *attributes* used, as well as the processing *application*. The process owner can also find out how business objects and attributes are implemented in different systems, with their respective *data objects* and *attributes*. In order to identify and remedy data defects, process owners could also apply data quality *metrics* or *KPIs*, and identify *roles* and *actors* responsible for defective *objects* or *attributes* to prompt rectifying measures.

The fourth use case focuses on regulatory issues, specifically on data protection, e.g., in the context of GDPR. Here, a data protection officer needs to compile a list of where personally identifiable data is stored, and how it is used, as well as document compliance requirements. The latter can be achieved through specification of the *regulations & guidelines* metadata object, e.g., by documenting business rules and policies. In order to identify affected *business objects* and *business object attributes*, *business terms* (e.g., specifying personal data), *data domains* (e.g., business partner inherently contains personal data) and *business processes* (e.g., customer account management processes necessarily deal with personal information) can be a starting point. Furthermore, thanks to the logical – physical *attribute* inheritance built into the metadata model, a link can be made to the specific *systems* processing the data of interest. In the context of data protection, this is crucial for identifying processing systems located in non-EU countries, to which additional GDPR requirements apply.

5 Discussion and outlook

In the course of the digital transformation, democratizing and generalizing the use of data is on top of the management agenda in many organizations that set out to enhance business processes, inform business decisions and implement data-driven business models. In this context, this research presents enterprises with a reference model meant to act as the foundation of data documentation and support the implementation of tools to empower data consumers to work with data. The suggested metadata model reflects the enterprise context and provides a business-oriented view on data. Compared to existing standards such as DC and MDR, which consist of flat lists of metadata attributes, it provides enterprises with a structure for organizing metadata objects and their relationships. In contrast to CWM, our model has not primarily been developed to standardize the interorganizational exchange of metadata information. Instead, its major focus is on providing relevant context information around data. While the available models may support organizations in finding (i.e., data discovery) data, the suggested metadata model introduces logical and physical views on data and comprises usage and governance contexts. Specifically, it enriches existing, generic models by addressing enterprise-specific contextual aspects, i.e., business processes, business glossary, business intelligence and analytics. It also integrates the governance perspective and suggests metadata objects representing both organizational roles and relevant guidelines and regulations. Our research thereby contributes to providing a holistic perspective on data in the enterprise context and links metadata concepts to enterprise (data) architecture literature.

As previously stated, the model has been assessed and partially tested in selected cases, e.g., by performing a mapping with existing, company-specific enterprise metadata models. Moreover, it has been tested for four concrete usage scenarios. While these steps provide evidence for the validity, consistency and completeness of the model, the next research steps include a broader evaluation of the final model, both artificial (e.g., questionnaire-based) and naturalistic (e.g., implementation of the metadata model in a corporate setting).

The metadata model presented in this paper constitutes a core model representing fundamental data documentation concepts, and thereby is also meant to provide a foundation for future research. Most importantly, we are interested in understanding how data documentation - and its publication in data catalogs - impacts data citizens' satisfaction, productivity and quality of work. Interesting avenues for future research relate to the design of extensions relying on the core, providing metadata objects and attributes more specifically catered to specific business areas (e.g., research and development, finance). Future research could also address overlaps of the presented metadata model with enterprise architecture repositories. Although we have observed these overlaps within our focus group activities, we have not yet elaborated on them in more detail. However, we have witnessed that in practice, data catalog tools can be populated with information maintained in enterprise architecture management tools, i.e. applications, infrastructure (incl. interfaces) and processes.

In addition, while unstructured data (e.g., video clips, audio files) was not prominent within our focus group discussions, we appreciate the growing need to include this aspect in future iterations of the metadata model.

Acknowledgements

This research was supported by the Competence Center Corporate Data Quality (CC CDQ). The authors would like to thank the experts that contributed to this research.

References

Baca, M. 2016. *Introduction to Metadata*, (Third Edition.), Los Angeles, CA, USA: The Getty Research Institute.

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. (https://www.forrester.com/report/Determine+Your+Datas+Worth+Data+Plus+Use+Equals+Value/-/E-RES127541).

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., and Venkatraman, N. 2013. "Digital Business Strategy: Toward a Next Generation of Insights," *MIS Quarterly* (37:2), pp. 471–482.

Brownlow, J., Zaki, M., Neely, A., and Urmetzer, F. 2015. "Data-Driven Business Models: A Blueprint for Innovation," University of Cambridge, Cambridge Service Alliance, May 14. (https://doi.org/10.13140/RG.2.1.2233.2320).

Burnett, K., Ng, K. B., and Park, S. 1999. "A Comparison of the Two Traditions of Metadata Development," *Journal of the American Society for Information Science* (50:13), pp. 1209–1217. (https://doi.org/10.1002/(SICI)1097-4571(1999)50:13<1209::AID-ASI6>3.0.CO;2-Y).

Chen, S., Alderete, K. A., and Ball, A. (n.d.). "RDA Metadata Standards Directory." (https://rd-alliance.github.io/metadata-directory/standards/, accessed November 19, 2019).

Clobridge, A. 2010. "Metadata," in *Building a Digital Repository Program with Limited Resources*, Chandos Information Professional Series, A. Clobridge (ed.), Chandos Publishing, pp. 85–109. (https://doi.org/10.1016/B978-1-84334-596-1.50005-5).

Dublin Core Metadata Initiative. (n.d.). "DCMI: DCMI Metadata Terms." (https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, accessed August 25, 2019).

El Kharbili, M. 2012. "Business Process Regulatory Compliance Management Solution Frameworks: A Comparative Evaluation," in *Proceedings of the Eighth Asia-Pacific Conference on Conceptual Modelling - Volume 130*, APCCM '12, Darlinghurst, Australia, Australia: Australian Computer Society, Inc., pp. 23–32. (http://dl.acm.org/citation.cfm?id=2523782.2523786).

Ferguson, S., and Hebels, R. 2003. "Access to Information Resources," in *Computers for Librarians (Third Edition)*, Topics in Australasian Library and Information Studies, S. Ferguson and R. Hebels (eds.), Chandos Publishing, pp. 81–109. (https://doi.org/10.1016/B978-1-876938-60-4.50009-4).

Goetz, M., Leganza, G., Hoberman, E., and Vale, J. (n.d.). "STIR Your Data For Context," Consortium Report, Consortium Report, Forrester Research. (https://www.forrester.com/report/STIR+Your+Data+For+Context/-/E-RES145620).

Hillmann, D. I., Marker, R., and Brady, C. 2008. "Metadata Standards and Applications," *The Serials Librarian* (54:1–2), pp. 7–21. (https://doi.org/10.1080/03615260801973364).

Inmon, W. H., O'Neil, B., and Fryman, L. 2010. *Business Metadata: Capturing Enterprise Knowledge*, Morgan Kaufmann.

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).

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.

Marco, D. 2000. *Building and Managing the Meta Data Repository - A Full Lifecycle Guide*, Hoboken, NJ, USA: John Wiley & Sons.

Martin, J. 1977. *Computer Data-Base Organization*, Engelwood Cliffs, NJ, USA: Prentice Hall.

Matt, C., Hess, T., and Benlian, A. 2015. "Digital Transformation Strategies," *Business & Information Systems Engineering* (57:5), pp. 339–343. (https://doi.org/10.1007/s12599-015-0401-5).

Mohr, N., and Hürtgen, H. 2018. "Achieving Business Impact with Data," Consultancy Report, Consultancy Report, Digital McKinsey, April. (https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/achieving-business-impact-with-data).

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).

Pentek, T., Legner, C., and Otto, B. 2017. "Towards a Reference Model for Data Management in the Digital Economy," in *Proceedings of the 12th International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, Karlsruhe, Germany, May. (https://doi.org/10.5445/IR/1000069452).

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.

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).

Schüritz, R., Seebacher, S., and Dorner, R. 2017. "Capturing Value from Data: Revenue Models for Data-Driven Services," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, Waikoloa Village, Hawaii, USA, January 4, pp. 5348–5357. (https://doi.org/10.24251/HICSS.2017.648).

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).

Vetterli, T., Vaduva, A., and Staudt, M. 2000. "Metadata Standards for Data Warehousing: Open Information Model vs. Common Warehouse Metadata," *SIGMOD Rec.* (29:3), pp. 68–75. (https://doi.org/10.1145/362084.362138).

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), p. 160018. (https://doi.org/10.1038/sdata.2016.18).

World Wide Web Consortium (W3C). (n.d.). "Data Catalog Vocabulary (DCAT)." (https://www.w3.org/TR/vocab-dcat/, accessed August 25, 2019).

Xiao, B., Zhang, C., Mao, Y., and Qian, G. 2015. "Review and Exploration of Metadata Management in Data Warehouse," in *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)*, , June, pp. 928–933. (https://doi.org/10.1109/ICIEA.2015.7334243).

FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs

Clément Labadie, Markus Eurich, Christine Legner, and Martin Fadler

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

Published in the Proceedings of the 22nd IEEE Conference on Business Informatics (CBI), 2020

Abstract. With increasing relevance of data as a strategic asset, companies strive to make data FAIR, i.e. findable, accessible, interoperable and reusable. Data catalogs are considered an important means to realize these aspirations. However, data catalogs are still a novel and loosely defined concept, which lacks empirical studies on their implementation. Against this background, this study aims at fostering the understanding of data catalogs as a means of increasing data exposure and usage in enterprises. Based on a qualitative and explorative study involving 12 multi-national enterprises, we assess data catalog initiatives based on their scope, goals, and users. We propose a taxonomy of data catalog initiatives and present 3 detailed case studies that illustrate typical approaches to data catalogs. Our findings contribute to the ongoing discourse on the FAIR principles by elaborating on their significance in the enterprise context and analyzing their operationalization by means of data catalogs.

Keywords: FAIR principles, data catalog, data democratization, data management, metadata

Table of contents

List of tables

1 Introduction

In the age of digital transformation, data has become a valuable resource and the keystone for new business models, decision-making and value creation (George et al. 2014). Encouraged by the huge business potential of data (Wixom and Ross 2017), enterprises strive for making better use of their data. They face two challenges: First, data quantities in their storage systems are ever increasing. For instance, according to analysts, the proportion of companies reporting data volumes exceeding 1'000 terabytes almost tripled between 2016 and 2017 (Goetz et al. 2018). Much of this data is never used – for instance, it has been reported that between 60% and 73% of enterprise data goes unused (Gualtieri et al. 2016). Second, companies can only unlock the value from data if it is well maintained, trusted, and used broadly by a wide range of employees. This implies making data available not only to data specialists, but also to so-called “data citizens” (employees who use data on a regular basis to fulfill their daily jobs). Prerequisites to encouraging data citizens to work with data and to unlocking value from data are reflected in the FAIR principles. The latter claim that data should be findable, accessible, interoperable and reusable and were originally proposed by over fifty researchers for scientific data (Wilkinson et al. 2016). Recently, the FAIR principles have gained much popularity in the enterprise context and companies started to turn to data catalogs as means to enforce the FAIR principles. Broadly defined, a data catalog "maintains an inventory of data assets through the discovery, description, and organization of datasets” (Zaidi et al. 2017). However, despite their increasing popularity and their usefulness as data documentation tools, data catalogs are still barely addressed in scientific literature, which does not yet discuss data catalogs as specific class of applications to enforce these principles. Particularly, empirical studies on how data catalogs are implemented are missing.

Against this background, this study aims at understanding how companies leverage data catalogs to make their enterprise data FAIR. In particular, we investigate the research question “How can data catalogs support the FAIR principles in the enterprise context?”. To answer our research question, we use an explorative and qualitative research design involving twelve multinational enterprises. In a first step, we assess their data catalog initiatives based on their scope, goals, and users and organize our findings in the form of a taxonomy. We then present three detailed case studies that illustrate typical approaches to data catalogs.

From an academic perspective, our findings shed light on the role of data catalogs in enterprises and their roles in bringing the FAIR principles to the enterprise context. From a practical perspective, they provide data managers and business decision makers with a means to understand key components if data catalogs to develop and drive sustainable data catalog initiatives.

The remainder of this article is structured as follows: after this introduction, we first provide an overview on related work namely on the FAIR principles and on the evolution of metadata management. In the section that follows, we explain the research methodology of this study. Subsequently, our empirical findings of data catalog initiatives are presented. The paper ends with concluding remarks and suggestions for future research.

2 Related work

Enterprise data is often hidden in silos, and many enterprises have started to engage in enhancing data usage. Data democratization in an enterprise context is a fairly recent topic, which has seen very few conceptualizations in research. It can be defined as "the process of empowering a group of users spanning beyond established data experts to access and use data, by removing obstacles to data exploration and sharing in enterprises" (Labadie et al. 2020), and represents the enterprise interpretation of the FAIR principles. In the following, we analyze it through the FAIR principles and elaborate on the role of data catalogs in enterprises.

2.1 The FAIR principles

The need for better data documentation has been echoed in both business and academia. In research, a 2016 paper co-authored by more than fifty researchers unveiled the FAIR principles (Wilkinson et al. 2016), according to which data should be findable, accessible, interoperable and reusable – they "describe distinct considerations for contemporary data publishing environments with respect to supporting both manual and automated deposition, exploration, sharing, and reuse." Each principle is specified by a set of data-related requirements.

The motivation for these principles is rooted in the following observation: when researchers require data to answer a given research question, they need to invest substantial amounts of time (i.e., several weeks or months) in the data gathering process. The FAIR principles were designed to alleviate this effort along the various steps of data gathering processes.

First, the *findable* element addresses the search activity itself, i.e., figuring out what data exists and where it exists. It is about ensuring that users are provided with the means to find data within the repository. This requires that data are associated with identifiers and are indexed in a searchable source, which presents them with a suitable description of their contents. Once suitable data sources have been identified, the *accessible* element requires that they are made available to users, e.g., through downloads. Here, it is about ensuring that users are provided with the means to access the data within the repository. The *interoperable* element further requires that data should be encapsulated in standardized, commonly used formats. The *reusable* element points to the documentation of the data, which goes beyond the description that is comprised in the *findable* element. Here, this description should not only enable users to find and identify data, but to provide all necessary contextual information so that users can understand them (e.g., detailed description about tables, columns, attributes).

From a technical perspective, the FAIR principles are closely related to metadata – often defined as "data about data" (Roszkiewicz 2010). The objectives of metadata are well aligned with those of the FAIR principles, as they "aim at facilitating access, management and sharing of large sets of structured and/or unstructured data (Kerhervé and Gerbé 1997)". In fact, the FAIR principles authors specifically mention metadata in 13 of the 15 requirements specifying the 4 principles (Wilkinson et al. 2016). Taken together, the FAIR principles are meant to enable data exploration, sharing, and reuse. They are formulated in both domain- and technology-independent terms but are applicable to a variety of contexts. In the next section, we address how the FAIR principles are translated to the enterprise context.

2.2 The FAIR principles in the enterprise context: evolution of metadata documentation towards data catalogs

In recent years, companies have started to view data as a strategic asset. To extract value from data, organizations are required to analyze and centralize their existing data assets with new technologies and architectures. Similar to the Findable element of the FAIR principles, indexing and documenting said data assets, as in a catalog, is a prerequisite to conducting those tasks.

What is called a "data catalog" nowadays is a result of the evolution of different concepts that companies have established for data documentation and data provisioning since the early days of data processing in the 1960s (s. Fig. 1). To this day, data documentation is most often associated with metadata, that are commonly defined as "data about data" (Sen 2004) and "aim at facilitating access, management and sharing of large sets of structured and/or unstructured data" (Kerhervé and Gerbé 1997). It all began with field names (1960s) and table definitions (1970s), before system-specific data dictionaries emerged during the 1980s. These data dictionaries were used for basic technical documentation of database tables. With the emergence of enterprise resource planning systems, more emphasis was put on business process integration and the system landscapes grew more and more complex. Over time, it became necessary to plan the data architecture as well as data integration more carefully and link them to the business needs. As it became increasingly important to analyze business data to facilitate decision-making, companies established data warehouses for central storage of business data. In addition, they began documenting these business data in metadata repositories. Subsequently, they introduced business data dictionaries as extension of the former technical-oriented data dictionaries as an answer for the increasing distribution of data storage due to market-ready cloud technologies. To define semantics and allow users correctly interpret data for different use cases, documentation approaches of using business glossaries were implemented.

Data catalogs can be viewed as the next step in the evolution of concepts for data documentation and data provisioning. Due to its novel nature, a widely accepted definition of the term "data catalog" is still missing. The first definition originates from the computer science community. Researchers made the observation that data ecosystems evolve from standalone databases to a heterogeneous set of systems to store and analyze data (Franklin et al. 2005). They argue that in spite of being stored across a variety of systems (i.e., multiple versions of the truth), data still need to be managed as though they were stored in a single database (i.e., single version of the truth). Based on these assumptions, they define a data catalog as follows: "A catalog is 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). Coming from the computer science domain, this definition mainly considers technical aspects, while it neglects the usage aspect.

Over ten years later, with data catalogs becoming of increasing interest for data managers, market research analysts widened the definition to also embrace users and usage contexts: "a data catalog maintains an inventory of data assets through the discovery, description and organization of datasets. The catalog provides context to enable data analysts, data scientists, data stewards and other data consumers to find and understand a relevant dataset for the purpose of extracting business value" (Zaidi et al. 2017). This definition emphasizes the discovery and understanding of datasets by different user groups. The same authors define key desired capabilities that data catalog solutions should enable and facilitate *curation/inventory* of data assets, built-in *collaboration* for accountability and governance, as well as *communication* for shared semantic meaning. However, this definition does not reflect how data can be obtained and the access controlled, which becomes more and more important in view of ever-increasing data protection concerns.

Another definition from the practitioner's community describes the data catalog as a modern approach that "takes metadata management from its backwater silos to a centralized cross-platform facility that is feature-rich and comprehensive" (Russom 2017). This definition emphasizes the platform character of a data catalog.

2.3 Research gap

As highlighted in this section, data catalogs are the latest abstraction of metadata documentation, but there is a lack of academic studies dedicated to this concept. Despite the relevance of data catalogs described in the practice-oriented literature, we have little empirical insights on how these emerging tools are being adopted and implemented. Furthermore, the FAIR principles are rooted in an academic perspective on data, which we argue differs from an enterprise perspective. Therefore, this study aims at fostering the understanding of data catalogs as a means to make enterprise data FAIR and support data democratization.

3 Research design

As data catalogs are a novel concept that has seen little coverage in the extant literature, we opted for an explorative research approach and employed mixed methods. Mixing methods has advantages over using one method as multiple modes of analysis are more likely to create new insights (Kaplan and Duchon 1988; Venkatesh et al. 2013).

For understanding the motivation and scope of data catalog initiatives in practice, we conducted an expert study and collected data from interviews, focus group meetings and a questionnaire-based survey. We consolidated our findings into a taxonomy, meant to synthetize and shed some light on data catalog implementation practices. It was designed using a taxonomy development method as described by (Nickerson et al. 2013), which is suitable for structuring emerging concepts in the Information Systems field (Beinke et al. 2018; Püschel et al. 2016).

3.1 Research context

Our research was conducted in a consortium research program (Österle and Otto 2010) and entailed close interactions with 17 senior data management experts from 12 multi-national organizations over almost one year. All participants were involved in either evaluating or implementing a data catalog in their enterprises – Table 1 presents an overview of participating organizations. Between January and November 2019, we conducted interviews with each company, collected additional data from questionnaire-based survey and held 5 focus group meetings to interpret and discuss the findings.

As part of our research activities, we also conducted desk research – we reviewed analyst reports and identified a broader range of tools that are used as data catalogs, including specialized data catalog tools (from vendors such as Collibra, Alation, Cambridge Semantics, Informatica, Waterline, Ab Initio, IBM, Oracle and SAP) as well as metadata management and data governance tools (Goetz et al. 2018; Peyret et al. 2017; Zaidi et al. 2017). Out of approximately 100 identified solutions, and following feedback from our research partners regarding their priorities, we isolated 15 for further analysis. This analysis provided insights into the target user groups and functionalities provided by data catalog solutions. The results of these activities were synthesized and presented to participants during the first focus group, in order to create a shared understanding of the data catalog concept.

Table 23. Participating organizations

Company	Industry	Revenue range	Expert (title)	Catalog purpose	Status
A	Adhesives	1 – 50 B €	Lead Data Architect	Metadata management	Implementation in progress
B	Automation	1 – 50 B €	Head of Corporate Master Data Management	Support for data governance	Tool selection
C	Chemistry	50 – 100 B €	Data Catalog Product Owner and Enterprise Architect	Support for data governance and analytics	Implementation in progress
D	Fashion and jewelry	1 – 50 B €	Data glossary manager	Data glossary	Proof-of-concept (PoC)
E	Information technology	1 – 50 B €	Solution Advisor Expert	Support for data governance, metadata management and data analytics	Implemented
F	Manufacturing	1 – 50 B €	Corporate Data Management	Metadata management	Transition to another tool
G	Manufacturing	50 – 100 B €	Enterprise Architect for IoT and Digitalization	Metadata management	Implementation in progress
H	Packaging	1 – 50 B €	Global Master Data Driver	Support of data governance, analytics, inventory and automation	Tool selection
I	Pharmaceuticals	1 – 50 B €	Associate Director Information Management	Support for data analytics	Tool selection
J	Pharmaceuticals	1 – 50 B €	Business Data Analyst, Global Data Team, Ops IT	Support for data governance and data analytics	Implementation in progress

Company	Industry	Revenue range	Expert (title)	Catalog purpose	Status
K	Retail	100+ B €	Team Lead Master Data Management	Support for data governance	Tool selection, PoC
L	Tobacco	50 – 100 B €	Manager Enterprise Data Governance System	Support for data governance	Implemented

3.2 Taxonomy development

To assess and compare the motivations and approaches of our participating companies in implementing a data catalog, we developed a taxonomy to characterize data catalog initiatives based on their scope, goals, and users.

According to the taxonomy development methodology (Nickerson et al. 2013), a meta-criterion must first be defined to steer the definition of the taxonomy's dimensions and characteristics. We defined the "prevalent aspects for characterizing data catalogs initiatives" as meta-criterion. We selected five meta-characteristics that describe data catalog implementation approaches: 1) scoping and goals, 2) user groups, 3) functionalities, 4) data documentation, 5) tools. Then, dimensions and characteristics for the taxonomy must be identified through iterations, which may be either conceptual-to-empirical (i.e., starting from literature, then evaluated with empirical data) or empirical-to-conceptual (i.e., starting from empirical data, then evaluated with literature). We conducted three iterations to construct our analysis framework for data catalog initiatives.

For the first iteration, an empirical-to-conceptual approach was applied – we leveraged early focus group discussions to design a structured questionnaire meant to describe the objectives, audiences, approaches and tool support of our participants' data catalog initiatives. This led to the definition of 12 initial dimensions representing the four meta-characteristics for scoping & goals, user groups, data documentation and tools.

The second iteration also featured an empirical-to-conceptual approach. We examined data catalog initiatives in each participating organization individually, based on interviews. We analyzed and consolidated their answers, which enabled us to define characteristics for each of the 12 initial dimensions. At this point, we evaluated termination conditions as defined by (Nickerson et al. 2013), both internally and with our participants group. We concluded that the taxonomy was missing a characterization of data catalog functionalities, which are strongly linked with the objectives, goals and implementation characteristics of data catalogs.

This triggered a third iteration, which we conducted following a conceptual-to-empirical approach. We added functionalities as a fifth meta-characteristic and revisited our market analysis of data catalog solutions. This led to the addition of 9 dimensions focused on data catalog functionalities, as well as related characteristics for each of them.

3.3 Taxonomy evaluation and case studies

After the taxonomy stabilized, participating companies positioned their approach by selecting, for each dimension, the characteristics that they considered most relevant and/or desirable for their company. We used a questionnaire to collect this information and consolidated the results in the form of a heat map cluster, which we present in section 4 below (s. Table 2).

Based on the data collected, the research team condensed common patterns in the implementation approaches and their respective challenges and success factors and presented it in iterative steps to the participants, who in turn provided feedback and complementary information. Besides the identification of common patterns among the data catalog initiatives, three distinctive cases were identified that are representative of typical implementation goals and approaches (see section 5). For each of these three approaches, we selected the company that was already at an advanced level of data catalog implementation. These cases were documented in a three-step approach: first, experts from each company presented their approach to the team. Then, each approached was summarized and characterized according to the analysis framework. Finally, the documentation of the results was sent back to each expert, who was asked to validate them and if necessary, to make changes and provide further insights and feedback. The three case studies showcase typical ways of using data catalogs and can be related to different aspects of the FAIR principles.

4 Data catalogs in practice

Since data catalogs remain a novel and loosely defined concept, we used the taxonomy to collect empirical insights on data catalog initiatives. The final version of the taxonomy comprises 19 dimensions, clustered along five meta-characteristics (categories), which we will present in the following: 1) scoping and goals, 2) user groups, 3) functionalities, 4) data documentation, 5) tools. We will present each category along with the empirical insights into data catalog initiatives from 11 companies (s. Table 2).

The *scoping and goals* category is meant to describe the reach, i.e., the scale at which an organization intends to deploy a data catalog, as well as the main objectives that it is pursuing through a data catalog implementation. In this category, we witnessed that most participants (82%) aimed to implement an enterprise-wide data catalog, i.e., one that would make data accessible to all employees across functions and divisions. This reflects a willingness to position data catalogs as over-arching platform and overcome the data silo effect that often occurs in large organizations with highly distributed operations and IT landscapes. This is in line with the Findable element of the FAIR principles, in that the data catalog platform acts as a single, indexed and searchable source for data. In that same spirit, transparency, accessibility and increased data usage were often nominated as implementation goals, while compliance and risk lag behind significantly.

As data catalogs address several user groups – from data scientists and other data experts to data citizens, we were interested in understanding the target audience of a data catalog (*user groups*). All participating companies expressed a broad-reaching objective that is reflected in the unanimous nomination of a broad audience for the data catalog, i.e., extending beyond data experts and also targeting business users. This is echoed by ongoing digitalization activities in enterprises, which result from strategic pushes towards creating business value through data-driven insights and/or moving towards data-based product offerings (e.g., internet of things and connected devices). This broad target audience is also in line with the Accessible element of the FAIR principle, in the extended sense of the term – data should not only be accessible from a technical perspective, but also from an organizational one. On a detailed perspective, participants were asked to select specific roles that they meant to be users of the data catalogs, ranging from traditional data expert roles, responsible for onboarding, structuring, and maintaining data (i.e., data architect, solution architect, data owner, data steward) to new, less specialized roles, consisting of data consumers who use data to support an organization's business purposes and strategic goals (i.e., data citizen, business analyst, data analyst, data engineer and chief data officer). Consistent with the broad audience aspect highlighted above, an equal number of roles representing both categories received high nomination counts (i.e., 73% and above).

Since functionalities of a data catalog solution are very broad, we presented the participants a set of typical functionalities and asked them about their *functional scope*.

Following our market analysis of data catalogs, we outlined the following function groups:

- Data discovery covers functionalities allowing users to find data, mirroring the FAIR principles. They comprise search, browsing and recommendation (Borgman 2003; Franklin et al. 2005; Halevy et al. 2016).

- Data inventory covers functionalities supporting the data documentation required to present users with a single view on enterprise data (Hellerstein et al. 2017; Sen 2004). It relates to the Findable aspect of the FAIR principles.
- Administration covers functionalities for user management and system configuration.
- Data assessment covers functionalities supporting the measurement of metrics related to enterprise data, such as data usage and data quality.
- Data governance covers functionalities supporting a governance organization, e.g., assigning roles and responsibilities (Otto 2011) and setting up workflows.
- Data collaboration covers functionalities enabling users to communicate, work together and enrich the data within the tool (Lagoze et al. 2005).
- Data visualization covers functionalities enabling users to display enterprise data and lineage in meaningful way and get a better grasp on them.
- Data analytics covers algorithms, tools and workflows to gain insight on datasets and generate new data (Wilkinson et al. 2016).
- Automation and machine learning spans over all previous function groups, providing ways to automate related tasks.

Overall, data search, data tagging and workflows are the most cited functionalities. In contrast, we observe that analytics and automation aspects are the least cited among participants. We can hypothesize that such tasks would be handled in dedicated applications – in this case, the data catalog would act purely as a way to find and retrieve relevant data. On the other hand, high nomination counts for search and tagging again highlight that a data catalog is primarily envisioned a means to expose data resources. This is in line with the findable element of the FAIR principles. As for workflows, they express the accessible and reusable elements of the FAIR principle. In fact, workflows may either intervene when a user requests access to a data source (accessible) or when changing data descriptions within the catalog (reusable). In both instances, approval processes would be triggered based on regulatory and governance requirements. In the automation category, the automatic classification and tagging of data was the most cited functionality. This aspect is also stressed by the authors of the FAIR principles (Wilkinson et al. 2016), as it contributes to making data documentation easier. In large enterprises, this classification aspect is all the more relevant, due to the multiplicity and distributed character of systems and databases. It would enable such enterprises to scale their data catalog more efficiently.

Regarding data documentation, data catalogs offer three layers at which data should be documented: logical, physical and conceptual. The physical layer reflects the implementation view on data and represents the way it is organized and stored in enterprise systems (e.g., database tables). The logical layer represents the business view on data, in which core data domains as well as related business object and their attributes are documented, along with the applications that create, change and use them. The conceptual layer comprises several views depicting the ways data is used and governed in an enterprise context, e.g., in terms of processes, capabilities, definitions, guidelines, roles and responsibilities. As data catalogs cover all three layers and are meant to act as an abstraction platform between them to offer a single access point on data to users, it is consistent that they have all received high nomination counts (i.e., 73% and above).

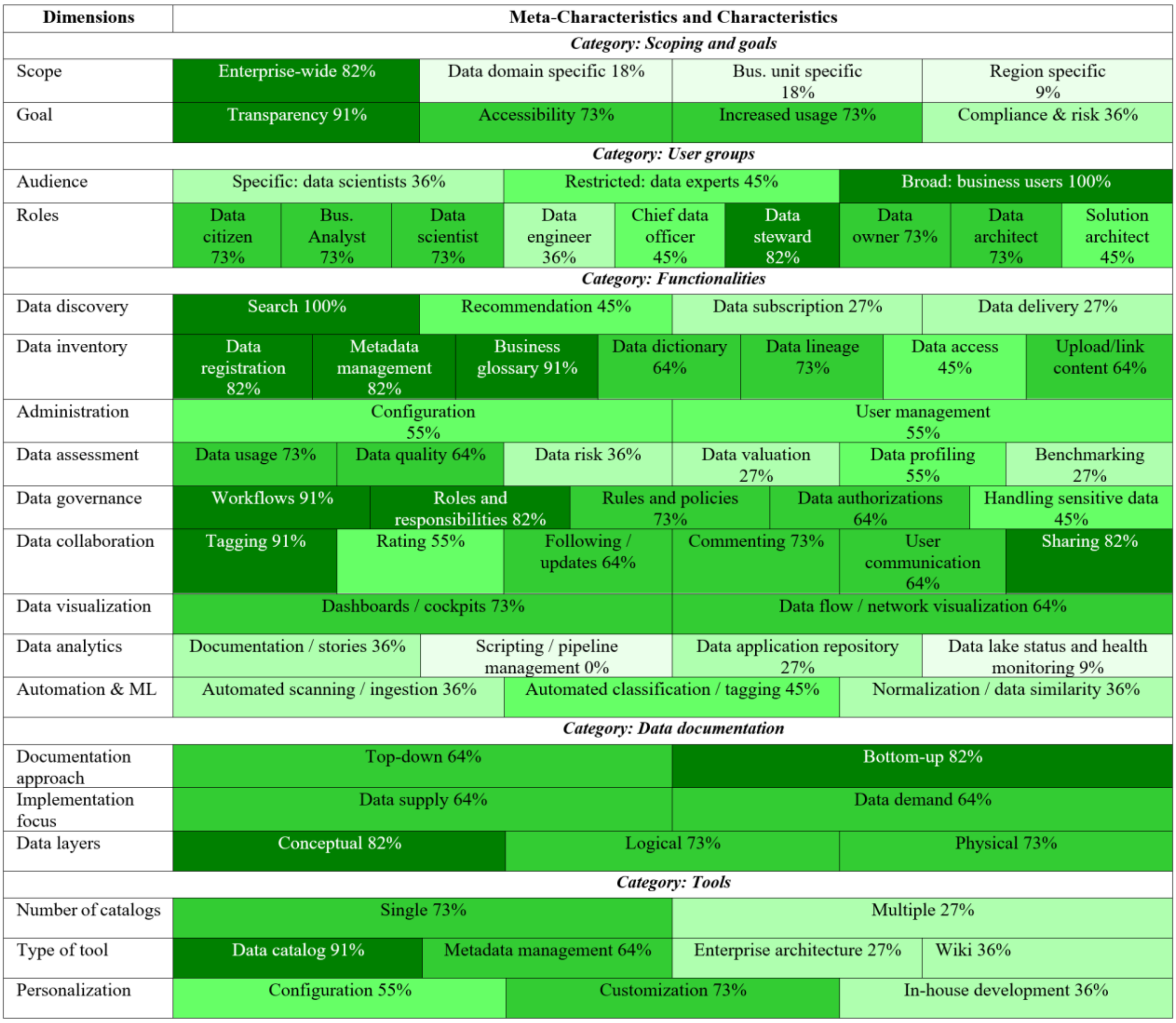
Since creating a complete data documentation is challenging and takes time, companies apply different approaches, based on the order in which tasks were performed. In the top-down approach, a company would start by defining the structure of the documentation and the high-level enterprise data model, and then proceed with collecting and sorting appropriate data to populate the platform. In the bottom-up, a company would start by importing bulk data resources into the tool and define documentation structure based on exploration and analysis of the data. Then, from a practical implementation perspective, a data supply-driven approach means that a company focuses its implementation efforts based on the input for the data catalog and prioritizes the requirements of users that will provide and maintain data and information in the data catalog. On the other hand, a data-demand approach signals a focus on the output of the data catalog and prioritizes the requirements of users that will use data and information contained in the data catalog for further business activities.

With regards to the technical implementation choices (tools), we asked about the number of catalogs, type of tool, and degree of personalization for tools. Here, nearly all participants stated that they envisioned to implement a single-catalog environment, powered by a dedicated data catalog solution, that they would either configure or customize. Metadata solutions were also nominated by many participants, but wikis and enterprise architecture tools are clearly behind.

These empirical findings show that participants' visions for a data catalog are well aligned with the rationale behind the FAIR principles and denote a willingness to position data catalogs as a vehicle for data democratization. The high nomination counts for a single data catalog environment, an enterprise-wide scope, a broad audience as well as objectives of transparency, accessibility and increased data usage all constitute testimonies to that effect.

In contrast, the documentation approach design areas are a key aspect where no definitive consensus appears from the empirical data (i.e., both were nominated by more than 64% of participants). Incidentally, we have found that implementation approaches as well as related common benefits and challenges mainly revolve around these dimensions. What participants have expressed is their vision, i.e., various possibilities or desired characteristics of their to-be situation. However, as implementation is a process, all of these target characteristics may not be introduced all at once.

Table 24. Taxonomy of data catalog initiatives (based on 11 companies[13])



Dimensions	Meta-Characteristics and Characteristics								
	Category: Scoping and goals								
Scope	Enterprise-wide 82%	Data domain specific 18%	Bus. unit specific 18%	Region specific 9%					
Goal	Transparency 91%	Accessibility 73%	Increased usage 73%	Compliance & risk 36%					
	Category: User groups								
Audience	Specific: data scientists 36%	Restricted: data experts 45%	Broad: business users 100%						
Roles	Data citizen 73%	Bus. Analyst 73%	Data scientist 73%	Data engineer 36%	Chief data officer 45%	Data steward 82%	Data owner 73%	Data architect 73%	Solution architect 45%
	Category: Functionalities								
Data discovery	Search 100%	Recommendation 45%	Data subscription 27%	Data delivery 27%					
Data inventory	Data registration 82%	Metadata management 82%	Business glossary 91%	Data dictionary 64%	Data lineage 73%	Data access 45%	Upload/link content 64%		
Administration	Configuration 55%	User management 55%							
Data assessment	Data usage 73%	Data quality 64%	Data risk 36%	Data valuation 27%	Data profiling 55%	Benchmarking 27%			
Data governance	Workflows 91%	Roles and responsibilities 82%	Rules and policies 73%	Data authorizations 64%	Handling sensitive data 45%				
Data collaboration	Tagging 91%	Rating 55%	Following / updates 64%	Commenting 73%	User communication 64%	Sharing 82%			
Data visualization	Dashboards / cockpits 73%	Data flow / network visualization 64%							
Data analytics	Documentation / stories 36%	Scripting / pipeline management 0%	Data application repository 27%	Data lake status and health monitoring 9%					
Automation & ML	Automated scanning / ingestion 36%	Automated classification / tagging 45%	Normalization / data similarity 36%						
	Category: Data documentation								
Documentation approach	Top-down 64%	Bottom-up 82%							
Implementation focus	Data supply 64%	Data demand 64%							
Data layers	Conceptual 82%	Logical 73%	Physical 73%						
	Category: Tools								
Number of catalogs	Single 73%	Multiple 27%							
Type of tool	Data catalog 91%	Metadata management 64%	Enterprise architecture 27%	Wiki 36%					
Personalization	Configuration 55%	Customization 73%	In-house development 36%						

While the analysis framework helps outlining the characteristics of data catalogs (concept), there is also a need to understand the why and how they are implemented (goals and approaches). In the next section, we will present our multi-case study.

[13] One entry per company, selection of multiple characteristics per dimension allowed, values converted to percentages and rounded for clarity.

5 Data catalogs implementation approaches

From our analysis, we were able to identify three cases that showcase key differentiating goals and approaches, revolving around the business context and goals as well as their relationship with the FAIR principles. Table 25 presents a summary of these cases.

Table 25. Summary of approaches

Case	Albaco	Strychem	Mom-and-Pop
Scope	Enterprise-wide	Enterprise-wide	Enterprise-wide
Goal	Transparency, Increased usage	Transparency, Increased usage	Transparency
Audience	Broad	Broad	Broad
Documentation approach	Top-down	Top-down Bottom-up	Bottom-up
Implementation focus	Data supply	Data supply and data demand	Data demand
Number of catalogs	Single	Single	Single
FAIR focus	F A	F A I R	I R
Key aspects	Enterprise-wide modeling and roll-out All user roles addressed Multitude of functionalities	Roll-out for prioritized data domains or business units Focus on prioritized functionalities with highest business value Focus on user roles involved in selected projects	Roll-out flexibility in accordance with user needs Functionalities as desired by users
Benefits	Fast user-base growth High level of IT-business alignment	Achievement of "quick wins" Opportunity to build up roles and processes	High level of user acceptance Showcasing of achievements
Challenges	User support (necessity to setup a helpdesk) Difficulty to promote an enterprise-wide tool when functions have varying levels of maturity.	Data supply in time and quality Need to for a prioritization concept to select usage scenarios Identification of key stakeholders	Data supply in time and quality Difficulty to showcase shared success stories across various business units

5.1 Albaco case: business transformation through data democratization

The first case is characterized as top-down (i.e., with an objective to model all enterprise-wide assets from the start) and data-supply oriented (i.e., with an emphasis on documenting data assets, as opposed to focusing on user needs). Albaco is a multinational tobacco manufacturing company. As a response to changing consumer preferences and the growing trend towards health consciousness across all industrialized countries, Albaco has decided to make a fundamental change regarding its business model and product portfolio. Consequently, Albaco has operated strategic shift from a business-to-business to a business-to-consumer model, with an emphasis on data-driven insights. This motivated the creation of the Enterprise Analytics & Data (EAD) function, that comprises two major pillars: Data Foundation and Analytics Delivery. It uses data-driven insights as the “carrot” illustrating data’s business value, and data governance as the “stick” ensuring an adequate data foundation. As such, Albaco’s data catalog is positioned as a strategic platform and is the cornerstone of its strategy to become a data-driven company. It is meant to provide the basis for governing the large quantities of data gathered enterprise-wide that constitute its new data foundation.

In light of this objective, it was decided early on to adopt a solution from Collibra and that the data catalog should be open to a wide audience, covering both data experts and business user. Focus was put on data inventory and data governance, which is reflected in the two main components of Albaco’s data catalog platform:

- A business glossary, providing specific information on data for business users, e.g., definitions of terms, business objects and business attributes
- A data dictionary, providing technical information on data models and systems.

Additionally, as a governance-oriented tool, Albaco also put an emphasis on defining governance roles, which were the basis for establishing workflows, i.e., approval processes necessary to ensure correct evolution and maintenance of information and definitions over time. Taken together, these three aspects characterize the essence of the top-down approach, in that initial efforts are targeted on designing extensive models covering all aspects of enterprise data, before the platform is populated.

To achieve this, two key business units were initially selected – one of them was at a high level of maturity with regards to data management, while the other was at the opposite end of the spectrum. Business experts were involved with the data experts from the EAD team to create a business glossary, i.e., defining requirements of what information should be included, defining data roles and ownership, as well as defining workflows for approval – this last aspect was a key requirement to enable onboarding additional users to fill these documentation structures in the next phase.

In a second phase, the data catalog was rolled-out to these two business functions. Access was given to users beyond initial experts, and focus was put on training activities, in order to enable these new users to start entering definitions. At the same time, preparation started for subsequent business units, integrating workflows and user experience improvements that were derived from the first round of implementation. In that sense, Albaco adopts a top-down, data supply-driven approach, starting with requirements and modeling for the documentation of the entire enterprise data scope, and following up with importing data into the platform. Once data definitions were collected and validated through workflows, access to the platform was further granted to data consumers within the company.

Thanks to its approach, Albaco was able to bring its data catalog to scale and reach a sizeable userbase in a timely manner. The data catalog currently covers 20 000 data assets and is used by over 1 200 employees. In addition, more than 200 business experts are routinely editing and approving business definitions. As a result, it successfully supported a better alignment between IT and business perspectives, and allows non-data experts to navigate data resources, learn about them and eventually use them to support daily tasks.

The main difficulty that Albaco faced was user support. On the one hand, business users who have an active role in the system are not always proficient with underlying data-related concepts (e.g., conceptual models, entities, attributes), which significantly increased the training effort. On the other hand, from a user experience perspective, data consumers who are using the system to find data encountered difficulties in using the tool, resulting in higher help requests, which is especially true since all functions are not equal when it comes to data management maturity. This is a direct consequence from the strong data supply focus that Albaco chose for its implementation. Therefore, it is now shifting its efforts towards user acceptance and empowerment, which is a key component of its push towards data-driven decisions and business.

What sets the Albaco case apart is its ambitious goal for introducing a data catalog, which constitutes a key part of redesigning its business models towards becoming more data-driven. For this transition to be successful, Albaco needs to emphasize data democratization to make sure that all employees consider data a fundamental part of their tasks. In doing so, Albaco is placing an emphasis on the Findable and Accessible aspects of the FAIR principles. The Albaco case also highlights the fact that user involvement is not a given in an enterprise context, and that the Accessible aspect also incorporates a need to incentivize users to use enterprise data. From a data catalog functionality perspective, data discovery and data collaboration functions support Finding and Accessing data, by providing a single and collaborative frontend to business users.

5.2 Strychem case: increasing trust in data

The second case emphasizes a balanced approach that combines data supply and demand perspectives. One of the major characteristics of this approach is the structured, step-by-step adoption of the data catalog, which is based on the premise of prioritizing data initiatives that create value (including enable business growth, drive efficiency, and avoid risks). Therefore, the approach combines top-down elements (documenting enterprise-wide data assets for the sake of data governance) with bottom-up elements (specific usage scenario needs) and it understands a data catalog as a platform to matches data demand and supply. Strychem, a chemical company with over 300 production sites worldwide has recently established a digital vision and set up a related program. The program emphasizes data-to-value to strive for transforming Strychem into a data-driven company. A part of the new Strychem working mode concerning data is the "need-to-share". If there is no legal restriction data ought to be shared with an increasingly amount of people within the company. To this end, a new infrastructure, consisting of a (enterprise) data lake and an enterprise-wide data catalog for structured data, was deployed. A data steward organization with more than 100 members ensures a business-driven implementation. Furthermore, a recently established data monetization group in data-to-value cares about the creation of data-driven services for the external market.

The major goals of the data catalog are to create transparency over enterprise-wide data assets while feeding usage scenarios with necessary and useful data. As a component of the overall infrastructure, the data catalog documents data at the conceptual, logical, and physical layer and incorporates a new standard for metadata management. While the data catalog is neither a data storage nor a data analysis tool, it aims at finding, understanding, and governing data. Thus, data governance, search, collaboration, business glossary, and data lineage are among the key functionalities of this data catalog solution.

Regarding the governance aspect of Strychem's approach, the data catalog adoption was accompanied by the definition of data governance roles, especially the role of a data steward (assesses the data quality continuously and implements improvement measures). These roles are of relevance to define workflows, e.g. to grant access rights or to manage the connection of new data sources. Thus, relevant content is pushed into the data catalog at a large scale to enable these data governance functions.

Concerning the demand-driven site, relevant data sources are supplied, and metadata information extended when needed. The data catalog adoption starts with an enablement project and eventually aims at provided as many self-services as possible. Rules and guidelines for the usage of the data catalog as well as skills need to be set up accordingly. In this sense, Strychem follows a step-by-step approach to roll-out the data catalog, both in terms of (meta-)data documented in the data catalog as well as in granting access rights. For instance, data owners (take care on data quality and control the access) and data stewards are addressed in early stages and the complete roll-out to all data citizens (working with data as part of the everyday work) may follow at later stages.

A big benefit of this approach stems from its balanced manner of implementation: it allows to facilitate sustainable data governance functionalities while generating data-driven business value at the same time. Therefore, data governance structures, like data roles, workflows, and data modelling, can evolve and be tested against real-world usage scenarios. Yet, the approach also ensures that the data which is document in the data catalog is also used and that necessary data is provided to the users. As much as the balance is an advantage, it also comes with some inherent challenges. This approach requires a concept on how to balance between exploration and productive usage as well as a prioritization concept to select usage scenarios and not to frustrate others. After prioritization, identifying the right stakeholders in the organization to successfully work with the data catalog is a major challenge. Success stories are particularly necessary in the early stages to maintain momentum and to support the further adoption of the data catalog. These success stories rely on getting the required data in time and quality and on providing self-service possibilities in the tool.

Strychem characterized its issues with “data mistrust” stemming from three main problematic areas: difficult data finding, doubtful data quality, unclear data meaning. These three challenges are in alignment with its three goals for data catalogs: care (i.e., ensuring data quality), connect (i.e., enabling easy access, transparency and sharing) and use (i.e., find, access, understand and act based on data). While Strychem shared the same user involvement challenges as Albaco, it has also identified data quality and data documentation issues as obstacles to increased data usage. Therefore, Strychem also envisions to tackle the Interoperable and Reusable aspects of the FAIR principles. This case shows that Interoperability in the enterprise context has less to do with using standardized formats, and more to do with achieving high data quality and data integration standards. Furthermore, a key component of data Reusability is the ability of users to understand them, for which data documentation is necessary. In fact, data will only be used (and reused) if these criteria are met are met.

From a data catalog functionality perspective, data assessment and functions can support organizations in maintaining data quality and integration standards, supporting the Interoperable aspect of the FAIR principles. As for the Reusable aspects, data inventory and data governance functions can support organizations in documenting data and establishing a shared understanding among stakeholders.

5.3 Mom-and-pop: optimizing data processes

The third case illustrates a user-driven approach. It is characterized as bottom-up (i.e., with an objective to drive data documentation structures from exploring bulk data in the catalog) and data demand-oriented (i.e., with an emphasis on the requirements of the data catalog's end-users). Mom-and-Pop is a European retail group that is currently engaged in unifying management and partially consolidating data from its individual brands. For this purpose, a comprehensive data strategy was defined that includes several tools as strategy building blocks, including a data catalog.

One of Mom-and-Pop's primary goals is to facilitate data-driven decision-making and value generation. The data catalog should serve as a means of accelerating data-based projects, e.g. marketing campaigns to improve cross-selling and up-selling, initiatives to reduce time-to-market. Thus, Mom-and-Pop's data catalog should support users in their data-driven endeavors with data in time and quality. Key functionalities of the selected data catalog solution, therefore, include advanced search, collaboration, business glossary, and automated services. The data catalog focusses on each user that needs data for decision-making and value creation. While Mom-and-Pop's data catalog basically focusses on all data roles, the main user group are data analysts (who identify new use cases, develop proof-of-concepts, and use data to deliver business value).

Mom-and-Pop's implementation approach is stepwise: all user groups, but especially data analysts are encouraged to articulate their data demand to exploit the potential of data. In accordance with the data demand, the data catalog is rolled out flexibly and step-by-step. An advantage of this approach is the relatively high user acceptance: users actively ask for data and are not forced to use the data catalog. Successful data-driven initiatives help to promote the data catalog and create positive network effects among users and data suppliers. Another benefit is that the data catalog can evolve over time, i.e. (meta-)data models and workflows can be developed based on the actual needs. Compared to a (data) push approach, only required data sources are managed in the data catalog, which helps to keep the solution rather lean.

However, it is a big challenge to provide the right data in a short time and in good quality. Without preparatory work this is quite difficult to achieve. For the sake of promoting the data catalog internally, success stories should be created and distributed, which can be difficult if the use cases are divers and spread over various business units.

By implementing a data catalog, Mom-and-Pop aim at creating a reliable and uniform understanding of their data assets and to foster data governance structures. The data catalog is meant to serve as basis for running data-based projects faster and more efficiently. Ultimately, Mom-and-Pops strive for cost reductions. For instance, by identifying and eliminating redundancies, fast discovery of relevant data, and re-use of formerly prepared and processed data, time-to-market can be fastened. While finding and accessing data are essential prerequisites, the lever for gaining increased efficiency is placed at the principles Interoperable and Reusable. In regard to the Interoperable principle, efficiency enhancement potential is associated with enabling machine readability and time savings for finding and accessing data. On that account, (meta)data must use a formal, accessible, shared, and broadly applicable language for knowledge representation (Wilkinson et al. 2016). As for the Reusable principle, cost reductions are related to using data for various purposes and to reducing repetitive tasks, e.g. regarding data preparation and processing. To this end, (meta)data must be described with a plurality of accurate and relevant attributes, meet domain-relevant community standards, and feature a clear data usage license (Wilkinson et al. 2016).

6 Discussion

As illustrated by the cases, one of the main issues to overcome is bringing users aboard. Data catalogs are provided as an over-arching layer to unite enterprise-wide data assets, in an attempt to increase data usage among all employees of the firm. In order to realize the aim of data democratization, the value of the data catalog must appear clearly to all users, on both, the data supply and the data demand side. This is a typical issue of multi-sided platforms, and implementation approaches must take these considerations into account, and be planned accordingly from the beginning (i.e., variety of purposes, variety of users, new roles, free choice of use). Therefore, network effects play a crucial role to keep the data catalog alive (Katz and

Shapiro 1985). The difficulty resides not only in creating positive same-sided network effects, but as data catalogs are meant to link data supply and demand, they must also target positive cross-sided network effects.

With regards to the FAIR principles, this difficulty highlights a crucial difference between the academic environment, from which the principles stem from, and the enterprise context. In fact, the FAIR principles have been created to support the needs of users (i.e., academic researchers) with a high motivation to find and use data. In the enterprise context however, as shown by the cases of Albaco and Strychem, actually getting users to use data is a significant challenge, even if said data is made FAIR, from a technical standpoint. Therefore, user incentivization is a key component of the Accessible aspect in the enterprise context.

Table 26. The FAIR principles in academia and the enterprise

FAIR element	Research context	Enterprise context
Findable	Repositories with search functions	Abstraction platform for enterprise-wide data
Accessible	Repositories store data resources or provide updated links to them. Repositories support the identification of users if required for use of sensitive data	Enterprise systems that store data are interfaced with the platform, linking the physical, logical and conceptual data layers Approval processes and/or access rights are implemented for sensitive data
Interoperable	Use of standardized formats References to other data	
Reusable	Rich documentation, relevant to intended users	

Additionally, the Interoperable aspect also poses differing challenges in academia and business. While the emphasis is put on using standardized and open formats for research, this would arguably be less problematic in an organization that can control its application landscape. In an enterprise context, the Interoperable aspect is to be linked with data-related metrics, such as data quality and integration. Table 26 depicts a summary of these contextual differences, based on insights presented in the previous sections.

Automation is one way of optimizing data catalog implementations, as well as their maintenance. There is a conflict between the need of having detailed documentation models that are able to capture the complexity of data assets (including unstructured data) and increase the value of data catalogs, and the need of keeping this complexity in check to ensure that maintenance efforts are not disproportionately large, as failure to properly maintain the catalog would decrease data quality and eventually, user involvement. Automation technologies would help solve this issue, especially in large organizations, and could also help reuse existing models more efficiently.

7 Conclusion and outlook

As organizations navigate the digital transformation, they need to treat data as a strategic resource and ensure that it is used to inform business decisions (Belissent et al. 2019; Mohr and Hürtgen 2018). In order to achieve this, their employees must adopt a "data mindset", which requires that they know about enterprise data, i.e., about where and how to find it, and about what it means. These issues have been conceptualized for the academic world with the FAIR principles and enterprises must follow suit.

In this context, our study enriches the ongoing academic discourse around data usage and sharing initiated by the FAIR principles, by specifying them in the context of enterprise data. In this study, we have presented a taxonomy of data catalog initiatives enriched with empirical insights and three cases of data catalog goals and implementations. By doing so, this study is among the first to investigate the novel concept of data catalogs, providing empirical insights and shedding light on the difficulties in involving a broad audience of enterprise users with data, as well as in creating positive network effects. It also positions data catalogs, being the latest evolution of metadata documentation tools, as a suitable platform to support such efforts. Furthermore, it provides a conceptual structure for analyzing data catalogs, which is an emerging topic in Information Systems research. It could also facilitate further research on data management and governance.

From a practical perspective, these findings provide data managers and business decision makers a means to grasp the fundamentals of this emerging type of platform. The empirical insights can also support them in developing and driving sustainable data catalog initiatives. The taxonomy can act as a checklist for managerial choices to be considered, ensuring that essential aspects are discussed and establishing common ground among stakeholders. Additionally, the exemplary cases provide real-world illustrations of the taxonomy's dimensions and characteristics.

Finally, potential future research activities could revolve around designing an implementation method for data catalogs, building on learnings from the cases presented in this study. They could also entail empirical studies on how data catalog adoption takes place (e.g., from behavioral and human-computer interaction perspectives) and investigate tool support (e.g., project management tools).

Acknowledgements

This research was supported by the Competence Center Corporate Data Quality (CC CDQ). The authors would like to thank the experts that contributed to this research.

References

Beinke, J. H., Nguyen, D., and Teuteberg, F. 2018. "Towards a Business Model Taxonomy of Startups in the Finance Sector Using Blockchain," in *Proceedings of the 39th International Conference on Information Systems (ICIS)*, San Francisco, California, USA, December 13. (https://aisel.aisnet.org/icis2018/crypto/Presentations/9).

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. (https://www.forrester.com/report/Determine+Your+Datas+Worth+Data+Plus+Use+Equals+Value/-/E-RES127541).

Borgman, C. L. 2003. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*, MIT Press.

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., Hoberman, E., and Hartig, K. 2018. "The Forrester Wave: Machine Learning Data Caralogs, Q2 2018," Consortium Report, Consortium Report, Forrester Research, June 21. (https://www.forrester.com/report/The+Forrester+Wave+Machine+Learning+Data+Catalogs+Q2+2018/-/E-RES140524#dialog-1574414272946-dialog).

Gualtieri, M., Yuhanna, N., Kisker, H., Curran, R., Purcell, B., Christakis, S., Warrier, S., and Izzi, M. 2016. "The Forrester Wave: Big Data Hadoop Distributions, Q1 2016," Forrester Research, January 19. (https://www.forrester.com/report/The%20Forrester%20Wave%20Big%20Data%20Hadoop%20Distributions%20Q1%202016/-/E-RES121574).

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 Conference on Innovative Data Systems Research 2017*, Chaminade, California, USA, p. 12.

Kaplan, B., and Duchon, D. 1988. "Combining Qualitative and Quantitative Methods in Information Systems Research: A Case Study," *MIS Quarterly* (12:4), pp. 571–586. (https://doi.org/10.2307/249133).

Katz, M. L., and Shapiro, C. 1985. "Network Externalities, Competition, and Compatibility," *The American Economic Review* (75:3), pp. 424–440.

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.

Labadie, C., Eurich, M., and Legner, C. 2020. "Data Democratization in Practice: Fostering Data Usage with Data Catalogs," in *Proceedings of the 20th Symposium of the Association Information et Management*, Marrakesh, Morocco.

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.

Mohr, N., and Hürtgen, H. 2018. "Achieving Business Impact with Data," Consultancy Report, Consultancy Report, Digital McKinsey, April. (https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/achieving-business-impact-with-data).

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. 2011. "Data Governance," *Business & Information Systems Engineering* (3:4), pp. 241–244. (https://doi.org/10.1007/s12599-011-0162-8).

Peyret, H., Cullen, A., McKinnon, C., Blissent, J., Iannopollo, E., Kramer, A., and Lynch, D. 2017. "Enhance Your Data Governance to Meet New Privacy Mandates," Consortium Report, Consortium Report, Forrester Research.

Püschel, L., Roeglinger, M., and Schlott, H. 2016. "What's in a Smart Thing? Development of a Multi-Layer Taxonomy," in *Proceedings of the 37th International Conference on Information Systems (ICIS)*, Dublin, Ireland, December 11. (https://aisel.aisnet.org/icis2016/DigitalInnovation/Presentations/6).

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).

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).

Venkatesh, V., Brown, S., and Bala, H. 2013. "Bridging the Qualitative–Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems," *Management Information Systems Quarterly* (37:1), pp. 21–54.

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), p. 160018. (https://doi.org/10.1038/sdata.2016.18).

Wixom, B., and Ross, J. 2017. "How to Monetize Your Data," *MIT Sloan Management Review* (58:3), p. 7.

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).

Building Data Management Capabilities to Address Data Protection Regulations: Learnings from EU-GDPR

Clément Labadie and Christine Legner

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

Published in the Proceedings of the 14th International Conference on Wirtschaftsinformatik (WI), 2019

Extended journal manuscript

Abstract. The European General Data Protection Regulation (EU-GDPR) has entered into force in May 2018. Its emphasis on individual control and organizational accountability constitutes a new paradigm has led to significant changes in the way organizations manage personal data. However, two years later, organizations are still facing difficulties when implementing EU-GDPR, such as establishing transparency on personal data collection, storage and retention, and enforcing granular user consent preferences when processing data. Anchored in the resource-based view theory (RBV), this paper argues that the regulation requires companies to build a dedicated data management capability. It presents a capability model that was developed in an iterative design science process, integrating both interpretation of legal texts, the continuous review of EU-GDPR-related publications and practical insights from focus groups with experts from 22 companies and 3 EU-GDPR projects. The paper makes two academic contributions: first, it advances the regulatory compliance management literature by translating and concretizing legal data protection concepts into a set of system and organizational capabilities. Second, the suggested capability model contributes to building common ground between legal and data management domains. Practitioners may use the capability model to assess their current status and set-up systematic approaches towards compliance with an increasing number of data protection regulations.

Keywords: EU-GDPR, Data Protection, Regulations, Compliance, Capabilities

Table of contents

List of figures

List of tables

1 Introduction

In 2020, the European Commission (EC) released a plan for a European data strategy, outlining the transformative importance of data in modern economies, and striving to position the European Union (EU) at the forefront of data-related innovation (European Commission 2020). One of the pillars of this strategy is the building of public trust in data processing activities, with the General Data Protection Regulation (EU-GDPR) enabling individuals to have control of their personal data. EC President Ursula von der Leyen even ambitions for EU-GDPR to set data protection "*standards for the rest of the world*" (von der Leyen 2020). In fact, since its enforcement in May 2018, EU-GDPR has initiated a paradigm shift in data protection, towards greater choice and sovereignty for individuals, and more accountability for organizations (De Hert and Papakonstantinou 2012). For organizations, it comes with the burden of proof related to whether, how and how well they protect personal data and increased fines for noncompliance. This requires them to fundamentally rethink the way they store and process personal data on an enterprise-wide level. As EU-GDPR applies at a scale larger than any previous data protection regulation, it has revealed implementation challenges for organizations, as well as conflicting interests between legal obligations, business drivers and innovation (Jakobi et al. 2020). Yet, even two years after the date of EU-GDPR's entry into force, many organizations are still facing compliance challenges. A study conducted in June 2019 among more than 1000 European and US companies, reported that organization had been over-optimistic regarding their ability to achieve timely compliance. When surveyed in March-April 2018, 78% of responding organizations expected to be compliant with EU-GDPR by June 2018, but only 28% of them reported to be compliant upon reassessment a year later (Capgemini Research Institute 2019). A study released in April 2020 reveals similiar difficulties among large multi-national enterprises: only 54% of them had achieved operational compliance, while 37% were still conducting "significant readiness actions", while 9% were still at a "project mode" (Dansac Le Clerc and Mannent 2020). According to this study, a majority of organizations are still implementing mechanisms to manage data protection rights, data storage and retention, and in-depth registries of data processing activities (Dansac Le Clerc and Mannent 2020).

From a research perspective EU-GDPR has been debated in both legal and IS communities (De Hert and Papakonstantinou 2012, 2016; Jakobi et al. 2020; Mitrou 2017). Although legal aspects of information privacy were not standing among the "topic areas closer to the interests of most IS researchers" (Bélanger and Crossler 2011) until then, IS researchers have started to investigate approaches to ease EU-GDPR compliance management, similar to the idea of regulatory technologies for financial regulations, so-called "RegTech" (Butler and O'Brien 2019). However, they have mostly proposed technical solutions, such as enterprise architecture (Burmeister et al. 2019, 2020; Huth, Burmeister, et al. 2020) and blockchain (Farshid et al. 2019; Guggenmos et al. 2020; Mejtoft et al. 2019; Rieger et al. 2019).

The difficulties in implementing EU-GDPR highlight not only the general lack of common ground between legal and IS research communities, but also between the professionals in both disciplines. In most companies, data protection topics have traditionally been addressed by legal departments by adapting contracts and general conditions, but without directly influencing data management practices. However, the new generation of data protection regulations does not allow for such a restricted approach, and companies see data processing related issues as the most challenging topics in EU-GDPR. Since the regulation remains generic and does not prescribe concrete implementation options, there is a need to translate it into data management concepts and practices.

Anchored in the resource-based view theory (RBV), this paper argues for utilizing capabilities as an interface between abstract compliance requirements and their concretization. Capabilities also support the “translation” between legal and IS perspectives and help to analyze compliance requirements and options, before deciding on concrete (technical) implementations. More specifically, this paper addresses the following research question: *what data management capabilities need to be built in order to address EU-GDPR's requirements?* Following a design science research approach (Peffers et al. 2007), we propose a capability model for EU-GDPR that integrates both interpretation of legal texts, insights from EU-GDPR related publications and practical insights from focus groups with experts from 22 companies as well as 3 EU-GDPR projects. The resulting capability model identifies and describes the required organizational and system capabilities from a data management perspective. The paper makes two academic contribution: first, it advances the regulatory compliance management literature by translating legal data protection concepts for the IS community. Second, the suggested capability model contributes to building common ground between legal and data management domains. In contrast to prior research on EU-GDPR that treats selected aspects of the regulation or proposes specific implementation solutions, our study thereby provides an integrated perspective on enterprise-wide data management practices. The capability model may also inform researchers that investigate specific aspects of the regulation and act as an overarching framework that outlines the links between capabilities and their materialization in recent IS studies. Practitioners may use the capability model to assess their current status and set-up systematic approaches towards compliance with an increasing number of data protection regulations.

The remainder of this paper is structured as follows: we first introduce the EU-GDPR and provide a synthesis of research on the topic and on regulatory compliance in general. After outlining the research methodology and process, we motivate the resource-based view (RBV) perspective, and present the capability model. We conclude by summarizing our contribution and discussing future research.

2 Background and related research

2.1 The European General Data Protection Regulation (EU-GDPR)

In January 2012, the European Commission published a proposal for an overhaul of data protection law, which would become EU-GDPR[14]. It thereby aimed to remedy the fragmented implementations of the preceding Data Protection Directive (95/56/EC), as well as to account for the significant changes introduced by the internet and digital services (Mitrou 2017; Nicolaidou and Georgiades 2017). As a result, EU-GDPR directly applies in every EU member state. Moreover, any organization that processes personal data of EU-citizen must comply with it, regardless of the geographical location of their operations. If it fails to do so, significant fines apply (i.e., up to 20 million euros or 4% of an organization's global revenues, whereas previous regulations averaged at ca. 500 000 euros). While EU-GDPR reinforces existing concepts, it also introduces new ones. Most notably, existing transparency mandates have been strengthened – organizations must now inform individuals about data processing in clear language and separately from general conditions, and are also required to present more granular consent options (Nicolaidou and Georgiades 2017). One of the major additions is the concept of accountability, which implies that organizations must be able to demonstrate compliance with the regulation. They must also appoint data protection officers (DPOs) and announce data breaches to both authorities and individuals (data breach notification). Privacy-by-design principles (i.e., implementing privacy from the ground up in systems and offerings) also appear in the regulation, along with new individual rights, such as data portability as well as a right to oppose automated decision making (Nicolaidou and Georgiades 2017). All of these evolutions constitute a paradigm shift in data protection, towards greater choice and sovereignty for individuals, and more accountability for organizations (De Hert and Papakonstantinou 2012, 2016; Mitrou 2017).

[14] Regulation (EU) 2016/679. Recitals (R.) and articles (art.) mentioned throughout the text refer to EU-GDPR unless otherwise specified.

2.2 EU-GDPR and data protection in IS literature

Although EU-GDPR was finalized in 2016 and presents a major paradigm shift in data protection, it only slowly gained the attention of IS researchers. This reflects the general reluctance of IS researchers regarding information privacy (Bélanger and Crossler 2011). In 2018, a query with the keyword "GDPR" on the AIS Electronic Library only returned 27 matches. This number has been multiplied by 10 within the last two years – the same query returns 262 matches as of November 2020. These studies can be categorized according to their type of contributions:

1. *Core contribution on the overall regulation* (19 studies, s. Table 27): in these studies, EU-GDPR is the central topic of interest and the outcomes are derived following an analysis of the regulation as a whole. This is where we position the study at hand.
2. *Core contribution on selected aspects of the regulation* (23 studies, s. Table 27): in these studies, EU-GDPR is the central topic of interest and the outcome relates to a specific aspect of the regulation, e.g. consent management or data deletion.
3. *Core contribution on neighbouring topics* (110 studies): these studies contribute to domains that are related to EU-GDPR and data protection, such as other regulations, general privacy research, cybersecurity, information ethics, information disclosure and data sharing. While they position EU-GDPR as motivating factor for their outcomes, they do not analyze the regulation.
4. *Core contribution on other topics* (94 studies): these studies relate to a variety of other domains of IS research, and only mention EU-GDPR to back up a specific, isolated argument, with no strong link to the main contribution.

Table 27 lists the contributions in the first and second category which we analyzed in more detail and categorized according to Bélanger and Crossler's (2011) taxonomy of topic areas. The EU-GDPR studies published until 2018 fell within the domains of information privacy practices and information privacy technologies and tools. Since then, as the regulation entered into force and received significant attention in organizations and the general public alike, some researchers have investigated the information privacy concerns and attitudes of individuals and specific stakeholder groups (e.g. software developers, researchers, business executives) with regards to the regulation. We observe the most significant uptake in studies that focus on technologies and tools for EU-GDPR compliance, which now constitute the majority of EU-GDPR-related studies (i.e. 41% in 2020, up from 28% in 2018). Four of these studies (out of 16) investigate blockchain as a technological basis for compliance solutions. On the opposite side, studies regarding information privacy practices were predominant in 2018 (i.e. 57%), but are now second to technology and tools-related research (i.e. down to 31% in 2020) – they predominantly consist of empirical studies of EU-GDPR-related practices. Lastly, the share of studies classified in the information privacy impact category has slightly dropped (i.e. 28% in 2018, down to 20% in 2020) – they investigate the impact of EU-GDPR on emerging technologies (e.g. advanced analytics and smart products) and business model innovation, as well as the economic/market impact of data portability.

While the majority of relevant studies has a narrow scope on one of EU-GDPR's requirements, those that consider the overall regulation tend design dedicated solutions to tackle compliance. They investigate, for instance blockchain-based personal data management solutions, evaluate existing practices, or analyze EU-GDPR's impacts on a specific domain (e.g. social media discourse, innovation, Big Data). There are two shortcomings in these approaches: first, most papers take the compliance requirements for granted and directly look into specific practices or solutions. Second, these studies do not provide insights into the entire regulation's implications on data-related practices from an enterprise-wide perspective. Hence, we are still lacking a broader and solution-agnostic understanding of data-related practices and the required changes with EU-GDPR. Russell et al. 2018 address this topic by proposing a Digital-Privacy Transformation "Gap-Map" that measures the organization's propensity for change. However, it exclusively takes a change management perspective, without investigating the compliance requirements and their implications on enterprise-wide data management practices.

Table 27. Overview of EU-GDPR related studies in IS literature

Study	**Type***	**Topic area†**	**Research focus**
Scope: overall regulation			
(Addis and Kutar 2018)	C	Impact	Impacts of EU-GDPR on emerging technologies
(Martin and Matt 2018)	E	Impact	Impacts of privacy regulations on startup innovation
(Pankowska 2018)	C+E	Practices	EU-GDPR mapping to privacy frameworks and awareness
(Russell et al. 2018)	C	Impact + practices	Transformation framework for digital privacy
(Tona et al. 2018)	C	Tech./Tools	Design of ethical Big Data artefacts
(Veiga et al. 2018)	E	Practice	Mapping of data protection regulations and benchmarking of practices
(Burmeister et al. 2019)	C	Tech./Tools	Enterprise Architecture meta-model for EU-GDPR
(Martinez et al. 2019)	C	Impact	Impacts of EU-GDPR on smart grid operations
(Rösch et al. 2019)	C	Tech./Tools	Translation of legal requirements into technical requirements
(Addis and Kutar 2020)	E	Tech./Tools	Data protection challenges arising from implementation of AI
(Burmeister et al. 2020)	C	Tech./Tools	Enterprise Architecture Management supporting EU-GDPR implementation
(Francis et al. 2020)	C	Practices	Comparison of principles behind privacy frameworks in 14 countries
(Grundstrom et al. 2020)	E	Practices	EU-GDPR impacts on access to data inside organizations
(Houta et al. 2020)	E	Concerns	Analysis of EU-GDPR discourse on social media

Study	Type*	Topic area†	Research focus
(Huth, Burmeister, et al. 2020)	E	Practices	Collaboration between legal and enterprise architecture teams during EU-GDPR implementations
(Jakobi et al. 2020)	C	Impact	Research contribution to conflicting business implications of EU-GDPR implementation
(Lindgren 2020)	E	Impact	Impact of EU-GDPR on (multi) business model innovation
(Maunula 2020)	C	Tech./Tools	Technoloy review for EU-GDPR
(Zhang et al. 2020)	E	Concerns	Impacts of EU-GDPR on consumer online trust
Scope: data protection rights			
(Engels 2016)	C	Impact	Impacts of data portability right on competition dynamics
(Farshid et al. 2019)	C	Tech./Tools	Blockchain prototype for data deletion
(Presthus and Sørum 2019)	E	Concerns	Privacy awareness and knowledge of consumers following EU-GDPR
(Rieger et al. 2019; Guggenmos et al. 2020)	E	Tech/Tools	Design principles and development of blockchain solution for asylum procedures in Germany
(Wohlfarth 2019)	C	Impact	Strategic aspects of data portability
Scope: Consent			
(Bergram et al. 2020)	E	Attitudes	Influence of phrasing and digital nudges on user consent and privacy awareness
(Kurtz et al. 2020)	E	Practices	Identification of consent related issues + design goals
(Proferes and Walker 2020)	E	Attitudes	Researchers attitudes towards consent in exploiting public data
Scope: transparency requirements			
(Alboaie 2017)	C	Tech./Tools	Privacy label for GDPR
(Diamantopoulou and Mouratidis 2018)	C	Tech./Tools	Reference architecture for privacy level agreements
(Fox et al. 2018)	C	Tech./Tools	Guidelines for compliant privacy notices
(Mejtoft et al. 2019)	E	Tech./Tools	Blockchain prototype for increased transparency of data processing
(Watson and Nations 2019)	E	Tech./Tools	Identification of factors influencing transparency of algorithms + recommendations
(Paul et al. 2020)	E	Impact	Impact of EU-GDPR on user privacy perceptions for wearable IoT devices
Scope: accountability requirements			
(Karyda and Mitrou 2016)	C	Practices	Information security / incident management
(Petkov and Helfert 2017)	E	Practices	Applying data breach notification to past infringements
(Kurtz et al. 2018)	E	Practices	Review of third-party data processors
(Vemou and Karyda 2018)	C	Practices	Evaluation of privacy impact assessment methods
(Kurtz et al. 2019)	E	Practices	Analysis of third-party data processing in service ecosystems
Scope: technical and organizational measures			
(Huth and Matthes 2019)	C	Tech./Tools	Privacy engineering approaches for software development
(Faber et al. 2020)	C	Tech./Tools	Blockchain-based personal data and identity management system
(Huth, Both, et al. 2020)	C	Tech./Tools	Tool prototype + approach for integrating privacy aspects in agile development methods
** C = conceptual, E = empirical*			
† Based on Bélanger and Crossler (2011)			

2.3 Regulatory Compliance Management (RCM)

So far, the academic discussion on EU-GDPR has not linked up to the regulatory compliance management (RCM) research domain, although the latter could inform how to analyze regulations and their influence on business practice. RCM is defined as "ensuring that enterprises are structured and behave in accordance with the regulations that apply, i.e., with the guidelines specified in the regulations" (El Kharbili 2012). RCM introduces useful definitions to delimit relevant legal concepts and distinguishes between regulations (i.e., binding document), regulatory guidelines and compliance requirements, as provided in the legal text. Following interpretation, this ultimately results in concretized compliance requirements as implementation (El Kharbili 2012).

Two review papers from 2009 analyze the coverage of RCM in IS research. Cleven and Winter (2009) isolate 26 relevant papers and analyze them through the lens of enterprise architecture. They found that while some RCM aspects have been prominently studied (e.g. organizational and behavioral impacts of regulations, compliance supporting IT solutions), others had been neglected. Specifically, they found no contributions on the operationalization of compliance objectives. Abdullah et al.'s (2009) review on RCM, revolves around the approaches (i.e., explanatory or solution) and context (i.e., region, type and domain) of the considered contributions. The the majority of the 45 papers concerns North America, whereas only 3 of them focus on Europe. Related to data protection, they identify 2 papers on Fair Information Practices, and only one on the European Data Protection Directive (95/46 EC), even though it had been enforced for 15 years. Furthermore, all identified contributions offer either preventive or detective solutions, but no corrective solutions. The authors hypothesize that corrective solutions are an outcome of legal analysis, which is why they were not addressed by the IS community.

Hence, there is a lack of RCM-related contributions that address data protection regulations, focus on regions other that North America (Abdullah et al. 2009) and provide guidance to concretize strategic compliance objectives (Cleven and Winter 2009). This last call is echoed by our literature review on EU-GDPR – although contributions exist around the topic, they all focus on specific aspects of the regulation, and lack a single integrating framework.

3 Research design

Given our goal to support companies in achieving EU-GDPR compliance, we adopt design science research (DSR) to develop a capability model, as an artefact "to solve identified organizational problems" (Hevner et al. 2004). Figure 11 depicts the research steps, following the iterative process suggested by Peffers et al. (2007). It outlines how different types of research activities informed the development of the capability model: analysis of legal texts, review of EU-GDPR related publications as well as close interactions between academics and practitioners, comprising 5 focus group meetings with 33 data management experts from 22 companies and insights from 3 EU-GDPR projects.

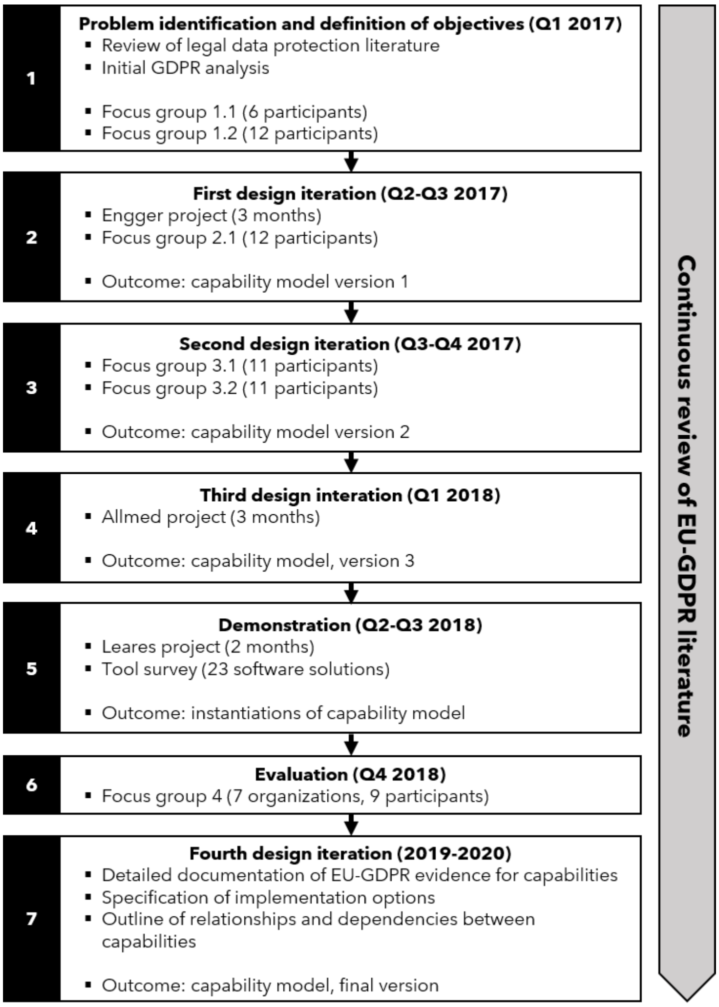

1
Problem identification and definition of objectives (Q1 2017)
▪ Review of legal data protection literature
▪ Initial GDPR analysis
▪ Focus group 1.1 (6 participants)
▪ Focus group 1.2 (12 participants)
2
First design iteration (Q2-Q3 2017)
▪ Engger project (3 months)
▪ Focus group 2.1 (12 participants)
▪ Outcome: capability model version 1
3
Second design iteration (Q3-Q4 2017)
▪ Focus group 3.1 (11 participants)
▪ Focus group 3.2 (11 participants)
▪ Outcome: capability model version 2
4
Third design interation (Q1 2018)
▪ Allmed project (3 months)
▪ Outcome: capability model, version 3
5
Demonstration (Q2-Q3 2018)
▪ Leares project (2 months)
▪ Tool survey (23 software solutions)
▪ Outcome: instantiations of capability model
6
Evaluation (Q4 2018)
▪ Focus group 4 (7 organizations, 9 participants)
7
Fourth design iteration (2019-2020)
▪ Detailed documentation of EU-GDPR evidence for capabilities
▪ Specification of implementation options
▪ Outline of relationships and dependencies between capabilities
▪ Outcome: capability model, final version
Continuous review of EU-GDPR literature

Figure 11. Research process based on (Peffers et al. 2007)

The **first phase** was meant to understand the challenges with EU-GDPR implementation and specify the research objectives. In an initial review of the regulation, we extracted EU-GDPR's compliance requirements and analyzed them according to foundational data protection principles in legal literature (i.e. personal data, informational self-determination, accountability and transparency). To that end, we selected reference text books that integrate legal texts and their related data protection foundations and preparatory works, as well as insights from case law and legal doctrine (Bensoussan et al. 2018; Debet et al. 2015; Voigt and Von Dem Bussche 2017). Early results of this analysis were discussed with practitioners through focus groups 1.1 and

1.2, allowing them to reflect on the regulation's impacts on their organizations and implementation challenges. These discussions revealed two main challenges with regards to EU-GDPR compliance. First, participants recognized a lack of understanding of the regulation itself, while anticipating significant changes to the current way of storing and processing personal data on an enterprise-wide level. Second, they cited a lack of common ground with legal departments. In their organizations, discussions around data protection and privacy regulations are often cut short due to a lack of common approaches and vocabularies, which blocks the identification of feasible and compliant solutions and hinders progress. This led to the research objective of defining a capability model for EU-GDPR that assists data management professionals to understand and implement the regulation, as well as collaborate with legal colleagues.

The next phases (2, 3 and 4) were iterative design cycles, involving insights from field projects and parallel research activities to design the capability model, as well as focus groups for collecting feedback. Research activities included a continuous analysis of EU-GDPR-specific legal literature (Bensoussan et al. 2018; De Hert and Papakonstantinou 2012, 2016; Guadamuz 2017; Mitrou 2017; Nicolaidou and Georgiades 2017; Voigt and Von Dem Bussche 2017), guidelines from official authorities (European Data Protection Board 2017, 2018a, 2018b) as well as interpretations from the private sector, including consortia (e.g. Iannopollo et al. 2017, 2018; Merlivat et al. 2017; Peyret et al. 2017) and industry stakeholders (e.g. Deutsche Telekom 2016).

Phase 2, the first design iteration phase, comprised a project at Engger[15], a global engineering company, and resulted in the initial version of the capability model. It had just started a multi-project around EU-GDPR-compliant personal data aiming at harmonizing business partner data management in a highly distributed landscape, i.e., with around 500 systems in different countries and subsidiaries. This project helped understanding issues and define capabilities related to collection and distribution of personal data and consent. It ultimately led to the first version of the capability model that was presented to and discussed with data management experts in focus group 2.1.

During **phase 3**, the discussions in the two focus group meetings 3.1 and 3.2 revolved around the scope of the model. Feedback from focus group 3.1 indicated that security is usually a distinct function, and supported the need for a data management-centric perspective. From an academic perspective, information security is a well-research field and the existing concepts may be translated to EU-GDPR, whereas there is little coverage of data management practices in regulatory compliance with data protection regulations. It was decided to set aside all security-

[15] All company names have been anonymized.

related considerations from the capability model and focus exclusively on data management capabilities.

Phase 4 comprised a project around consent management at Allmed, a global pharmaceutical company. Its technical team had designed an MVP solution, which we analyzed based on the second version of the capability model. Insights from the project resulted in the capability model's third version with a stable set of capabilities and sub-capabilities.

Phase 5 included a demonstration with the EU-GDPR activities at Leares, a small consulting firm. The capability model proofed to be applicable and useful for assessing the current capabilities, identifying the required capabilities and prioritizing compliance activities. In parallel, we analyzed software tools from major vendors claiming to support EU-GDPR compliance (s. Appendix 1). To that end, we used the capability model to analyze and classify 23 tools from major vendors that fall into the common categories of data management, compliance and identity & access management (CIAM), security, and enhancement. This analysis allowed us to further validate and demonstrate the capabilities.

In **phase 6**, we conducted additional expert interviews to evaluate the artefact's simplicity, understandability, fidelity and completeness (evaluation criteria as suggested by (Prat et al. 2015). We conducted interviews with the data protection officer as well as a data management specialists of two major insurance companies in Switzerland, Versuisse and Svizzance. Both companies are among the top 10 providers of life and non-life insurance, and are also operating in EU countries. Interviews consisted in a walkthrough of each individual capability, in order to discuss and evaluate the company's standing and practices. At the end of each interview, we asked participants to rate the capability model's simplicity, understandability and completeness using a 5-level likert scale (where 1 = fully disagree, 3 = neutral and 5 = fully agree). Our respondents rated the capability model's simplicity, understandability and completeness were all rated with a minimum of 4 out of 5. The fidelity dimensions was the only one with no rating of 5, as respondents rated it with 3 and 4. Respondents with a legal education indicated that although the capabilities seemed the adequately reflect EU-GDPR requirements, they were missing assignments of each capability to the regulation's principles. Similarly, data management expressed that although capabilities matched the requirements that they discussed with members of their organizations' legal teams, they pointed a lack of explicit reference to the regulation.

Finally, for **phase 7**, we updated and refined the capability model in light of expert feedback, and mapped each capability, sub-capability and software features with relevant EU-GDPR recitals and articles. We also monitored EU-GDPR-related studies until Q4 2020, refreshing the literature review, and integrating insights from selected publications as support for relevant capabilities. The combination of practitioner and research insights also enabled us to specifiy relationships and dependencies between capabilities, as well as isolate enabling ones.

4 Data management capabilities for EU-GDPR

4.1 Capabilities for regulatory compliance

As theoretical foundation, we chose to rely on the RBV, as regulatory compliance is a component of firm performance, and contributes to an organization's control objectives (as defined by Sadiq et al. 2007). Building on Zhang et al.'s (2013) definition of an IT capability, we define data management capabilities for regulatory compliance as a firm's ability to acquire, deploy, and leverage its data resources in combination with other resources and capabilities in order to achieve an organization's compliance objectives.

Table 28. Positioning capabilities among RCM concepts

RCM concept	Definition (based on (El Kharbili 2012))	Illustration in EU-GDPR
Regulatory guideline	Stipulates a set of obligation to comply to.	Art. 6 – "Lawfulness of processing": enumerates conditions in which data processing is legal.
Compliance requirement (CR)	Pieces of text extracted from the regulatory guideline specifying an expected behavior / a specific condition to fulfill.	Extraction of requirements bearing data management relevance. E.g. art. 6 § 1 a and art. 7 § 1 require that data be processed according to individuals expressed consent.
Capabilities	*Result of the interpretation of CRs in terms of capabilities that are to be implemented or improved.*	*Manage consent and sub-capabilities: implement consent items, collect consent instances, distribute consent, enforce consent-based processing.*
Concretized compliance requirement (CCR)	Implementation of a CR in an enterprise model, fulfilling its legal specification.	A concrete measure implemented in a specific organization to operationalize CRs. E.g. "In company X, consent data should be first recorded in system 1 and pushed to other systems every 12 hours".

The capability model complements RCM concepts (El Kharbili 2012) and acts as an abstraction layer between the normative aspects of the regulation, i.e. the regulatory guidelines and compliance requirements (CR), and the concretized compliance requirements (CCR), i.e. the concrete implementation of a CR. Introducing capabilities allows describing results from the interpretation of CRs and translate them into what organizations should do, as opposed to how they should do it. Table 28 depicts this articulation.

4.2 Capability model: structure and overview

EU-GDPR art. 24 § 1 states the overall responsibility of organizations with regards to the regulation as the implementation of "*appropriate technical and organizational measures to ensure and be able to demonstrate that processing is performed in accordance with this Regulation*". In line with capability conceptualizations[16], we derived our two main capability groups, i.e., system and organizational capabilities (see Table 29), reflecting their predominant aspect. Correspondingly, **system capabilities** are mainly enabled by data-processing systems, while **organizational capabilities** predominantly rely on data protection processes and responsibilities (but can be supported by tools). Capabilities were derived from EU-GDPR's underlying principles, as described by legal literature, and reflect the "pillars" of the regulation. Sub-capabilities are the result of the analysis and express compliance requirements. In the following sections, we present each of the suggested capabilities, along with its justification, the empirical evidence and the sub-capabilities.

Table 29. Capability model for EU-GDPR

(*A*) System capabilities				
(*A1*) Define protected data scope	(*A1.1*) Identify data objects	(*A1.2*) Classify data attributes	(*A1.3*) Locate data records	
(*A2*) Manage consent	(*A2.1*) Implement consent items	(*A2.2*) Collect consent instances	(*A2.3*) Distribute consent	(*A2.4*) Enforce consent-based processing
(*A3*) Enable data processing rights	(*A3.1*) Delete data	(*A3.2*) Pseudonymize data	(*A3.3*) Transmit data in standardized form	
(*B*) Organizational capabilities				
(*B1*) Orchestrate data protection activities	(*B1.1*) Assume data protection responsibilities	(*B1.2*) Oversee data protection activities	(*B1.3*) Control compliance of external processors	
(*B2*) Demonstrate compliant data processing	(*B2.1*) Maintain records of processing activities	(*B2.2*) Maintain documentation of system landscape	(*B2.3*) Supervise sensitive processing activities	
(*B3*) Disclose information	(*B3.1*) To individuals	(*B3.2*) To authorities		

[16] In the RBV, capabilities "involve complex patterns of coordination between people and between people and other resources" (Grant 1991). Authors relying on the RBV in the IS literature usually demarcate technological and organizational aspects that underpin IS capabilities (Baiyere and Salmela 2014; Bharadwaj 2000).

4.3 System capabilities

4.3.1 Define protected data scope

This capability is based on art. 1 § 1 and 4 § 1 and denotes the ability to clearly identify, classify and locate personal data. Personal data is defined as "data enabling direct or indirect identification of a single physical person, data that is specific to a single physical person without enabling identification, data that can be linked to a physical person, data regarding which anonymization techniques cannot completely mitigate the risk of re-identification" (Debet et al. 2015).

Focus groups 1.1 and 1.2 indicated that companies generally had no overview on the personal data collected and used during processes, especially in terms of storage location. A participant of focus group 3.2 asked: "How do you identify personal data in a heterogeneous IT-System landscape?" Follow-up questions revolved around means to identify personal data. The project at Engger provided significant insight regarding this capability group. One of its main objectives was making sure that personal data was consistently kept up-to-date within all systems, which proved difficult due to multiple overlapping systems managed in independent subsidiaries. Overall, companies faced two main challenges: determining what kind of personal data they were processing, and where such data was stored. The resulting capability may be best summarized by Bensoussan et al. (2018), stating that "organizations must have perfect knowledge of personal data". Iannopollo et al. 2017 recommend two actions that mirror these issues (e.g. data inventory and system mapping) and suggest that personal data should not only be identified, but also classified. This is required as EU-GDPR prescribes higher protection levels for data that is considered sensitive (R. 51). From a tool perspective, this capability group is well aligned with the functional scopes of solutions in the data management and security/protection tool categories, as most of them provide functionalities supporting data discovery, i.e., retrieving data across the organization's entire system landscape, and classification (e.g., using data crawlers). The resulting sub-capabilities are:

- Identify data objects: identify data domains and related data objects that fall within EU-GDPR's scope of applicability, for instance **customer, employee, or job applicant.**
- Classify data attributes: assign levels of sensitivity to data attributes contained within personal data objects.
- Locate data records: identify all storage instances of personal data objects and have the ability to access and retrieve them.

Practitioner insights indicate that this capability group is intertwined with documentation and classification activities. Svizzance and Versuisse, for instance, were in the process of re-aligning existing data and system landscape documentation with their databases and systems, by specifying additional details and updating it when necessary. For this purposes, both organizations turned to enterprise architecture tools – this approach has been investigated by recent research, which confirms its adequacy in this context (Burmeister et al. 2019, 2020; Huth, Burmeister, et al. 2020). In the case of Leares, the company did not possess a similar existing documentation, and it became clear that this capability group should be the focus of initial efforts, as other capabilities could not be realized without an understanding of the protected data scope.

This capability can be viewed as a prerequisite to other system capabilities, which consist in performing operations on previously identified data records – for instance, it is not possible to delete a data object if it has not been discovered and indexed.

Table 30. Capability overview: Define protected data scope

Sub-Capability	Description	Specification	Implementation	Evidence
Identify data objects *(A1.1)*	Ability to inventory data objects within EU-GDPR scope contained in enterprise systems	Identify relevant data domains that contain personal data (e.g. customer, business partner, employee, job applicant) Distinguish between identifying data (i.e. that can be linked to a specific individual) and non-identifying data (i.e. that cannot be linked to a specific individual)	Data crawlers with machine learning for personal data retrieval and identification Identification based on pre-existing documentation: enterprise architecture models, data models, application inventory or data catalog Manual scanning	Art. 2 § 1 Art. 4 § 1 Art. 15 R. 26, 27, 30, 57
Classify data attributes *(A1.2)*	Ability to assign levels of data sensitivity to identifying data attributes	Personal data (e.g. identity, contact, personal history, finan-cial, location, content) Sensitive data (e.g. opinions, health, biometry, ethnic origin, criminal history) Data related to children	Data crawlers with machine learning for personal data retrieval and classification Classification based on pre-existing documentation: enterprise architecture models, data models, application inventory or data catalog Manual scanning	Art. 4 Art. 8-10 R. 34, 35, 38, 51
Locate data records *(A1.3)*	Ability to inventory all systems containing personal data and retrieve relevant objects	Internal location (systems, storage media, responsible or-ganizational unit) External location (third party processors, geographical location)	Automated queries System landscape documentation Data flows documentation Data lineage	Art. 4 Art. 15 Art. 28 R. 101, 103, 107, 108, 110

4.3.2 Manage consent

This capability comprises the prerequisites for collecting consent and ensuring consent-based processing of information. The principle of consent (Art. 7, Bensoussan et al. 2018; Nicolaidou and Georgiades 2017; Voigt and Von Dem Bussche 2017) is arguably one of the pivotal concepts of EU-GDPR and an expression of the right to informational self-determination. It can be defined as ability for each individual to determine whether and to what ends information about themselves can be processed (Mitrou 2017). The related concepts of conditionality, granularity and specificity are the most challenging for data management (European Data Protection Board 2018a) compared to practices before EU-GDPR, when consent was mostly obtained through the bulk acceptance of general conditions. Conditionality (art. 7 § 4) means that consent for processing activities cannot be bundled in general conditions, and that a difference should be made between necessary and optional processing activities for a given purpose. Granularity (R. 43) implies that each processing activity and related consent item must be presented separately. Specificity prescribes a 1:1 relationship between processing types and consent items (i.e. yes/no question that relates to a personal data processing activity).

Consent management found a significant echo in our focus groups. During focus group 3.1, none of the participants reported solutions either in final stages nor operational. During focus group 3.2, more questions were asked regarding consent management than all other capabilities combined. The Allmed project goal was making consent information accessible and readable by all systems, which mirror capabilities “distribute consent” and “enforce consent-based processing”. However, difficulties arouse in two areas. First, the system would need to be connected to every system storing and processing personal data – identification of such systems proved difficult and the existing system landscape documentation was deemed insufficient (see the capability “define protected data scope”). Second, the team struggled to identify consent items, as they were usually contained in unstructured form (e.g. within general conditions, contracts, webpages). A specific sub-capability was added to reflect this issue, and is a prerequisite to all other consent-related capabilities. The resulting sub-capabilities are:

- Implement consent items: define and implement consent items that mirror data processing activities performed throughout business processes.
- Record consent instances: collect and record consent expressed by individuals.
- Distribute consent: ensure consent items updates in all affected processing systems.
- Enforce consent-based processing: ensure that data processing activities are performed in accordance with consent expressed by individuals.

While the concept of consent itself was not new, the granularity requirements pose a significant technical challenge. They imply that additional data reflecting consent must be collected and taken into account when processing the related data objects, in both manual and automated processing scenarios. From a tool perspective, data management, compliance management and access management solutions started offering specific modules to record user consent, especially in scenarios involving a web-based user frontend. However, even though such tools provide technical means to communicate and acquire consent, organizations are still faced with the issues of defining and implementing consent items. In that regard, the difficulties encountered by Allmed's team in defining what items should be recorded once again highlights that technical implementations are dependent on the availability of clear documentation regarding processing activities. This difficulty is echoed in Kurtz, Wittner, et al. (2020)'s recent study, which highlights inadequacies between stated and effective processing purposes in digital service offerings, and provides remediating design goals to be considered in the development of IT artefacts. Huth, Both, et al. (2020) also explore the necessity for technical and development teams to recognize consent items and subsequently investigate implementation modalities – they suggest a prototype to support related discussions between development and business teams in the context of agile software engineering. Both studies suggest that consent requirements, due to their impact on final designs, should be integrated in early stages of IT development, which can explain difficulties in bringing pre-existing solutions to compliance.

Table 31. Capability overview: Manage consent

Sub-Capability	Description	Specification	Implementation	Evidence
Implement consent items *(A2.1)*	Ability to define and implement consent items that mirror data processing activities performed by business entities	Define consent items (i.e. yes/no questions reflecting processing purposes requiring explicit consent) Translate consent items into machine-readable data attrib-utes	Data attributes Metadata Consent management tool	Art. 5-7 R. 33, R. 50
Collect consent instances *(A2.2)*	Ability to collect and record consent from targeted individuals	Enable individuals to provide consent for specific processing types and to change it (e.g. adding new, modifying or withdrawing existing)	Self-service portal Request-based approach Consent management tool	Art. 5-7 R. 32, 42
Ditribute consent *(A2.3)*	Ability to keep consent instances updated throughout all im-pacted systems	Changes of consent recorded in one system should be propagated within all other systems containing personal data about the targeted individual	Data centralization Data integration platform	Art. 5-7

Sub-Capability	Description	Specification	Implementation	Evidence
Enforce consent-based processing *(A2.4)*	Ability to control (i.e. enable or restrict) data processing activi-ties on the basis of consent items	Ensuring that consent items are actually taken into account throughout business processes	Meta-data attributes readable by processing systems Data catalogs Authorization concept	Art. 5-7 R. 40, 42

4.3.3 Enable data processing rights

This capability denotes the ability to process data according to EU-GDPR's data rights and principles. It was derived from the principle of accountability (art. 24 § 1), but covers only the technical aspects to reach compliance, document them, and provide proof of compliance (Bensoussan et al. 2018; Nicolaidou and Georgiades 2017; Voigt and Von Dem Bussche 2017).

Art. 17 provisions a "right of erasure", according to which individuals can request that organizations delete their personal data (provided that they have no other obligation to keep said data). From a technical perspective, enterprise systems usually prevent users from deleting data and practitioners expressed a difficulty in that regard. When asked about it, none of the participants of focus group 3.1 reported that they had operational deletion processes or mechanisms, and participants expressed a lack of well-established solutions at this level. Art. 25 mandates privacy by design / by default approaches, including the principle of minimization (Voigt and Von Dem Bussche 2017), i.e. processing as little personal data as possible. One way of operationalizing it is pseudonymization, which is a rare occurrence of EU-GDPR mentioning a specific technological approach (R. 28-29). Pseudonymization or anonymization are a form of data processing – in that sense, while they may, for instance, enable organizations to lighten the depth of privacy assessment requirements or to use a legal basis other than consent, processing of anonymized/pseudonymized data should still have a legal basis and does not free organizations of all data protection responsibilities (Groos and Veen 2020) - for this reason, pseudonymization was added as second order capability. Art. 20 introduces a "right do data portability" – organizations are required to transmit personal data records "in a structured, commonly used and machine-readable format" to individuals, and, in some cases, directly to other organizations. Researchers have highlighted that facilitating customer movement between (competing) organizations is a prime example of conflicting interests between compliance and business mandates (Engels 2016; Wohlfarth 2019). From a technical standpoint, during focus group 3.1, only a quarter of respondents declared that the provision of data in standardized formats was mature, and none of them reported working communication channels. The resulting sub-capabilities can be summarized as follows:

- Delete data: permanently remove data records from their systems.
- Pseudonymize data: use pseudonymization techniques in order to adhere to the principle of minimization.
- Transmit data in standardized form: transmit personal data to external parties using standard formats and set up communication channels with other organizations.

From a tool perspective, features related to data deletion and data transfer seem to be the least frequent – they are only enabled by two solutions, and only one of them supports the enforcement of a data retention policy. Data anonymization and data pseudonymization are also rather uncommon due to the fact that encryption mechanisms, which are generally already in use for security purposes, seem to be favored. Furthermore, unless the keys are kept by a trusted third party, symmetric encryption is a not a valid means for data pseudonymization, as it can be reversed. These shortcomings have also been relayed in research, and a number of studies investigate blockchain as technological foundation to design systems that enable such capabilities (Faber et al. 2020; Farshid et al. 2019; Guggenmos et al. 2020; Mejtoft et al. 2019; Rieger et al. 2019).

Table 32. Capability overview: Enable data processing rights (art. 24 § 1)

Capability	Description	Specification	Implementation	Evidence
Delete data *(A3.1)*	Ability to permanently remove data records	Deletion of all storage instances of data objects / attributes should be carried out if they no longer serve a purpose, or upon request, pending other legal obligations	Deletion functionality Automated deletion processes Blockchain (Farshid et al. 2019)	Art. 17 R. 65, 66
Pseudonymize data *(A3.2)*	Ability to process personal data in a way that they cannot be linked to a specific individual without complementary information	Data that is pseudonymized, e.g. that is stripped of any link to a specific individual, does not fall into the scope of EU-GDPR, provided it is not reversible. In case of reversibility, information enabling linkage to individuals must be kept outside the organization, by a trusted third party Pseudonymization should be used whenever possible.	Cryprography tools Hash functions	Art. 4 § 5 Art. 5 § 1(c) Art. 11 Art. 25 R. 28, 29, 78

Capability	Description	Specification	Implementation	Evidence
Transmit data in standardized form *(A3.3)*	Ability to transmit data to third parties using free and/or interoperable formats	*Right of access*: communicate complete list of data records in a freely readable format / media *Right of portability*: communicate data records in a machine-readable way (to the individual or directly to a designated third party)	*Right of access*: PDF, OpenDocument *Right of portability*: XML-based formats (+ dedicated communication channels if applicable) Blockchain (Faber et al. 2020)	Art. 15 Art. 20 R. 59, 68

4.4 Organizational capabilities

4.4.1 Orchestrate data protection activities

This capability denotes the organizational ability to coordinate and execute data protection activities, involving different roles and responsibilities. It was derived from the organizational component of the principle of accountability (Bensoussan et al. 2018; Nicolaidou and Georgiades 2017; Voigt and Von Dem Bussche 2017). As stated, focus group feedback indicated that data managers often are at a loss as of who to consult when faced with data protection inquiries. This became particularly clear during the Allmed project – when the team needed to obtain information regarding data protection matters, they did not have a clearly designated contact person. On several occasions, responsibilities (e.g., for defining consent items) were not clearly defined. Art. 37-39 requires that organizations of a certain size appoint a “Data Protection Officer” (DPO). The DPOs should monitor compliance by acquiring an overview of processing activities, serve as advisory contact person (European Data Protection Board 2017), oversee record keeping and cooperation with authorities.

EU-GDPR also makes a distinction between data controllers and processors, and art. 28 orders the former to control compliance of the latter. This distinction is relevant to organizations when they outsource data processing to third party companies – the use of cloud services also falls into this situation, as merely storing data is considered processing. This became apparent during the Allmed project (cloud CRM) and especially in the case of Leares, which exclusively relies on cloud services (e.g., CRM, content management, websites) for the storage and processing of data. A corresponding capability was therefore added. The resulting sub-capabilities are:

- **Assume data protection responsibilities:** responsibilities for data protection-related tasks in all business functions that routinely process personal data.
- **Oversee data protection activities:** a leading role should oversee, organize, control and coordinate data protection activities.

- **Control compliance of external processors:** monitor that data processing conducted by third parties for EU-GDPR compliance.

The DPO is a role that existed before EU-GDPR in some organizations (e.g., “Privacy officer”) and that has been made mandatory by the regulation. However, even though the responsibilities are generally described in the regulation, the specificities remain unclear. In their recent EU-GDPR readiness study, Dansac Le Clerc and Mannent (2020) find that most DPOs have a legal background (62%), and only 21% of them are experts in IT/digital domains. They also report that DPOs “have less experience in IT and security than they judge necessary”. This corroborates statements from representatives of Versuisse and Svizzance, where DPOs were both coming from legal teams. Dansac Le Clerc and Mannent (2020) also highlight that EU-GDPR efforts should involve a “chain of compliance and responsibilities”, extending beyond legal teams towards IT, security, HR and business units. At Svizzance, the DPO reached out to the data management team and formed a partnership with one team member, who acts as a technical advisor to the DPO and coordinates decisions with the data management team. One of the outcomes of this partnership was a decision to refine existing enterprise architecture documentations of system and processes in light of data protection requirements. These empirical evidences highlight the need for better communication and alignment (“common ground”) between legal and data management teams. In fact, Huth, Burmeister, et al. (2020) analyzed the collaboration between enterprise architecture and data protection teams, and found evidence that using enterprise architecture as basis for joint documentation was a successful course of action. In addition, tools from the data management, compliance management and security/protection categories can also assist organizations in defining governance and centralizing workflows and policies.

When it comes to making sense of third-party data processing, enterprise architecture may also be used to document “external” processing systems (e.g., cloud storage). We have also identified two software solutions that assist organizations in maintaining an inventory of all vendors utilized. In research, two studies have been published on the matter, focusing on the issues of third-party data processing (Kurtz et al. 2018), as well as an investigation of third party data dissemination in digital service ecosystems (Kurtz et al. 2019). The latter illustrates the challenges from both legal and technical perspectives in the seemingly straightforward use case of a weather app on a smartphone, triggering data transmissions to operating system provider, the app developer and an underlying API provider.

Table 33. Capability overview: Orchestrate data protection activities (art. 24 § 1)

Sub-Capability	Description	Specification	Implementation	Evidence
Assume data protection responsibilities *(B1.1)*	Ability to establish data protection responsibilities in business functions	Each business function should have people: (1) acting as main contact for data protection matters (2) carrying out data protection-related processes and provid-ing input for data protection tasks	Role model Company-wide directory Collaborative / communication platforms	Art. 26-28 Art. 37-39 R. 74
Oversee data protection activities *(B1.2)*	Ability to oversee, organize, control and coordinate data protection activities	Role entailing advisory, control and cooperation with authorities	Appointment of data protection officer (over-arching, coordination role)	Art. 31 Art. 37-39 R. 48, 97
Control compliance of external processors *(B1.3)*	Ability to ensure that processing activities conducted by external processor are compliant with legal requirements	Only processors providing sufficient guarantee of EU-GDPR compliance should be selected Documentation of processing activities Collaboration between organization and processors to guarantee the exercise of rights and proof of compliance Deletion or restitution of data at the end of the contract	Contract with processors with enhanced data protection terms Vendor inventory tools Enterprise architecture tools	Art. 24 § 2 Art. 28 R. 58, 74, 78, 81-83, 101, 108, 111

4.4.2 Demonstrate compliant data processing

This capability comprises the ability to record and evaluate sensitive processing activities, as well as to document system landscapes. It was derived from the documentation component of the principle of accountability (Nicolaidou and Georgiades 2017; Voigt and Von Dem Bussche 2017). Art. 30 orders organizations to “maintain a record of processing activities under its responsibility” and details the contents of such documentation. It was identified as a significant difficulty by Iannopollo et al. (2018), and all participants of focus group 3.2 acknowledged that documentation represented a significant effort. Maintaining system landscape documentation was identified as another sub-capability, as the experts indicate that most organizations have difficulties locating data – this was the very motivation for the Engger project, and one significant roadblock for Allmed’s solution implementation. Art. 35-36 further require organizations to conduct and document in-depth data protection impact assessments (DPIA) when performing sensitive processing activities. The resulting sub-capabilities are:

- **Maintain records of processing activities:** inventory and document personal data-related activities performed throughout business processes.
- **Maintain documentation of system landscape:** inventory and document systems that store and process personal data on a regular basis.
- **Supervise sensitive processing activities:** identify and evaluate sensitive data processing activities.

As mentioned previously (s. section 4.3.1), documentation of systems that store and process personal data, as well as of processing **purposes,** is a pillar of EU-GDPR compliance. From a tool perspective, solutions offer functionalities that can support and streamline the documentation process. Tools from most categories offer functionalities to detect irregularities in data use and detect data breaches, with tools in the protection/security category taking the lead. While the majority of surveyed solutions offer logging capabilities, which can prove useful to investigate suspected or known incidents, only three of them assist organizations in running DPIAs. This is partly due to the fact that DPIAs are about interpreting the purpose of specific processing activities and evaluating potential nefarious real-world consequences for individuals. Researchers have highlighted the need to refine currently available methods, and further investigate tool implementation (Vemou and Karyda 2018). Another aspect lies in the sensitivity of certain processing activities due to their novel technological underpinnings. There is an ongoing debate on the data protection-related risks of technologies such as Big Data Analytics and Artificial Intelligence (Addis and Kutar 2020, 2018), and EU-GDPR specifically considers decisions that are the result of automated decision-making processes (art. 22). Conversely, researchers are investigating ways to incorporate ethics and transparency considerations into the design of big data artefacts and algorithms (Tona et al. 2018; Watson and Nations 2019)**.**

Table 34. Capability overview: Demonstrate compliant data processing

Sub-Capability	Description	Specification	Implementation	Evidence
Maintain records of processing activities *(B2.1)*	Ability to inventory and document personal data-related activities	Inventory and describe all processing activities in terms of: basis of processing, data used, purpose, means (e.g. use of analytics), consent items requested	Enterprise Architecture Documentation templates Process maps Data flows / lineage Review processes	Art. 25 Art. 30 R. 39, 44-50, 78

Sub-Capability	Description	Specification	Implementation	Evidence
Maintain documentation of system landscape *(B2.2)*	Ability to inventory and document systems storing and processing personal data	Inventory and describe all systems processing personal data in terms of stored data types, and processing capabilities	Enterprise Architecture Documentation templates Enterprise architecture tools / maps Review processes	Art. 30
Supervise sensitive processing activities *(B2.3)*	Ability to assess and document the privacy consequences of sensitive processing activities in details	Required when the scale of processing, and/or the sensitivity of data processed or technology used pose high privacy risks (e.g. advanced analytics, profiling). Subjected to specific authorization if satisfactory privacy measures cannot be implemented	Documentation templates Review processes	Art. 35 Art. 36 R. 51-56, 84, 90-92, 94

4.4.3 Disclose information

This capability involves the ability to disclose information to individuals (R. 58) and authorities (art. 31). It was derived from the principle of transparency, which requests data protection measures to be clearly exposed (Bensoussan et al. 2018; Nicolaidou and Georgiades 2017).

Transparency requirements apply in two cases (European Data Protection Board 2018b). First, at the point of data collection, organizations must present related information separately, in a manner (e.g., language, illustrations) that can be easily comprehended. Transparency also refers to communications with individuals after data is collected, when organizations are faced with right-related requests (e.g., access, rectification, deletion). Art. 31 specifies that organizations "shall cooperate, on request, with the supervisory authority in the performance of its tasks". This implies that organizations set up a contact person for authorities (usually the DPO), and the ability to present relevant information / documentation as proof of compliance.

These capabilities may be seen as the operationalization of the principle of accountability, which is materialized by documentation. Since such documentation should contain all relevant information regarding an organization's data protection practices, these capabilities are about presenting that information to the interested parties (i.e., individuals and authorities). The resulting sub-capabilities are:

- **Disclose information to individuals:** provide individuals with complete and understandable information regarding the processing of their personal data and respond to their data protection-related requests.

- **Disclose information to authorities:** collaborate with designated data protection authorities and communicate relevant information upon request.

Clear documentation of processing bases, including consent, is mandatory to demonstrate compliance (s. section 4.4.2) and ensure that consent is reflected in terms of data (s. section 4.3.2). A variety of tools among all considered categories exhibit audit trail functionalities, which record and gather information about data processing and related events. However, transparency requirements also relate to the way these bases and consent items are presented to users, and there is evidence of discrepancies between the way processing purposes are communicated and the actual data processing occurs (Kurtz et al. 2020). Bergram et al. (2020) also analyze common methods for acquiring consent in digital settings, and both studies derive design recommendations for enlightened user consent.

Table 35. Capability overview: Disclose information

Sub-Capability	Description	Specification	Implementation	Evidence
Disclose information to individuals *(B3.1)*	Ability to respond to data protection-related requests from data subjects, and communicate data processing activities in clear terms	Data records must be disclosed to data subjects in exercise of the following rights: access, rectification, portability and deletion Information about processing activities and consent items must be presented in clear, everyday language, separately from other terms and agreements Data breach notification	Self-service portal Request-based approach Contact person for individuals Incident response processes	Art. 12-23 Art. 34 Art. 36 Art. 38 § 4 R. 39, 58-73, 86
Disclose information to authorities *(B3.2)*	Ability to collaborate with designated government bodies and communicate requested information	Records of processing activities Evidence of compliance and security measures Data breach notification	Contact person for authorities Incident response processes Documentation material	Art. 17 Art. 31 Art. 33 Art. 39 § 1(d) R. 85, 87, 89, 94

5 Building the capabilities

The focus groups, the case study and expert interviews helped us validating the capabilities, but also provided insights into how companies build these capabilities and the relationships between them. Figure 12 depicts the articulation of the capabilities – arrows describe dependencies, where the source is a prerequisite to the target. Through our exploration, we identified that documentation-related capabilities, as well as the ability to inventory and locate data objects are central to achieve EU-GDPR compliance, and can be considered enabling capabilities (highlighted with black borders in Figure 12) – they are interdependent and should be the starting point of EU-GDPR implementation approaches.

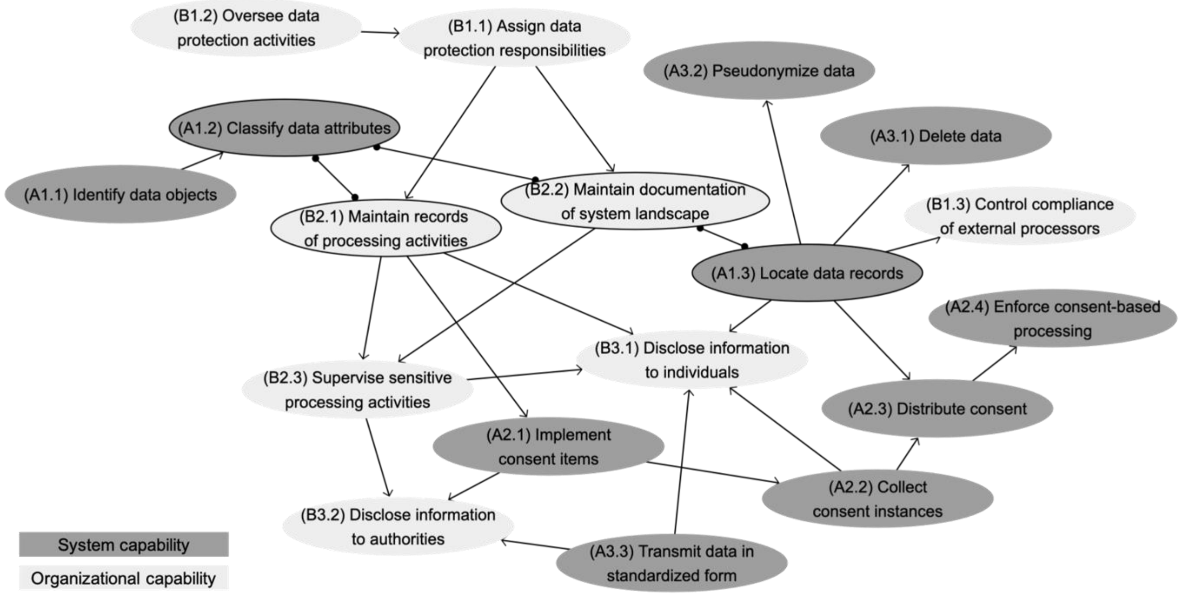

(B1.2) Oversee data protection activities
(B1.1) Assign data protection responsibilities
(A3.2) Pseudonymize data
(A1.2) Classify data attributes
(A3.1) Delete data
(A1.1) Identify data objects
(B2.2) Maintain documentation of system landscape
(B1.3) Control compliance of external processors
(B2.1) Maintain records of processing activities
(A1.3) Locate data records
(A2.4) Enforce consent-based processing
(B3.1) Disclose information to individuals
(B2.3) Supervise sensitive processing activities
(A2.3) Distribute consent
(A2.1) Implement consent items
(A2.2) Collect consent instances
(B3.2) Disclose information to authorities
System capability
(A3.3) Transmit data in standardized form
Organizational capability

Figure 12. Relationships between capabilities

This was especially apparent in the case of Svizzance – the company had achieved major progress on the documentation of processing activities and system landscape, and was in the process of linking this documentation with data records (i.e., identifying data objects, classifying data attributes and locating data records). However, it had not yet started activities related to the "manage consent" and "integrate data processing requirements" capabilities, as having a clearly defined protected data scope was viewed as a prerequisite by our respondents. A similar pattern was identified at Versuisse, which was more mature on defining the protected data scope and had just started venturing into realizing the "manage consent" and "integrate data processing requirements" capabilities at an enterprise-wide level. Being insurance companies, Versuisse and Svizzance both operate in a highly regulated market. Such organizations traditionally put an emphasis on control activities and maintain a thorough documentation of their operations, which can explain the top-down approach that we have observed (i.e., starting from high-level documentation and investigating links and relationships with data objects). However, we argue that organizations operating in markets with lower regulatory pressure, or who are, by nature, data-driven, could also adopt a bottom-up approach, i.e., starting from an inventory and classification of data objects to build or enhance their documentation of processing activities and system landscape. Leares constitutes an example of the bottom-up approach, as such documentation did not exist and was built alongside inventorying systems processing data in the protected data scope.

6 Conclusion and outlook

This paper introduces a data management perspective to EU-GDPR and argues that the regulation requires companies to build dedicated data management capabilities, comprising both technical and organizational capabilities. The suggested capability model was developed in an iterative design science process, integrating both interpretation of legal texts, analysis of academic publications and practical insights from focus groups with more than 30 experts and from 3 EU-GDPR projects. By translating compliance requirements into organizational and system capabilities, we make two academic contributions: first, we analyze EU-GDPR - representing the latest generation of data protection regulations - using concepts from the regulatory compliance management literature (El Kharbili 2012). This allows us to systematically interpret and translate data protection compliance requirements for the IS community. Second, the suggested capability model contributes to building common ground between legal and data management domains and extending the scope beyond the isolated investigation of specific implementation options. The capability model thereby complements the emerging body of research on EU-GDPR, by classifying and integrating these focused research efforts into an enterprise-wide perspective, and by proposing an all-encompassing perspective on the *what* of EU-GDPR implementation, rather than the *how*. Our capability model, taken together with practical insights and the analysis of available tools, reveals that no single tool or approach is able to remediate all of the regulation's requirements. While specific solutions are availabe to tackle selected aspects of the regulation, the capability model acts as a framework to determine the scope of these solutions and assess gaps and priorities.

For practice, the capability model supports companies in developing a systematic approach towards achieving EU-GDPR compliance and monitoring progress, instead of "fire-fighting". The capability model can be used in different ways: first, as basis for assessing the current capabilities, identifying the required capabilities and prioritizing them; second, for analyzing software features that help in realizing the capabilities, and mapping them to existing market offerings, further informing EU-GDPR implementation initiatives; third, the capability model also supports the development of capabilities by monitoring progress towards achieving compliance.

As implications for future research, our findings reveal that implementing EU-GDPR is not a one-time effort, but an ongoing process of building system and organizational capabilities. The suggested capability model may serve as a basis for studying in more details how the capabilities are being built and how they can be maintained efficiently. Since companies have to comply with an increasing number of data protection regulations (for instance, the California Consumer Privacy Act (California State Senate 2018), the upcoming revised Swiss Federal Data Protection Act (Métille and Raedler 2017)), our findings may serve as basis for analyzing commonalities and differences between regulations as well as identifying a "common core". Research should also study the benefits of capabilities beyond compliance, and specify how they may serve other purposes in organizations, as a way to outweigh potential conflicting interest that some researchers have identified (Engels 2016; Grundstrom et al. 2020; Jakobi et al. 2020; Lindgren 2020; Martin and Matt 2018; Wohlfarth 2019). Some of the related difficulties, such as achieving transparency on data flows, storage and usage, are not specific to EU-GDPR and can also negatively impact value-adding data-driven activities in the enterprise, suggesting that the dichotomy between compliance and business value should be challenged.

References

Abdullah, N. S., Indulska, M., and Shazia, S. 2009. "A Study of Compliance Management in Information Systems Research," in *Proceedings of the 17th European Conference on Information Systems (ECIS)*, Verona, Italy, June 8, p. 5.

Addis, C., and Kutar, M. 2020. "General Data Protection Regulation (GDPR), Artificial Intelligence (AI) and UK Organisations: A Year of Implementation of GDPR," in *Proceedings of the 25th UK Academy for Information Systems Annual Conference (UKAIS)*, Oxford, United Kingdom, April 31. (https://aisel.aisnet.org/ukais2020/24).

Addis, M., and Kutar, M. 2018. "The General Data Protection Regulation (GDPR), Emerging Technologies and UK Organisations: Awareness, Implementation and Readiness," in *Proceedings of the 23rd UK Academy for Information Systems Annual Conference (UKAIS)*, Oxford, United Kingdom, March 20, p. 29. (https://aisel.aisnet.org/ukais2018/29).

Alboaie, L. 2017. "Towards a Smart Society through Personal Assistants Employing Executable Choreographies," in *Proceedings of the 26th International Conference on Information Systems Development (ISD)*, Larnaca, Cyprus, September 6, p. 5.

Baiyere, A., and Salmela, H. 2014. "Towards a Unified View of Information System (IS) Capability," in *Proceedings of the 18th Pacific Asia Conference on Information Systems (PACIS)*, Chengdu, China, June 24, p. 329.

Bélanger, F., and Crossler, R. E. 2011. "Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems," *MIS Quarterly* (35:4), pp. 1017–1042.

Bensoussan, A., Avignon, C., Bensoussan-Brulé, V., Forster, F., and Torres, C. 2018. *Règlement Européen Sur La Protection Des Données: Textes, Commentaires et Orientations Pratiques*, (2nd ed.), Brussels: Bruylant.

Bergram, K., Bezençon, V., Maingot, P., Gjerlufsen, T., and Holzer, A. 2020. "Digital Nudges for Privacy Awareness: From Consent to Informed Consent?," in *Proceedings of the 28th European Conference on Information Systems (ECIS)*, Marrakesh, Morocco, June 15, p. 64. (https://aisel.aisnet.org/ecis2020_rp/64).

Bharadwaj, A. 2000. "A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Investigation," *MIS Quarterly* (24:1), pp. 169–196.

Burmeister, F., Drews, P., and Schirmer, I. 2019. "A Privacy-Driven Enterprise Architecture Meta-Model for Supporting Compliance with the General Data Protection Regulation," in *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, January 8. (https://doi.org/10.24251/HICSS.2019.729).

Burmeister, F., Huth, D., Drews, P., Schirmer, I., and Matthes, F. 2020. "Enhancing Information Governance with Enterprise Architecture Management: Design Principles Derived from Benefits and Barriers in the GDPR Implementation," in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, January 7. (https://doi.org/10.24251/HICSS.2020.688).

Butler, T., and O'Brien, L. 2019. "Understanding RegTech for Digital Regulatory Compliance," in *Disrupting Finance: FinTech and Strategy in the 21st Century*, Palgrave Studies in Digital Business & Enabling Technologies, T. Lynn, J. G. Mooney, P. Rosati, and M. Cummins (eds.), Cham: Springer International Publishing, pp. 85–102. (https://doi.org/10.1007/978-3-030-02330-0_6).

California State Senate. 2018. *California Consumer Privacy Act*.

Capgemini Research Institute. 2019. “Championing Data Protection and Privacy - A Source of Competitive Advantage in the Digital Century,” Consultancy Report, Consultancy Report, , September. (https://www.capgemini.com/wp-content/uploads/2019/09/Report_Championing-Data-Protection-and-Privacy.pdf).

Cleven, A., and Winter, R. 2009. “Regulatory Compliance in Information Systems Research - Literature Analysis and Research Agenda,” in *Enterprise, Business Process and Information Systems Modeling*, Lecture Notes in Business Information Processing, Berlin, Heidelberg: Springer-Verlag, pp. 174–186.

Dansac Le Clerc, M., and Mannent, P. 2020. “GDPR Survey - Benefits Beyond Compliance,” Consultancy Report, Consultancy Report, Baker McKenzie & BearingPoint, April. (https://www.bakermckenzie.com/-/media/files/insight/publications/2020/04/gdpr_survey.pdf?la=en).

De Hert, P., and Papakonstantinou, V. 2012. “The Proposed Data Protection Regulation Replacing Directive 95/46/EC: A Sound System for the Protection of Individuals,” *Computer Law & Security Review* (28:2), pp. 130–142. (https://doi.org/10.1016/j.clsr.2012.01.011).

De Hert, P., and Papakonstantinou, V. 2016. “The New General Data Protection Regulation: Still a Sound System for the Protection of Individuals?,” *Computer Law & Security Review* (32:2), pp. 179–194. (https://doi.org/10.1016/j.clsr.2016.02.006).

Debet, A., Massot, J., and Métallinos, N. 2015. *Informatique et Libertés: La Protection Des Données à Caractère Personnel En Droit Français et Européen*, Les Intégrales, Issy-les-Moulineaux: Lextenso.

Deutsche Telekom. 2016. “Binding Interpretations: General Data Protection Regulation (GDPR),” Deutsche Telekom, November 18.

Diamantopoulou, V., and Mouratidis, H. 2018. “Evaluating a Reference Architecture for Privacy Level Agreement’s Management,” in *Proceedings of the 12th Mediterranean Conference on Information Systems (MCIS)*, Corfu, Greece, September 28, p. 28. (https://aisel.aisnet.org/mcis2018/28).

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).

Engels, B. 2016. “Data Portability and Online Platforms The Effects on Competition,” in *Proceedings of the 29th Bled EConference (BLED)*, Bled, Slovenia, June 19, p. 39.

European Commission. 2020. “A European Strategy for Data,” Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committe of the Regions No. COM(2020) 66 final, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committe of the Regions, Brussels: European Commission. (https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593073685620&uri=CELEX%3A52020DC0066).

European Data Protection Board. 2017. “Guidelines on Data Protection Officers (WP243 Rev.01),” Regulatory Guidelines, Regulatory Guidelines, European Union, April 5.

European Data Protection Board. 2018a. “Guidelines On Consent Under Regulation 2016/679 (WP259, Rev.01),” Regulatory Guidelines, Regulatory Guidelines, European Union, April 10.

European Data Protection Board. 2018b. “Guidelines on Transparency under Regulation 2016/679 (WP260 Rev.01),” Regulatory Guidelines, Regulatory Guidelines, European Union, April 11.

Faber, B., Michelet, G. C., Weidmann, N., Mukkamala, R. R., and Vatrapu, R. 2020. “BPDIMS:A Blockchain-Based Personal Data and Identity Management System,” in *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, January 8. (https://doi.org/10.24251/HICSS.2019.821).

Farshid, S., Reitz, A., and Roßbach, P. 2019. “Design of a Forgetting Blockchain: A Possible Way to Accomplish GDPR Compatibility,” in *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, January 8. (https://doi.org/10.24251/HICSS.2019.850).

Fox, G., Tonge, C., Lynn, T., and Mooney, J. 2018. “Communicating Compliance: Developing a GDPR Privacy Label,” in *Proceedings of the 24th Americas Conference on Information Systems (AMCIS)*, New Orleans, Louisiana, USA, August 16, p. 30.

Francis, M., Covert, Q., Steinhagen, D., and Streff, K. 2020. “An Inventory of International Privacy Principles: A 14 Country Analysis,” in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*, , January 7. (https://doi.org/10.24251/HICSS.2020.534).

Grant, R. M. 1991. “The Resource-Based Theory of Competitive Advantage: Implications for Strategy Formulation,” *California Management Review* (33:3), p. 114.

Groos, D., and Veen, E.-B. van. 2020. “Anonymised Data and the Rule of Law,” European Data Protection Law Review (6:4), Lexxion Publisher, pp. 498–508. (https://doi.org/10.21552/edpl/2020/4/6)

Grundstrom, C., Väyrynen, K., Iivari, N., and Isomursu, M. 2020. “Making Sense of the General Data Protection Regulation—Four Categories of Personal Data Access Challenges,” in *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, Jaunary -11. (https://doi.org/10.24251/HICSS.2019.605).

Guadamuz, A. 2017. “Developing a Right to Be Forgotten,” in *EU Internet Law: Regulation and Enforcement*, T.-E. Synodinou, P. Jougleux, C. Markou, and T. Prastitou (eds.), Cham: Springer International Publishing, pp. 59–76. (https://doi.org/10.1007/978-3-319-64955-9_3).

Guggenmos, F., Lockl, J., Rieger, A., Wenninger, A., and Fridgen, G. 2020. “How to Develop a GDPR-Compliant Blockchain Solution for Cross-Organizational Workflow Management: Evidence from the German Asylum Procedure,” in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*, , January 7. (https://doi.org/10.24251/HICSS.2020.492).

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.

Houta, S., Meschede, C., Beeres, K., Surges, R., and Klötgen, M. 2020. “User-Centered Design and Evaluation of Standard-Based Health Technologies for Epilepsy Care,” in *Proceedings of the 28th European Conference on Information Systems (ECIS)*, Marrakesh, Morocco, June 15, p. 16. (https://aisel.aisnet.org/ecis2020_rp/16).

Huth, D., Both, A., Ahmad, J., Sauer, G., Yilmaz, F., and Matthes, F. 2020. “Process and Tool Support for Integration of Privacy Aspects in Agile Software Engineering,” in *Proceedings of the 26th Americas Conference on Information Systems (AMCIS)*, Salt Lake City, Utah, USA, August 15, p. 6. (https://aisel.aisnet.org/amcis2020/systems_analysis_design/systems_analysis_design/6).

Huth, D., Burmeister, F., Matthes, F., and Schirmer, I. 2020. “Empirical Results on the Collaboration Between Enterprise Architecture and Data Protection Management during the Implementation of the GDPR,” in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, January 7. (https://doi.org/10.24251/HICSS.2020.715).

Huth, D., and Matthes, F. 2019. “‘Appropriate Technical and Organizational Measures’: Identifying Privacy Engineering Approaches to Meet GDPR Requirements,” in *Proceedings of the 25th Americas Conference on Information Systems (AMCIS)*, Cancún, Mexico, August 15, p. 4. (https://aisel.aisnet.org/amcis2019/info_security_privacy/info_security_privacy/5).

Iannopollo, E., Balaouras, S., and Harrison, P. 2017. “The Five Milestones to GDPR Success,” Executive Brief, Executive Brief, Forrester Research, April 25.

Iannopollo, E., Balaouras, S., Pikulik, E., and Dostie, P. 2018. “The State of GDPR Readiness,” Executive Brief, Executive Brief, Forrester Research, January 31.

Jakobi, T., von Grafenstein, M., Legner, C., Labadie, C., Mertens, P., Öksüz, A., and Stevens, G. 2020. “The Role of IS in the Conflicting Interests Regarding GDPR,” *Business & Information Systems Engineering* (62:3), pp. 261–272. (https://doi.org/10.1007/s12599-020-00633-4).

Karyda, M., and Mitrou, L. 2016. “Data Breach Notification: Issues and Challenges for Security Management,” in *Proceedings of the 10th Mediterranean Conference on Information Systems (MCIS)*, Paphos, Cyprus, September 4, p. 60.

Kurtz, C., Semmann, M., and Böhmann, T. 2018. “Privacy by Design to Comply with GDPR: A Review on Third-Party Data Processors,” in *Proceedings of the 24th Americas Conference on Information Systems*, , August 16. (https://aisel.aisnet.org/amcis2018/Security/Presentations/36).

Kurtz, C., Wittner, F., Semmann, M., Schulz, W., and Böhmann, T. 2019. “The Unlikely Siblings in the GDPR Family: A Techno-Legal Analysis of Major Platforms in the Diffusion of Personal Data in Service Ecosystems,” in *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, January 8. (https://doi.org/10.24251/HICSS.2019.607).

Kurtz, C., Wittner, F., Vogel, P., Semmann, M., and Böhmann, T. 2020. “Design Goals for Consent at Scale in Digital Service Ecosystems,” in *Proceedings of the 28th European Conference on Information Systems (ECIS)*, Marrakesh, Morocco, June 15, p. 69. (https://aisel.aisnet.org/ecis2020_rp/69).

von der Leyen, U. 2020. *Statement by the President at ‘Internet, a New Human Right’*. (https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_20_1999).

Lindgren, P. 2020. “The Impact on Multi Business Model Innovation Related to GDPR Regulation,” in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, January 7. (https://doi.org/10.24251/HICSS.2020.537).

Martin, N., and Matt, C. 2018. “Unblackboxing the Effects of Privacy Regulation on Startup Innovation,” in *Proceedings of the 39th International Conference on Information Systems (ICIS)*, San Francisco, California, USA, December 13, p. 11. (https://aisel.aisnet.org/icis2018/security/Presentations/11).

Martinez, J., Ruiz, A., Puelles, J., Arechalde, I., and Miadzvetskaya, Y. 2019. “Smart Grid Challenges through the Lens of the European General Data Protection Regulation,” in

Proceedings of the 28th International Conference on Information Systems Development (ISD), Toulon, France, August 28, p. 4. (https://aisel.aisnet.org/isd2014/proceedings2019/Society/4).

Maunula, G. 2020. "Advancing Technological State-of-the-Art for GDPR Compliance: Considering Technology Solutions for Data Protection Issues in the Sharing Economy," *Journal of the Midwest Association for Information Systems (JMWAIS)* (2020:2). (https://doi.org/10.17705/3jmwa.000060).

Mejtoft, T., Hellman, D., and Söderström, U. 2019. "Reclaiming Control over Personal Data with Blockchain Technology : An Exploratory Study," in *Proceedings of the 32nd Bled EConference (BLED)*, Bled, Slovenia, June 16, pp. 411–425. (http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-160870).

Merlivat, S., Iannopollo, E., Parrish, M., Khatibloo, F., Oesterreich, M., Liu, S., and Turley, C. 2017. "Digital Advertising under GDPR Hinges on Data Management," Executive Brief, Executive Brief, Forrester Research, November 15.

Métille, S., and Raedler, D. 2017. "Swiss Data Protection Act Reform Set in Motion," *Data Protection Leader* (14:2), pp. 14–16.

Mitrou, L. 2017. "The General Data Protection Regulation: A Law for the Digital Age?," in *EU Internet Law: Regulation and Enforcement*, T.-E. Synodinou, P. Jougleux, C. Markou, and T. Prastitou (eds.), Cham: Springer International Publishing, pp. 19–57. (https://doi.org/10.1007/978-3-319-64955-9_2).

Nicolaidou, I. L., and Georgiades, C. 2017. "The GDPR: New Horizons," in *EU Internet Law: Regulation and Enforcement*, T.-E. Synodinou, P. Jougleux, C. Markou, and T. Prastitou (eds.), Cham: Springer International Publishing, pp. 3–18. (https://doi.org/10.1007/978-3-319-64955-9_1).

Pankowska, M. 2018. "Privacy Awareness in the GDPR Implementation Circumstances," in *Proceedings of the 27th International Conference on Information Systems Development (ISD)*, Lund, Sweden, August 22, p. 5. (https://aisel.aisnet.org/isd2014/proceedings2018/Transforming/5).

Paul, C., Scheibe, K., and Nilakanta, S. 2020. "Privacy Concerns Regarding Wearable IoT Devices: How It Is Influenced by GDPR?," in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, January 7. (https://doi.org/10.24251/HICSS.2020.536).

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).

Petkov, P., and Helfert, M. 2017. "Identifying Emerging Challenges for ICT Industry in Ireland: Multiple Case Study Analysis of Data Privacy Breaches," in *Proceedings of the 23rd Americas Conference on Information Systems (AMCIS)*, Boston, Massachussetts, USA, August 10, p. 40.

Peyret, H., Cullen, A., McKinnon, C., Blissent, J., Iannopollo, E., Kramer, A., and Lynch, D. 2017. "Enhance Your Data Governance to Meet New Privacy Mandates," Consortium Report, Consortium Report, Consortium Report, Forrester Research.

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).

Presthus, W., and Sørum, H. 2019. "Consumer Perspectives on Information Privacy Following the Implementation of the GDPR," *International Journal of Information Systems and Project Management* (7:3), pp. 19–34.

Proferes, N., and Walker, S. 2020. "Researcher Views and Practices around Informing, Getting Consent, and Sharing Research Outputs with Social Media Users When Using Their Public Data," in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii , USA, January 7. (https://doi.org/10.24251/HICSS.2020.290).

Rieger, A., Guggenmos, F., Lockl, J., Fridgen, G., and Urbach, N. 2019. "Building a Blockchain Application That Complies with the EU General Data Protection Regulation," *MIS Quarterly Executive* (18:4), pp. 263–279.

Rösch, D., Schuster, T., Waidelich, L., and Alpers, S. 2019. "Privacy Control Patterns for Compliant Application of GDPR," in *Proceedings of the 25th Americas Conference on Information Systems (AMCIS)*, Cancún, Mexico, August 15. (https://aisel.aisnet.org/amcis2019/info_security_privacy/info_security_privacy/27).

Russell, K. D., O'Raghallaigh, P., O'Reilly, P., and Hayes, J. 2018. "Digital Privacy GDPR: A Proposed Digital Transformation Framework," in *Proceedings of the 24th Americas Conference on Information Systems (AMCIS)*, New Orleans, Louisiana, USA, August 16, p. 36.

Sadiq, S., Governatori, G., and Namiri, K. 2007. "Modeling Control Objectives for Business Process Compliance," in *Proceedings of the 5th International Conference on Business Process Management (BPM)*, Brisbane, Australia, September 24, pp. 149–164.

Tona, O., Someh, I. A., Mohajeri, K., Shanks, G., Davern, M., Carlsson, S., and Kajtazi, M. 2018. "Towards Ethical Big Data Artifacts: A Conceptual Design," in *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS)*, Waikoloa Village, Hawaii, USA, January 3. (https://doi.org/10.24251/HICSS.2018.571).

Veiga, A. D., Vorster, R., Li, F., Clarke, N., and Furnell, S. 2018. "A Comparison of Compliance with Data Privacy Requirements in Two Countries," in *Proceedings of the 26th European Conference on Information Systems (ECIS)*, Portsmouth, United Kingdom, June 23, p. 86. (https://researchportal.port.ac.uk/portal/en/publications/a-comparison-of-compliance-with-data-privacy-requirements-in-two-countries(4d4793b4-5c62-475b-b9b4-801e93639944)/export.html).

Vemou, K., and Karyda, M. 2018. "An Evaluation Framework for Privacy Impact Assessment Methods," in *Proceedings of the 12th Mediterranean Conference on Information Systems (MCIS)*, Corfu, Greece, September 28, p. 4. (https://aisel.aisnet.org/mcis2018/5).

Voigt, P., and Von Dem Bussche, A. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*, Cham: Springer International Publishing.

Watson, H., and Nations, C. 2019. "Addressing the Growing Need for Algorithmic Transparency," *Communications of the Association for Information Systems* (45:1), pp. 488–510. (https://doi.org/10.17705/1CAIS.04526).

Wohlfarth, M. 2019. "Data Portability on the Internet," *Business & Information Systems Engineering* (61:5), pp. 551–574. (https://doi.org/10.1007/s12599-019-00580-9).

Zhang, J., Hassandoust, F., and Williams, J. 2020. "Online Customer Trust in the Context of the General Data Protection Regulation (GDPR)," *Pacific Asia Journal of the Association for Information Systems* (12:1). (https://doi.org/10.17705/1pais.12104).

Zhang, M., Sarker, Saonee, and Sarker, Suprateek. 2013. “Drivers and Export Performance Impacts of IT Capability in ‘Born-Global’ Firms: A Cross-National Study,” *Information Systems Journal* (23:5), pp. 419–443. (https://doi.org/10.1111/j.1365-2575.2012.00404.x).

Appendix 1: List and categorization of analyzed software solutions

Vendor	Product	Category	Integrated / Toolkit	Standalone / Enhancement
Amazon Web Services	AWS	Enhancement	Integrated	Enhancement/Features of existing product
CA	CA Technologies solutions: - CA content discovery - CA identity suite - CA test data manager - CA live API manager - CA API developer portal - CA API gateway - CA PAM - CA PAM server control - CA SSO - CA cleanup - CA compliance event manager	Data Management	Toolkit	Standalone
Citrix	Citrix Workspace	Enhancement	Integrated	Standalone
Collibra	Collibra data governance platform	Data Management	Integrated	Standalone
Druva	Druva (Cloud platform)	Data Management	Integrated (platform)	Standalone

Vendor	Product	Category	Integrated / Toolkit	Standalone / Enhancement
ForgeRock	ForgeRock Identity Platform - FR Access Management - FR Directory Services - FR Identity Management - FR Identity Gateway - FR Common Services - FR Edge Security	Access Management	Integrated	Standalone
Gigya	Gigya CIAM	Access Management	Integrated	Standalone
IBM	IBM Cloud Secure Virtualization	Security /Protection	Integrated	Enhancement/Features of existing product
Imperva	Imperva - SecureSphere - CounterBreach - Camouflage	Security /Protection	Toolkit	Standalone
Informatica	Informatica-Secure@Source - Axon - Data Masking - Data Archiving - Master Data management	Data Management	Toolkit	Standalone
Janrain	Janrain Identity Cloud	Access Management	Integrated	Standalone
Microsoft	Microsoft 365: - Office 365 - Windows 10 - Enterprise Mobility + Security	Enhancement	Toolkit/Suite	Enhancement/Features of existing product

Vendor	Product	Category	Integrated / Toolkit	Standalone / Enhancement
OneTrust	OneTrust privacy management software	Compliance Management	Integrated	Standalone
Oracle	Oracle Database security products	Security /Protection	Toolkit	Enhancement of Oracle products
PingIdentity	PingIdentity: - PingDirectory - PingDataGovernance - PingID, MFA	Access Management	Integrated	Standalone
Protegrity	Protegrity - Data security platform	Security /Protection	Integrated	Standalone
Salesforce	Salesforce Shield	Security /Protection	Integrated	Enhancement (of Salesforce platform)
SAP	"SAP solutions"	Enhancement	Toolkit/Suite	Enhancement/Features of existing product
Skyhigh (McAfee)	McAfee Skyhigh Security Cloud	Security /Protection	Integrated	Standalone
Symantec	Symantec Control Compliance Suite	Compliance Management	Integrated	Enhancement/Features of existing product
TrustArc	TrustArc Platform	Compliance Management	Integrated	Standalone
VMware NSX	VMware NSX network virtualization platform	Enhancement	Integrated (enhance a toolkit)	Enhancement of VMware cloud services
Wickr	Wickr enterprise	Enhancement	Integrated	Standalone

Appendix 2: Taxonomy of EU-GDPR tool functionalities

Capability	**Function group**	**EU-GDPR reference**	**Software feature**	**Tool category**					
				Data Management (4)	Compliance Management (3)	Access Management (4)	Security/Protection (6)	Enhancement (6)	Total (23)
Define protected data scope	Data inventory	Art. 2	Identify data objects	3	2	1	4	0	10
			Classify data attributes	3	2	0	4	0	9
			Locate data records	3	2	0	4	0	9
		Art. 5	Identify related processes	3	2	1	3	0	9
		Art. 5 § 1b, recital 39	Identify who has access	3	2	1	3	0	9
Manage consent	Consent management	Art. 5-7	Collect consent	2	2	4	0	0	8
		Art. 7 § 3	Give access to data subject	0	2	1	0	0	3
Enable data processing rights	Portability	Art. 20	Data transfer	0	0	2	0	0	2
	Deletion	Art. 17	Delete/restrict upon request	1	0	1	0	0	2
		Art. 5 § 1e	Enforce retention policy	0	0	0	1	0	1

Capability	Function group	EU-GDPR reference	Software feature	Tool category					
				Data Management (4)	Compliance Management (3)	Access Management (4)	Security/Protection (6)	Enhancement (6)	Total (23)
	Obfuscation	Recital 83	Data encryption	2	0	4	5	5	16
		Art. 33-34	Key management	0	0	0	2	1	3
		Art. 33-34	Data anonymization	0	0	0	1	0	1
		Art. 25 § 1	Data pseudonymization	0	0	0	2	0	2
Orchestrate data protection activities	Policy management	Art. 32, recital 78	Define/Assign governance	2	0	0	1	0	3
			Centralize policies/workflows	2	2	0	1	0	5
	Monitoring	Art. 24, 26-29	Vendors' inventory	0	1	0	0	1	2
Demonstrate compliant proessing		Art. 35	Run DPIA	0	2	0	1	0	3
		Art. 33-34	Detect intrusion/abnormal use	2	0	0	4	2	8
			Detect data breach	1	1	0	3	1	6
	Audit trail	Art. 30-32	Log activities	4	1	2	6	3	16
Disclose information		Art. 24	Assessment	1	3	0	1	1	6

Personal data management inside and out – Integrating data protection requirements in the data life cycle

Clément Labadie and Christine Legner

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

Published in Enterprise Modelling and Information Systems Architecture – International Journal of Conceptual Modeling (EMISAJ), 2020

Abstract. Personal data is increasingly positioned as a valuable asset. While individuals generate and expose ever-expanding volumes of personal information online, certain tech companies have built their business models on the personal data they gather. In this context, lawmakers are revising data protection regulations in order to provide individuals with enhanced rights and set new rules regarding the way corporations collect, manage, and share personal information. We argue that recent data protection regulatory frameworks such as the European Union's General Data Protection Regulation (EU-GDPR) or the California Consumer Privacy Act (US-CCPA) are fundamentally about data management. Yet, there have been no attempts to analyze the regulations in terms of their implications on the data life cycle. In this paper, we systematically analyze EU-GDPR and US-CCPA, and identify their implications on the data life cycle. To synthesize our findings, we propose a semi-formal notation of the resulting changes on the personal data life cycle, in the form of a process and data model governed by business rules, consolidated in a reference personal data life cycle model for data protection. To the best of our knowledge, this study represents one of the first attempts to provide a data-centric view on data protection regulatory requirements.

Keywords: data life cycle, data protection, personal data, regulatory compliance

Table of contents

List of figures

List of tables

1 Introduction

The idea of a right to privacy is not a novel one in the 19th century, the attorney Samuel Warren and the lawyer Louis Brandeis described a “right to be left alone” (Warren and Brandeis 1890). Organizations’ ever-enhancing ability to acquire and process personal data makes this increasingly relevant in our current reality. Through customer relationship management (CRM), personal data has become of strategic relevance for enterprises to improve interactions with their customers and create mutual benefits (Payne and Frow 2005). As individuals generate and expose ever-expanding volumes of personal information online, “digital native” enterprises assemble individualized profiles to target consumers and deliver personalized content and services. In fact, personal information processing is the very foundation of some of the last decade’s most successful corporations. From a privacy perspective, this leads to a redefined threat landscape. When it comes to data, the idea of misuse traditionally refers to security concepts and expresses the risk of unauthorized access, meaning that a malevolent, external party might access data. The increasing scope of personal data processing has also enlightened a new threat: that of “unintended inferences” (Burt 2019), which occurs when a rightful custodian of personal data uses it for unauthorized purposes.

Addressing this threat is the objective of all data protection regulations, and it is no surprise that they started appearing in Europe in the early 1980s, following the widespread adoption of information systems in enterprises (Hirschheim and Klein 2012). Having been introduced before the democratization of the internet, these regulations needed to be substantially revisited to cope with the exploitation of personal information by certain tech companies. This was the motivation for major revisions of data protection regulations (Mitrou 2017; Nicolaidou and Georgiades 2017), such as the European Union’s General Data Protection Regulation (EU-GDPR – (European Parliament and Council of the European Union 2016)), and the State of California’s Consumer Privacy Protection Act (US-CCPA – (California State Senate 2018)).

While these regulations aim to impose restrictions on corporate behaviors, they are fundamentally about data management, bearing technical as well as organizational impact for data management organizations (Hakim et al. 2018). They introduce data-related rights and data transparency requirements (both internal and external) that force organizations to substantially rework their data management practices. To comply with emerging data protection regulations, organizations must gain a precise overview of and change the way they manage personal data from beginning (gathering) to end (archiving or even deletion).

Over the years, in research as well as practice, data-centric life cycle models have been developed with this objective in mind, of which the model built by (Levitin and Redman 1993) was one of the first. These models describe all necessary steps to manage data elements form start to finish. They stem from a variety of domains and address diverse data types (e.g. product data, scientific/research data), but very few are applicable to personal data. When such models consider privacy aspects at all, the information is usually derived from non-legal definitions of privacy, and is not aligned with the precise legal requirements. Similarly, privacy research in IS has not focused on regulatory matters (Bélanger and Crossler 2011) and neither has customer relationship management or consumer research. Based on this observation, this paper addresses two research questions (RQs):

- RQ 1: What is the impact of data protection regulations on the personal data life cycle?
- RQ 2: How could data life cycle models be amended in order to address regulatory requirements for data protection?

To address RQ 1, we analyze two recent data protection regulation frameworks (the EU-GDPR and the US-CCPA). We find that these requirements directly impact the way data objects are created, processed, and maintained. From our analysis, we propose a classification of legal requirements from data protection legislation and show how they impact the data life cycle stages.

As an answer to RQ 2, we propose a reference personal data life cycle model for data protection, which comprises a data life cycle notation for data protection, outlining how general data management activities and steps are impacted by the aforementioned regulations. The notation is complemented by data model extensions to capture compliance-relevant attributes, as well as business rules to operationalize the life cycle process.

We start the detailed content by presenting perspectives on the regulatory context and the notion of personal data and reviewing existing research related to data protection and the data life cycle. We then outline our research methodology and process. Finally, we present a classification of legal requirements and derive a data life cycle notation with process and data models, as well as related business rules. We conclude with a summary and outlook on future research.

2 Background

2.1 Data protection regulatory landscape

Since May 25, 2018, the EU-GDPR directly applies to every European Union (EU) member state (Art. 99), repealing the preceding Data Protection Directive (95/56/EC, Art. 94). It addresses the need to remedy the fragmented implementations of the Data Protection Directive and accounts for the significant changes introduced by the mainstream adoption of the internet and the digital transformation (Mitrou 2017; Nicolaidou and Georgiades 2017). Any organization that processes EU citizens' personal data must comply with it, regardless of its geographical location. Violations are punishable by substantially higher fines (up to 20 million euros or 4% of an organization's global revenue, when previous regulations averaged about 500,000 euros). The EU-GDPR constitutes a landmark regulation for data protection in the EU and similar regulations are being introduced in other parts of the world. In Europe, Switzerland is currently undergoing an overhaul of its data protection legal framework – after several delays, it is set to be enforced in the beginning of 2022 and is expected to incorporate the measures the EU-GDPR (Métille and Raedler 2017) introduced. In 2017, China introduced its cyber security legislation, which covers data protection aspects such as personal information protection and rules for transnational data transmission. In 2018, following a supreme court judgment that declared privacy a fundamental right, India introduced a draft for a Personal Data Protection Bill (Parliament of the Republic of India 2018), with the objective of acting as a reference template for developing countries to introduce similar regulations (Palanisamy and Nandle 2018). The United States of America still does not have a single, general data protection regulation. Instead, several sector-specific laws co-exist, such as the Children's Online Privacy Protection Rule, the Federal Privacy Act (which only applies to federal agencies), and HIPAA (introduced in 1996, it contains requirements similar to the EU-GDPR 's, but is restricted to health-related data). Since the Facebook-Cambridge Analytica data scandal of 2018, there have been calls for a federal EU-GDPR-inspired data protection regulation (Rubio 2019). So far, only the state of California has passed its own EU-GDPR-inspired data protection law (California State Senate 2018), which became effective on January 1, 2020.

Although these regulations originate from different legislative bodies, they all address the same issues, and some are directly inspired by the EU-GDPR. Therefore, even if their requirements are positioned at differing levels of severity, the underlying concepts (such as personal data, data processing, consent, organizational and technical measures, and processes) remain the same, allowing for comparisons. Most importantly, existing transparency mandates have been strengthened. Organizations must now inform individuals about data processing in clear language and separately from general conditions, at the point of data collection. This means that organizations must define processing purposes for collected data elements before they gather such data. In the EU-GDPR, they are additionally required to present granular consent options as opt-in for non-mandatory processing activities. Both the EU-GDPR and the US-CCPA introduce the concept of accountability, which prompts organizations to be able to demonstrate compliance with the regulation. As they process data, they must also operationalize data rights, referring to access, rectification and restriction. When processing is no longer necessary or desired, individuals may request that their data records be deleted from enterprise systems.

2.2 Defining personal data

From a regulatory perspective, personal data can be defined as "data enabling direct or indirect identification of a single physical person, data that is specific to a single physical person without enabling identification, data that can be linked to a physical person, data regarding which anonymization techniques cannot completely mitigate the risk of re-identification" (Debet et al. 2015). In practice, most companies collect personal data about their customers, and it is often referred to as consumer or customer data. In that regard, it can be defined as "a set of data that represents and is associated with the identity, activities and service offering associated with a unique individual" (Tapsell et al. 2018). The aspect of service offering is prevalent in the consumer/customer data literature, and has been emphasized in the broader customer relationship management (CRM) field. In CRM, customer data is considered as an opportunity to understand the customer and co-create customer value (Payne and Frow 2005). The related contributions focus on collecting, organizing, and using customer data in order to build longterm relationships with customers (Saarijärvi et al. 2015). In this study, we consider personal data as data that contains personally identifiable information, meaning that it identifies a specific individual and/or provides information about them.

2.3 Data lifecycle management

In order to reflect the changes in data management practices induced by recent data protection regulation frameworks, this study uses the data life cycle as a frame of reference. On a high level of abstraction, "the life cycle of something [. . .] is the series of developments that take place in it from its beginning until the end of its usefulness" (Collins English Dictionary n.d.). The life cycle concept has been applied to various data-related domains (e.g. product data, scientific/research data) and has enjoyed a renewed interest in the context of big and open data landscapes. Four overview studies provide a comprehensive analysis and synthesis of the data life cycle and will be summarized in the next paragraphs. Out of the multitude of data life cycle models covered, we have identified only one that specifically deals with personal data.

(Möller 2013) conducted an extensive metaanalysis of life cycle models to derive the Abstract Data Lifecycle Model (ADLC) for the semantic web. He reviewed life cycle models from media production, e-learning, digital libraries, knowledge and content management and databases – the last two are the ones closest to our research field. In the database domain, the data life cycle is often associated with four basic operations of persistent storage known as CRUD (create, read, update, and delete - (Möller 2013)). In the knowledge and content management domain, which represents the largest subset in (Möller 2013)'s study, seven models outline the steps that enable organizations to capture implicit knowledge, structure it in a way that fits the need of the target audience, and maintain it as it evolves. These models put an emphasis on ontology development (Staab et al. 2001), roles, processes and tools for metadata generation (Greenberg 2003), web content management systems (McKeever 2003), digital curation (Higgins 2008) and semantic applications (Modritscher 2009), among others. They put an emphasis on data creation/authoring, distribution, maintenance, and preservation, but do not specifically target personal data. These steps, especially the latter, are not highly relevant with regards to personal data, in the sense that the data is generally collected "as is" and is not the result of a dedicated creation/authorship process. Furthermore, the preservation aspect contradicts legal requirements that emphasize data deletion.

In the same year, (Ofner et al. 2013) proposed a framework for data life cycle models in the context of master data management. Although the study approaches the topic from a product data point of view, the authors surveyed general life cycle models in the master data domain. One of them (Levitin and Redman 1993) puts the data life cycle in three main activity clusters: the acquisition cycle, the usage cycle, and assessment activities that intervene in both cycles, and include data deletion. This perspective is aligned with data protection, and the argument can be made that all life cycle models, regardless of the domain, can be described according to this structure. This also holds true for the general steps the professional association (DAMA International 2009) outlined which additionally suggest that "when effectively managed, the data lifecycle begins even before data acquisition, with enterprise planning for data, specification of data, and enablement of data capture, delivery, storage, and controls." This perspective is in line with informational duties prescribed by data protection regulations.

The studies by (Sinaeepourfard, Garcia, et al. 2016; Sinaeepourfard, Masip-Bruin, et al. 2016) present a metaanalysis of 17 data life cycle models. They stem from a variety of domains and there is a significant overlap with those (Möller 2013) and (Ofner et al. 2013) analyzed. According to (Sinaeepourfard, Garcia, et al. 2016), the observed large number and topical variety of data life cycle models can be explained by the fact that they are meant to address the specific requirements of a particular field, which is not aligned to the authors' goal of establishing a "scenario-agnostic" model.

Among this abundant literature, we found only one data life cycle model that specifically addresses personal data management (Alshammari and Simpson 2018). It is based on the ADLC model (Möller 2013) proposedand usesthe Global Privacy Standard as reference to incorporate privacy by design aspects into the data life cycle. Although it specifically mentions the EU-GDPR, the authors approach the topic through a wider set of principles to prevent limiting the scope of their model to a specific regulatory framework. The study elaborates on the various roles involved in the life cycle stages and describes the associated activities and dependencies in terms of input and output. It also explains the steps through a concrete case study. This contribution is much closer to our research objective, although it is not meant to express regulatory requirements.

2.4 Research motivation

We can summarize the literature as follows. First, we observe an increasing number of data protection regulations that build on similar concepts and seek to extend data protection requirements toward increased transparency and control for individuals. Implementing these requirements prompts companies to revise data management practices.

Second, prior research on personal data management mostly focuses on customer/consumer data and does so either from a CRM perspective, or investigates the non-legal aspects of privacy.

Third, the data life cycle research domain is a prolific one, and comprises a large number of domain-specific contributions, as well as a few attempting to generalize life cycle concepts. Among these contributions, only one data life cycle model addresses the topic of personal data. However, even though it mentions the EU-GDPR as exemplary motivation, it does not formally integrate a regulatory compliance point of view.

To address this gap in the literature, our study contributes a regulation-focused and data-centric approach to personal data management by analyzing and expressing data protection regulatory requirements in the data life cycle, and proposing data objects, attributes, and business rules to operationalize data life cycle steps. We adopt an end-to-end lens on the data life cycle, with a starting point prior to data collection. In that sense, our approach constitutes an answer to the call for data protection by design and by default formulated in the EU-GDPR (Art. 25).

Our data-centric focus means that we have excluded other aspects of data protection regulations. We do not cover organizational requirements such as appointing a data protection officer or adopting a code of conduct. We also do not address information security requirements, which are related to a different research domain, and are generally addressed separately in practice.

3 Research approach

In order to develop the data life cycle model in a rigorous scientific research process, we follow the established design science research guidelines (Hevner et al. 2004) and the methodological steps (Peffers et al. 2007) suggested. As depicted in Figure 13, our research process comprised three design iterations and involved four focus group meetings. These focus groups were held with more than 25 data management experts from 20 multinational organizations. Each focus group lasted approximately two hours and focused on either problem identification (Focus groups 1 and 2) or evaluating different versions of the artifact (Focus groups 3 and 4).

During the first step, we analyzed the implementation challenges induced by data protection regulations. For this purpose, we analyzed the complete EU-GDPR regulations, based on foundational data protection principles. Our primary sources consisted of legal textbooks (Bensoussan et al. 2018; Debet et al. 2015; Meier 2011; Voigt and Von Dem Bussche 2017). We then discussed the EU-GDPR's requirements with experienced practitioners in Focus groups 1 and 2, and collected questions as well information about implementation challenges or difficulties in their organizations. This resulted in the observation that regulatory requirements are formulated in a way that does not immediately translate in data management terms. As a result, practitioners were unsure of the impact of such requirements on their activities.

During the second step, we determined the objectives of our study. To address the gaps identified in the first phase, we set out to develop a data life cycle model for the specific purpose of representing data protection regulatory requirements.

The third step consisted of the first design cycle. To develop the data life cycle model, we analyzed legal literature that is focused on the EU-GDPR in order to derive key principles that underpin the emerging regulations. We then analyzed the EU-GDPR according to these principles and extracted requirements that impact data management. For this purpose, we looked into EU-GDPR-specific literature (De Hert and Papakonstantinou 2012, 2016; Guadamuz 2017; Mitrou 2017; Nicolaidou and Georgiades 2017; Tikkinen-Piri et al. 2018) as well as guidelines from official authorities (Commission Nationale de l'Informatique et des Libertés n.d.; European Data Protection Board 2018a, 2018b).

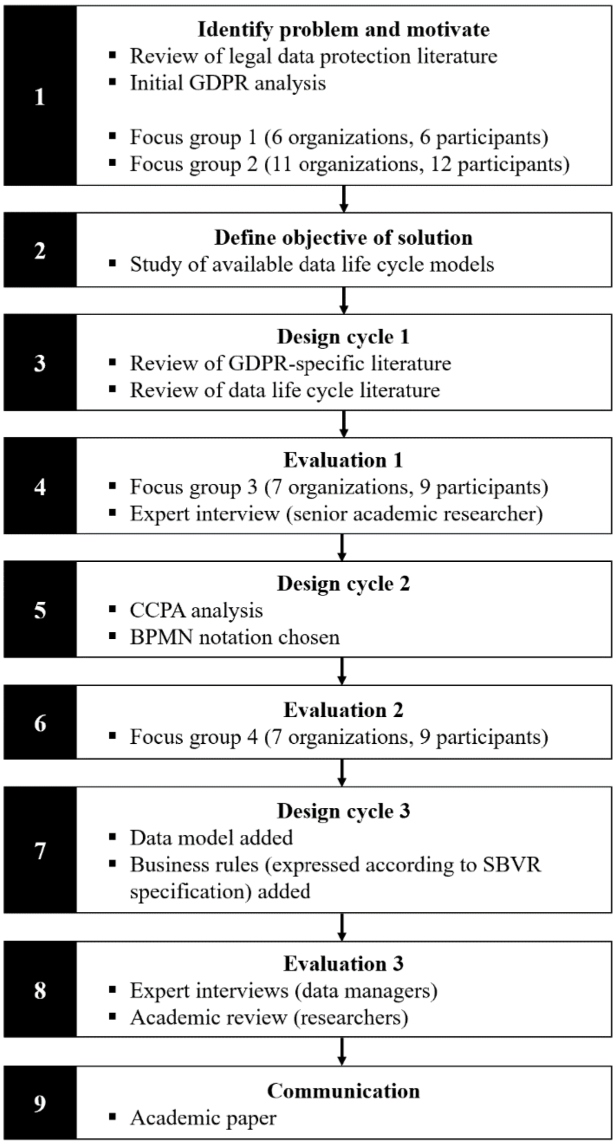

1
Identify problem and motivate
▪ Review of legal data protection literature
▪ Initial GDPR analysis
▪ Focus group 1 (6 organizations, 6 participants)
▪ Focus group 2 (11 organizations, 12 participants)
2
Define objective of solution
▪ Study of available data life cycle models
3
Design cycle 1
▪ Review of GDPR-specific literature
▪ Review of data life cycle literature
4
Evaluation 1
▪ Focus group 3 (7 organizations, 9 participants)
▪ Expert interview (senior academic researcher)
5
Design cycle 2
▪ CCPA analysis
▪ BPMN notation chosen
6
Evaluation 2
▪ Focus group 4 (7 organizations, 9 participants)
7
Design cycle 3
▪ Data model added
▪ Business rules (expressed according to SBVR specification) added
8
Evaluation 3
▪ Expert interviews (data managers)
▪ Academic review (researchers)
9
Communication
▪ Academic paper

Figure 13. Research process

We then conducted a literature review on data life cycle models. This enabled us to understand the typical articulation of data life cycle steps, and to confirm that existing models did not account for the specific requirements imposed by data protection regulations. Based on this review, we designed a first iteration of our data life cycle model, which extended existing models by amending and adding steps.

In a fourth step, our proposed model was evaluated through the third focus group with nine practitioners from seven organizations. Participants confirmed the general structure and commented on the consistency of the model (specifically, the order in which to position one of the newly designed steps), as well as on its level of detail. We also gathered concrete difficulties and roadblocks participants encountered along each life cycle step.

The fifth step consisted of the second design cycle, and we included the US-CCPA in our analysis of legal requirements in the light of the foundational principles identified during the first step. In parallel, we proceeded to rework our data life cycle model, based on the feedback. At this stage, BPMN was chosen as a reference notation, and the data life cycle reference model was redesigned accordingly.

In a sixth step, the redesigned data life cycle model was evaluated by means of an expert interview with a senior information systems researcher with knowledge of privacy and compliance topics. Additionally, we used a questionnaire distributed during Focus group 4 (with nine participants from seven organizations). We selected relevant evaluation criteria as described by (Prat et al. 2015) with regards to the model's structure (criteria: fidelity to modeled phenomenon, simplicity, completeness, and consistency) and fit to the target audience's environment (criteria: usefulness and ease of use). Our expert, as well as focus group participants, agreed that the proposed model was valid (accurately depicting data protection legal requirements), had an appropriate level of detail, consistent steps, and was easy to understand. In fact, these dimensions received evaluation scores of 4 and above (with 2 holdouts for the fidelity criterion). However, we observed lower scores with regards to the simplicity and usefulness of the model, with a consensus at around 3 out of 5, and one 2 out of 5. In their comments, participants noted that the model did not provide sufficient guidance about how to handle these steps in practice, especially on a technical level. Examples of remarks from the evaluation questionnaire included the following:

- "Detailed model is more tangible and sophisticated."
- "Provide a pragmatic proposal – how to manage it in reality."
- "Elaborate on details, e.g. rules, technical handling, etc."
- "Propose a business data model with the whole meta-data management in it, including end-oflife information."

To alleviate these concerns, the seventh phase consisted of an additional design cycle to enrich the data life cycle model with a simple, compliance-oriented data model, containing data elements (objects and attributes) that need to be recorded in order to operationalize the proposed steps on a technical level. The data model is supported by business rules that were designed following the Semantics of Business Vocabulary and Business Rules (SBVR) standard. A set of structural business rules specify the content of data objects in the data model, and a set of operational business rules steer the life cycle process and specify data operations throughout process steps, using the CRUD set of operations.

The eighth step consisted of an evaluation of the data model and business rules, by means of three expert interviews with representatives of two multinational organizations, as well as researcher feedback from the academic review. At this stage, we sought to guarantee the understandability, completeness, consistency, and effectiveness of the models and business rules (Prat et al. 2015). In addition, we evaluated the adequacy, usefulness, and applicability of the overall approach, referring to the combination of the life cycle model, data model, and business rules taken together. These evaluations were carried out as semi-structured interviews, and questions were evaluated based on a 5-level Likert scale. The experts consisted of a data architect and a data community manager from an organization active in the life-sciences industry, as well as a data architect from an organization active in the fashion & jewelry industry. All of them have over ten years of experience in the data management domain. The experts viewed the overall approach, data model, and business rules positively - their average rating of each dimension was above 3 (5 indicating full agreement) for all criteria except one.

Specifically, the data life cycle model's (s. Figure 15) ability to show the impact of data protection regulations on data management activities was rated with 4 out of 5 by all experts. The understandability, completeness, consistency and efficacy of the data model and business rules in the onboarding and usage phase were all rated with a minimum of 4 out of 5. Regarding the business rules for the end-of-life phase, two of the experts rated all criteria with a minimum of 4. One of them agreed that they were understandable and consistent (4 out of 5), but questioned their completeness and, as a result, efficacy. We made minor adjustments to the model to account for this feedback.

Following comments from academic reviewers, we also amended the data life cycle model to better reflect the EU-GDPR's right to restriction (art. 18) and breach notification requirement (art. 33). In order to maintain consistency with the updated data life cycle model, attributes registering the provenance of personal data and its recipients for a given processing purpose were added to the data model.

This paper constitutes the ninth and final step.

4 Integrating data protection requirements in the data life cycle

4.1 Data life cycle for personal data

In order to analyze the way data protection requirements impact the data life cycle, we started by synthesizing the steps described by existing data life cycle models. To that end, based on the literature review presented in Sect. 2, we selected those with a connecting link to personal data management. We therefore included the abstract data lifecycle model (Möller 2013) suggested, as it synthesizes and generalizes several other existing models. The model (Alshammari and Simpson 2018) proposed is included as well, since it is derived from (Möller 2013), and is the only one that specifically focuses on personal data. Because personal data manifests itself in the consumer / customer data domains in organizations, we also included data life cycle models related to master data management (DAMA International 2009; Levitin and Redman 1993).

Most models, except for (Levitin and Redman 1993), start with a planning phase, prior to data acquisition. It serves various purposes – for (Möller 2013), it defines the intent for creating the data and the internal requirements, such as data ownership, that will apply to the data post-collection. (Alshammari and Simpson 2018) phrase this intent in terms of the planned use of the data, referring to the purpose of data processing, while (DAMA International 2009) frames it as a preparatory phase to ensure proper alignment with an organization's system design processes.

All models then describe the step of collecting data and bringing it into an organization's system, which is referred to as creation, collection, obtaining values, acquisition, or publication. By mentioning the CRUD set of operations, (Möller 2013) suggests that these steps might be broken down further, in which case a first step would consist of acquiring and collecting the data. A second step would be translating it into an organization's data structure and making it consistently available in its systems. Although not explicitly stated, a similar inference can be drawn from (Alshammari and Simpson 2018), who mention a conceptual modeling step at the very beginning, prior to the planning phase, in order to specify the required data, the purposes for which personal data is to be processed, and the logical and physical data models.

The next step revolves around the usage of data and accompanying activities. Here, (Möller 2013) leans towards knowledge management / sharing and introduces steps closely related to getting feedback from internal as well as external users, while (Levitin and Redman 1993) emphasize quality control and generating related results. At this stage, (Alshammari and Simpson 2018) distinguish access and usage, but also outline privacy-related steps, such as retention and review/disclosure. (DAMA International 2009) centers on data maintenance.

All models also address the end-of-life of data, and comprise a step that corresponds to their removal from processing systems. In that same phase, before removal, (Möller 2013) and (DAMA International 2009) introduce an archiving step during which data can still be retrieved, while (Levitin and Redman 1993) focus on evaluation activities prior to removal.

Table 36. Data life cycle stages and steps

	Steps		
	Onboarding	**Usage**	**End-of-life**
(Möller 2013)	Ontology development, planning, creation	Publication, refinement, access, external use, feedback	Archiving, termination
(Levitin and Redman 1993)	Define view, implement view, obtain values, update records	Define subview, retrieve, manipulate, present results, use, assessment, analysis, adjust, update records	Obtain values, assessment, analysis, discard
(DAMA International 2009)	Plan, specify, enable, create and acquire	Maintain and use	Archive and retrieve, purge
(Alshammari and Simpson 2018)	Conceptual modeling, initiation, collection	Retention, access, usage, review, disclosure	Destruction

As synthesis, and in order to clearly structure the remainder of this study, we can derive three main stages showing through the aforementioned data life cycle models: onboarding (comprising the planning and collection/creation of data), usage (comprising all steps to be performed as data is stored and processed in systems), and end-of-life (comprising archival and deletion). Table 36 presents a summary of data life cycle steps as they appear in the models, classified according to these three stages.

4.2 Data-centric legal requirements

As per our research process, we started by investigating requirements from the EU-GDPR to analyze how data protection regulations impact the data life cycle. We formulated the assumption that principles underpinning these requirements are representative of data protection to a broader extent, and would apply to other data protection regulations. We verified this assumption by integrating the US-CCPA in our analysis in a second step.

In analyzing the EU-GDPR, we excluded the first chapter, which contains definitions and defines the material and territorial scopes of application. Chapters II and III, which contain principles and data rights, were included, as well as the first section of Chapter IV, which indicates organizations' data processing duties. The following sections of Chapter IV were excluded, as they cover security aspects (which we purposefully excluded from our study) and organizational aspects (such as impact assessments, data protection officers, codes of conduct, and certifications). The remaining chapters were not considered, as they deal with legal and judicial aspects that have no impact on data management activities.

With the selected chapters, further legal dispositions were set aside. Art. 10 targets information processing related to criminal convictions, which is a specific case that only applies to legal authorities. Similarly, Art. 23 gives national authorities a possibility to enact stricter rules regarding specific processing cases, such as homeland security, defense, and enhanced protection of individuals, which are also prerogatives of legal authorities. Art. 11 relates to the scope of the EU-GDPR in that it confirms that the processing of data that does not require the identification of individuals falls outside the scope of the regulation. Art. 12 defines modalities according to which organizations are expected to interact with individual requests, for example in terms of responsiveness and clarity, and states that the communication of information regarding data processing should occur without financial retribution. Finally, Art. 31 simply states that organizations must collaborate with supervisory authorities upon request.

Table 37. Regulatory requirements throughout the data life cycle

Requirement	EU-GDPR	US-CCPA	Onboarding	Usage	End-of-life
Right of information	Art. 7, 13, 14	§1798.100	X	X	
Right of access	Art. 15, 18, 20	§1798.110, 1798.115		X	
Right of deletion	Art. 15, 17	§1798.105			X
Right of rectification	Art. 7, 16, 21	N/A		X	
Right of restriction	Art. 18	N/A		X	
Right of consent	Art. 7, 8, 22	§1798.120	X	X	
Accountability requirement: Documentation	Art. 19, 24-30	§1798.130		X	X
Accountability requirement: Authorization	Art. 5, 6, 9	§1798.130		X	

To assess the impact of the data protection regulations on data management practices along the data life cycle, we have synthesized the relevant requirements into six categories of rights, and two categories of accountability requirements. These findings are consistent with the analysis provided in the legal literature, regarding rights (Bensoussan et al. 2018, pp. 30–31, Voigt and Von Dem Bussche 2017, pp. 31–38) as well as accountability principles (Nicolaidou and Georgiades 2017, Bensoussan et al. 2018, p. 12, Voigt and Von Dem Bussche 2017, p. 44). Table 37 provides an overview of the coverage of each category in the EU-GDPR and the US-CCPA, and outlines the impacted data life cycle stage, according to the main stages derived in Sect. 4.1. We will present these categories in the following paragraphs, and map the US-CCPA 's requirements to each of them.

Right of information. These rights are related to the principle of transparency (European Data Protection Board 2018b), and require that data processing measures be clearly communicated (Nicolaidou and Georgiades 2017, Bensoussan et al. 2018, p. 17). Concretely, organizations must inform individuals about the data elements they collect and detail the purposes for which they will be used in a clearly identifiable and understandable manner. This applies at the time of data collection, as well as during the entire personal data life cycle (as long as the organization processes related data elements). In these latter cases, information rights are complemented by access rights. These rights are expressed in a similar manner in both the EU-GDPR and the US-CCPA.

Right of access. As mentioned in the previous paragraph, access rights are similar to those of information, but relate to the disclosure of information during the data use stage only. In that sense, individuals may request to access their data at any time, and organizations must communicate the related data records.

Right of rectification. In the EU-GDPR, a right of rectification complements the right of access (Bensoussan et al. 2018, p. 31), and enables individuals to request that organizations update the data related to them. While the right of access is also present in the US-CCPA, the right of rectification is not stated in the regulation.

Right of restriction. In the EU-GDPR, a right of rectification enables individuals to contest the processing of the data related to them by an organization (e.g., due to inaccurate data or unauthorized processing). Art. 19 EU-GDPR states that they can request that organizations stop processing the related data until the dispute is resolved. In this case, organizations must effectively "freeze" the processing of data related to the individual. Art. 19 EU-GDPR also states that potential third-party recipients of the related data must be informed of the restriction.

Right of consent. In the EU-GDPR, consent is a foundational principle (Bensoussan et al. 2018; European Data Protection Board 2018a; Voigt and Von Dem Bussche 2017) that requires organizations to collect explicit authorizations from individuals as opt-in. It applies when processing is not based on other available processing bases (such as contract, legitimate interest, or legal obligation), and goes beyond their scope. It should also be collected in case data about children is collected (Art. 8), and when automated decisions will be made based on the collected data (Art. 22). In the US-CCPA, the right of consent is also present, albeit with a restricted scope

– the regulation only enables individuals to opt out of selling their personal data (§1798.120). The right of consent applies in conjunction with the rights of information (at the point of data collection) and of access (during the usage stage).

Right of deletion. Both regulations provide individuals with a right to request that organizations delete personal data that relates to them. This right is not absolute, in the sense that organizations may need to keep said data, or at least parts of it, for other purposes. In the EU-GDPR, these purposes are clearly laid out and refer to the authorization accountability requirements (see below). For instance, organizations may be required to retain personal data in order to comply with other regulations. Once the deletion has occured, thirdparty recipients of the related data must also be notified (Art. 19 EU-GDPR). In its §1798.105, US-CCPA enumerates the situations in which organizations are authorized to retain personal data.

Authorization (accountability requirement). In the EU-GDPR, any type of data processing must satisfy one (or several) of the bases for processing specified in Art. 5, referring to explicit consent (which is required for all automated decision-making), contract execution, compliance with another regulation, safeguarding the individual's vital interests, performance of public interest tasks/exercise of official authority, and legitimate interests. The US-CCPA does not provide a specific list of processing bases, but §1798.105 nevertheless lists cases in which organizations are allowed to continue data processing, even if deletion has been requested. These cases are similar to the EU-GDPR 's list of processing bases, with an emphasis on fraud prevention and scientific research (which could be construed as legitimate interests), as well as enforcing free speech and other legal requirements.

Documentation (accountability requirement). The documentation requirement appears in both regulations. It stipulates that organizations must be able to demonstrate the lawfulness of their processing activities (authorization), as well as the fulfillment of the abovementioned rights (Nicolaidou and Georgiades 2017, Voigt and Von Dem Bussche 2017, p. 44). This is necessary, for instance, to enforce the right of access: organizations must have defined and recorded the base(s) and purpose(s) of processing in order to communicate them upon individual request. From a broader perspective, and in case of official inquiry, organizations must be able to demonstrate that they process data according to legal requirements, meaning that they systematically collect the necessary data to ensure proper enforcement.

5 Reference personal data life cycle model for data protection

In this section, we reconcile the data life cycle steps with the data-centric regulatory requirements that we have isolated. We start by introducing a data model for data protection and structural business rules, which concretize the accountability requirements outlined in Sect. 4.2. Then we introduce a process model to articulate the relationship between data life cycle steps and data rights, organized around three subviews corresponding to the life cycle stages previously identified (onboarding, usage, and end-of-life).

Taken together, the data model, the process model and the business rules form the overarching reference personal data life cycle model for data protection, which we introduce in the following sections.

The data life cycle model has been designed using a semi-formal notation approach, based on the Business Process Modeling Notation (BPMN). We argue that the data life cycle can be expressed as a process, with different steps that create, read, update, and delete data objects, in accordance with the CRUD set of data operations. We chose bpmn due to its popularity for process modeling in both academia and practice. It was also suggested by participants of Focus group 2, so as to make the model approachable.

Both models are complemented by business rules expressed using the semantics of business vocabulary and business rules (SBVR) specification, a standard from the Open Management Group (OMG). It has been designed specifically to formalize compliance rules, and SBVR rules are adequate to support and complement bpmn process models (Cheng et al. 2011; Kluza and Honkisz 2016; Mickeviciute et al. 2017; Skersys, Tutkute, and Butleris 2012; Skersys, Tutkute, Butleris, et al. 2012). In our case, they ensure that legal requirements are met, and support the alignment of the life cycle process with data management.

5.1 Data model for data protection

In Sect. 4.2, we established that data protection regulations express accountability requirements, according to which organizations must be in a position to demonstrate compliance with data protection regulations by making sure all personal data processing is authorized and documented. In order to reach this objective, we argue that organizations must define, collect, and maintain personal data objects and attributes, as described by our proposed data model for data protection (s. Figure 14).

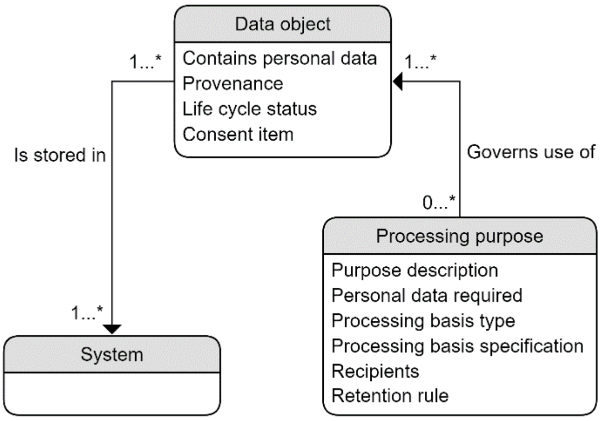


Figure 14. Personal data model for data protection

First, organizations must document the various purposes for which they process personal data, hence we suggest a first data object: *processing purpose*. Purpose documentation should be built around the information regarding the goal, authorization and type(s) of data required. To this end, we propose the following attributes:

- **`Purpose description`**: This should contain an explanation of the purpose at hand - to what end is personal data processed. For instance, an e-commerce retailer could outline that they need to collect personal data from their customers in order to process orders.
- **`Personal data required`**: This should list the personal data components that are required by the purpose, ideally in reference to data objects and / or attributes in the organization's data model. For instance, the e-commerce retailer would require a customer's name, address, birth date and credit card information in order to process orders.
- **`Processing basis type`**: This should indicate possible processing bases (e.g. a contract according to Art. 5 EU-GDPR) which must be specified to authorize data processing.
- **`Processing basis specification`**: This should describe the specific basis for the purpose at hand (e.g. details of a specific contract). In case the processing basis is consent, the specification should reflect the yes/no question that individuals will be asked for permission granting.
- **`Recipients`**: If the purpose entails transmitting data to third-party recipients, they should be specified here, so that organizations can notify such recipients in case restriction or deletion requests occur (Art. 19 EU-GDPR).
- **`Retention rule`**: If the purpose entails a specific retention rule (e.g. duration for keeping financial documents), it should be specified here.

For the *processing purpose* data object, all attributesmustberecorded, excepttheretention rule, as it is not mandatory to specify ending conditions for processing activities. Consequently, the following stuctural business rules apply to the *processing purpose* data object:

- It is necessary that *processing purpose* has purpose description and personal data required and processing basis type and processing basis specification.
- It is possible that *processing purpose* has **`recipients`**.
- It is possible that *processing purpose* has **`retention rule`**.
- It is necessary that processing purpose refers to personal data object.

When it comes to information recorded inside data objects, we suggest adding the following attributes to existing *data objects*:

- **`Contains personal data`**: This should be a Boolean value specifying whether a data object contains personal data.
- **`Provenance`**: This should specify whether a data object has been directly collected from the data subject themselves, or whether it was transmitted by a third-party. Organizations may also choose to introduce a more fine-grained classification. This would, for instance, enable organizations to clarify their data selling duties as per §1798.115d US-CCPA, which states that they cannot resell data that was itself sold to them without the data subject's agreement.
- **`Consent item`**: This should be a Boolean value, specifying whether an individual has opted in or out regarding consent-based processing purposes.
- **`Life cycle status`**: This should specify whether a data object is available for regular use or whether it has been marked for archival (e.g. in the case of a deletion request occurring while a processing purpose's retention rule is still ongoing) or restriction (in the case of a restriction request).

Consequently, the following business rules apply to the *data object* data object:

- It is necessary that *data object* has **`contains personal data`**.
- It is necessary that *data object* has **`provenance`**.
- It is necessary that *data object* has **`life cycle status`** if **`contains personal data`** is **`true`**.
- It is possible that *data object* has **`consent`** item if **`contains personal data`** is **`true`**.
- It is necessary that *data object* refers to *processing purpose* and *system* if **`contains personal data`** is **`true`**.

Finally, we suggest documenting *systems* of storage using a distinct data object. The ability to locate storage instances of data records is crucial for the disclosure and deletion activities, and has been cited as a significant difficulty in several of the focus groups we conducted. This aspect is confirmed by (Bensoussan et al. 2018, p. 23), highlighting the need for detailed mapping of collected data. (Peyret et al. 2017) points in the same direction.

5.2 Data life cycle model for data protection

In this section, we present the reference personal data life cycle model with its detailed subviews, corresponding to each of the data life cycle main stages (onboarding, usage, and end-of-life). Figure 15 depicts the data life cycle reference model, which comprises 12 steps. For each step, we specify CRUD operations that should be conducted on data objects and / or attributes in order to operationalize the regulatory rights and accountability requirements, as well as related operational business rules.

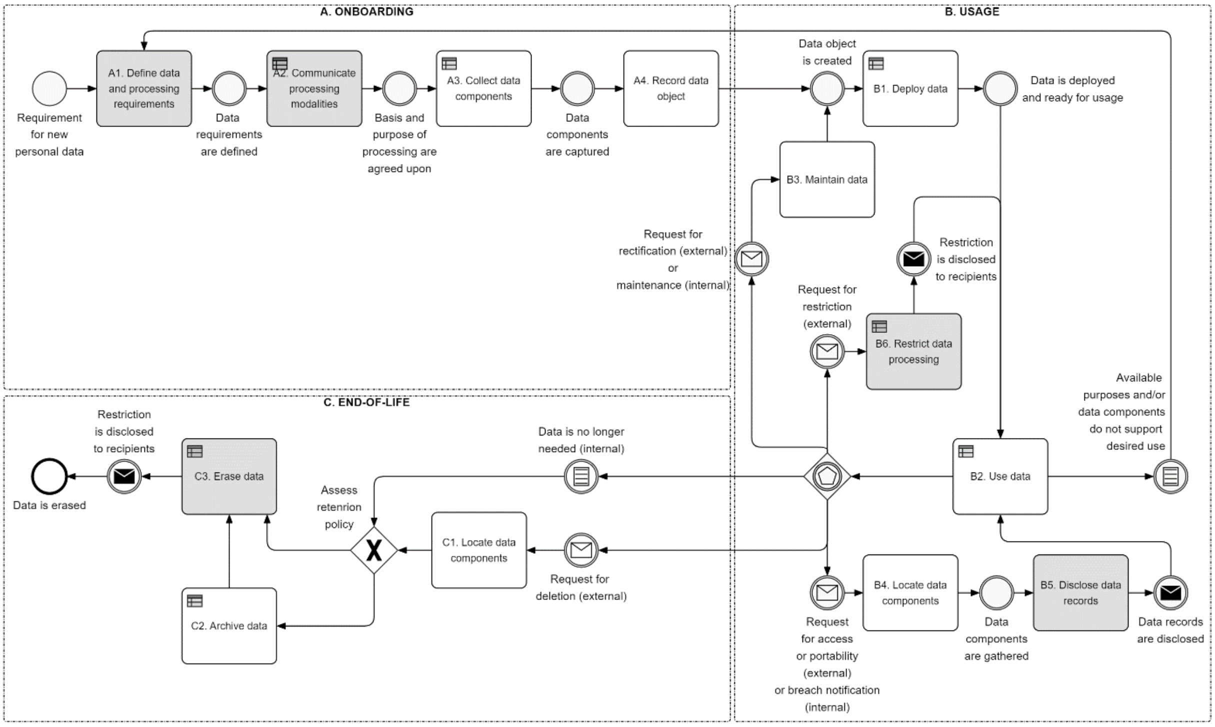

A. ONBOARDING
Requirement for new personal data
A1. Define data and processing requirements
Data requirements are defined
A2. Communicate processing modalities
Basis and purpose of processing are agreed upon
A3. Collect data components
Data components are captured
A4. Record data object
B. USAGE
Data object is created
B1. Deploy data
Data is deployed and ready for usage
B3. Maintain data
Request for rectification (external) or maintenance (internal)
Restriction is disclosed to recipients
Request for restriction (external)
B6. Restrict data processing
Available purposes and/or data components do not support desired use
B2. Use data
Request for access or portability (external) or breach notification (internal)
B4. Locate data components
Data components are gathered
B5. Disclose data records
Data records are disclosed
C. END-OF-LIFE
Restriction is disclosed to recipients
Data is erased
C3. Erase data
Data is no longer needed (internal)
Assess retenrion policy
C1. Locate data components
Request for deletion (external)
C2. Archive data

Figure 15. Reference personal data life cycle model for data protection[17]

5.2.1 Subview: onboarding

The entry point in the data life cycle is a requirement for new personal data (A1) - this step mirrors the planning step, as expressed in the models we analyzed in Sect. 4.1. It is related to the

[17] Steps highlighted in grey mirror data protection requirements and are additions to traditional data life cycle models.

right of information, as organizations must define and expose in advance the bases and purposes of processing related to the personal data they intend to collect. In the first step, organizations must define in advance the bases and purposes of processing for personal data, which is a key departure from previous data life cycle models. At this stage, the data in *processing purpose* is **`created`**.

The next three steps reflect activities necessary to bring data into an organization's system, and synthesize those described by the models we reviewed in Sect. 4.1. *Communicate processing modalities* and *collect data components* correspond to the data collection. *Record data object* corresponds to the moving of data into an organization's database/data model.

In the second step (A2), the processing basis and purposes should be displayed when collecting data from individuals (e.g. on a website). For that purpose, the data in *processing purpose* is **`read`**, and the following business rule applies, concretizing informational duties:

- *R1*: It is obligatory that *processing purpose* is disclosed.

In the third step (A3), data components are collected from the individual. It specifies which specific data components should be collected according to the processing purpose. The data in *processing purpose* is therefore **`read`**, and the following business rule apply:

- *R2*: It is obligatory that **`personal data required`** is collected.
- *R3*: It is obligatory that **`consent item`** is collected if **`processing basis type`** is **`consent`**.

In the fourth step (A4), data components are translated into the organization's structured data model. At this stage, the data in *data object* is created in the system.

5.2.2 Subview: usage

After the data is created, it must be deployed in the appropriate systems. *Deploy data* (B1) is derived from the publication step (Möller 2013) suggested, in the sense of making data available for usage. Participant feedback from focus groups indicated that organizations often operate with highly distributed system landscapes and need to deploy the data in several other systems that require it. For instance, in a multinational organization, data created in one ERP system may need to be deployed in other ERP systems (e.g. region-specific). Another example is data transfer in data lakes or other advanced analytics systems, separate from traditional enterprise systems. Therefore, recording target systems at the step of deployment could ease the creation of a system map for personal data processing, as (Bensoussan et al. 2018, p. 23) recommended.

To that effect, at this step (B1), the data in *system* is **read** and the *data object* is **updated**. Additionally, the following business rule applies:

- *R4*: It is obligatory that deployment *system* is referred to in *data object*.

Next, *use data* (B2) reflects the *processing purposes*, and was included in all analyzed models. At this stage, *processing purpose* and *data object* are **read**, and the following business rules applies:

- *R5*: It is obligatory that usage of *data object* conforms to *processing purpose*.

If the desired purpose cannot be fulfilled using available data objects and purposes, the life cycle process starts again at Step A1, as data and processing requirements need to be reviewed and extended, leading to a new instance of processing purpose. Following this, the new purpose must be communicated and/or additional data must be collected accordingly.

Contrary to (Alshammari and Simpson 2018), we did not include a retention step, and argue that retention is better described in terms of the **retention rule** attribute (defined in Sect. 5.1) than as a distinct step. The disclosure step (Alshammari and Simpson 2018) suggested represents the exchange of personal data between different organizations. We did not include a similar step in our model. From a regulatory standpoint, both the EU-GDPR and the US-CCPA stipulate that data exchanges must be announced as distinct processing purposes, and are therefore encapsulated in the use data (B2) step.

(Alshammari and Simpson 2018) suggest two steps for data communications – review and disclosure. Their review step originates from individuals and corresponds to the rights of access and rectification, which we have chosen to model as two distinct steps, namely *disclose data records* and *maintain data*. This articulation reflects the two rights individually, and that a disclosure does not necessarily prompt subsequent action from an individual. We also argue that a change in the data does not always originate from an individual request and can be triggered internally, for instance as part of data quality checks or routine data maintenance.

When maintaining data (B3), the data in *processing purpose* and *data object* is **updated**. At this point, relationships with certain instances of processing purpose may be removed (following a right of restriction request) or added (following authorization of further processing).

Following a right of access request, data must be located (B4) - for that purpose, the data in *data object* and system is **read**. In order to disclose data to individuals (B5), the data in *data object* and *processing purpose* is **read** and formatted for communication. Disclosure can also occur following a data breach. In this case, it is triggered internally when the breach is discovered, and the organization must communicate a list of compromised *data objects* (B4) to individuals (B5).

In the context of this study, we treat the right of portability as a variant of data disclosure, as it is also about communicating data records, with the added requirement of doing so in a standard, machine-readable format.

Following a right to restriction request under the EU-GDPR, organizations must stop processing the related data until further notice (B6), and inform third-party recipients of the restriction. At this point, *data object* and *processing purpose* are **`updated`**. We also suggest that organizations document a "restriction" processing purpose to effectively freeze data processing with the following business rule:

- *R6*: It is obligatory that **`life cycle status`** of *data object* is updated if *processing purpose* of *data object* is "restriction".

5.2.3 Subview: end-of-life

The end-of-life stage is triggered by a deletion request from the individual or by the ending of a predefined **`retention rule`**. Here again, related data objects must be located in the organization's system landscape (C1) before they can be archived or deleted. At this point, the data in *data object* and *system* is **`read`**.

As previously mentioned, when faced with a deletion request, it should be determined whether data can be erased, or whether it should be kept. This retention aspect has been cited as a significant difficulty during Focus groups 1 and 2, and participants mentioned that retention checks were not automated in their organizations.

Depending on the outcome of the retention check based on the **`retention period`** attribute, the related data elements would either be archived (C2) or removed (C3).

In the case of archival (C2), data in *processing purpose* is **`read`** and *data object* is **`updated`** (specifically, the **`life cycle status`** attribute). In addition, the following business rule applies:

- *R7*: It is obligatory that **`life cycle status`** of *data object* is changed if **`retention rule`** of *processing purpose* is valid.

In the case of erasure (C3), data in *processing purpose* is **`read`** and *data object* is **`deleted`**. In addition, the following business rule applies:

- *R8*: It is obligatory that *data object* is erased if `retention rule` of *processing purpose* is void.

By erasure, we mean both deletion and anonymization. As data is anonymized, meaning that all personally identifiable information is deleted, it no longer falls under the scope of data protection regulations, and thus exits the personal data life cycle.

Table 38 provides a condensed overview of CRUD operations carried-out throughout life cycle step.

Table 38. Overview of CRUD operations applied to data objects at eacg data life cycle step (related business rules mentioned when applicable)

	Create	Read	Update	Delete
A1. Define data and processing requirements	Processing purpose (R1)			
A2. Communicate processing modalities		Processing purpose		
A3. Collect data components		Processing purpose (R2, R3)		
A4. Record data object	Data object			
B1. Deploy data		System (R4)	Data object (R4)	
B2. Use data		Data object, Processing purpose (R5)		
B3. Maintain data		Processing purpose	Data object	
B4. Locate data components		Data object System		
B5. Disclose data records		Data object Processing purpose		
B6. Restrict data processing		Processing purpose (R6)	Data object (R6)	
C1. Locate data components		Data object System		
C2. Archive data		Processing purpose (R7)	Data object (R7)	
C3. Erase data		Processing purpose (R8)		Data object (R8)

6 Summary and discussion

The purpose of this study was to analyze the impact of data protection regulations from a data management perspective. To that end, we have investigated two distinct aspects of these regulations to develop a reference personal data life cycle model for data protection, which constitutes our main contribution.

First and foremost, data protection regulations grant a set of rights to individuals, designed to foster transparency about data processing, and clearly set the scope of data processing activities. The enablement of these rights translates into requirements for organizations, which we have represented using the concepts of life cycle process and business rules, in order to show how these requirements affect data management practices. This objective mirrors our first research question (RQ 1).

In addition to enabling data protection rights, organizations must be in a position to demonstrate that their processing of personal data has been lawful and authorized, and that it occurs within the contours of the regulatory rights and requirements. This obligation reflects the new principle of accountability, according to which organizations must document the compliance of their processing activities. To that effect, we have proposed a set of data objects and attributes that should be recorded along the steps of the data life cycle in order to provide a basis for such documentation. Furthermore, our model shows that such documentation (especially as it relates to processing purposes) should begin before any data is collected. In that sense, it matches the requirements for privacy by design and by default, particularly as formulated in the EU-GDPR, which states that measures should be implemented "both at the time of the determination of the means for processing and at the time of the processing itself" (Art. 25). This mirrors our second research question (RQ 2).

This study contributes to both research and practice. For research, it complements existing studies on the data life cycle by elaborating on the under-researched domain of personal data and by bringing in a regulatory perspective. By suggesting a semi-formal notation, we translate the emerging regulatory requirements into a set of rules. It also links up with related studies from the business process management domain. For instance, (Agostinelli et al. 2019) also use bpmn to model processes triggered by the exercise of individual rights (e.g. access, rectification), aiming to tackle them from process management perspective, which our study supplements by outlining data-related requirements to support compliance processes. Practitioners may benefit from the standardized notation of data protection issues to better understand data protection requirements, and identify potential blind spots in their operations.

7 Limitations and future research outlook

In this study, we purposefully bound our analysis of data protection regulations to the concept of the data life cycle. This informs organizations about critical steps that need to be addressed to tackle data protection requirements along the main stages of the data life cycle (onboarding, usage, end-of-life).

However, we only consider usage from the data provisioning perspective and did not analyze the actual data usage in detail. We argue that usage would be better described using concepts other than the data life cycle, such as data lineage, data flows or information supply chains. Such concepts would help analyze the information products and subsequent insights that can be derived from personal data, which would be useful, for instance, in the context of data protection impact assessments of data analytics activities. Similarly, we limited the suggested data attributes to the ones that strictly relate to legal requirements on an abstract level. A formal data model enriched with business rules could be developed for typical personal data objects, incorporating our suggested attributes. A classification of typical usage patterns could also be described in order to enhance the mapping of the retention policy to groups of data objects. Organizations would benefit from further describing the way data is used, for example in terms of roles, access control and permissions, and processes. In that regard, future research should make the link with responsibility definitions from the data governance domain, as well as with the abundant business process management literature stemming from the regulatory compliance management domain.

Acknowledgements

This research was supported by the Competence Center Corporate Data Quality (CC CDQ). The authors would like to thank the experts that contributed to this research.

References

Agostinelli, S., Maggi, F. M., Marrella, A., and Sapio, F. 2019. "Achieving GDPR Compliance of BPMN Process Models," in *Information Systems Engineering in Responsible Information Systems*, Lecture Notes in Business Information Processing, C. Cappiello and M. Ruiz (eds.), Cham: Springer International Publishing, pp. 10–22. (https://doi.org/10.1007/978-3-030-21297-1_2).

Alshammari, M., and Simpson, A. 2018. "Personal Data Management: An Abstract Personal Data Lifecycle Model," in *Business Process Management Workshops*, Lecture Notes in Business Information Processing, E. Teniente and M. Weidlich (eds.), Springer International Publishing, pp. 685–697.

Bélanger, F., and Crossler, R. E. 2011. "Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems," *MIS Quarterly* (35:4), pp. 1017–1042.

Bensoussan, A., Avignon, C., Bensoussan-Brulé, V., Forster, F., and Torres, C. 2018. *Règlement Européen Sur La Protection Des Données: Textes, Commentaires et Orientations Pratiques*, (2nd ed.), Brussels: Bruylant.

Burt, A. 2019. "Privacy and Cybersecurity Are Converging. Here's Why That Matters for People and for Companies.," *Harvard Business Review*. (https://hbr.org/2019/01/privacy-and-cybersecurity-are-converging-heres-why-that-matters-for-people-and-for-companies).

California State Senate. 2018. *California Consumer Privacy Act*.

Cheng, R., Sadiq, S., and Indulska, M. 2011. "Framework for Business Process and Rule Integration: A Case of BPMN and SBVR," in *Business Information Systems*, Lecture Notes in Business Information Processing, W. Abramowicz (ed.), Berlin, Heidelberg: Springer, pp. 13–24. (https://doi.org/10.1007/978-3-642-21863-7_2).

Collins English Dictionary. (n.d.). "Life Cycle Definition and Meaning," *Collins English Dictionary*. (https://www.collinsdictionary.com/dictionary/english/life-cycle, accessed May 24, 2019).

Commission Nationale de l'Informatique et des Libertés. (n.d.). "RGPD : passer à l'action." (https://www.cnil.fr/fr/rgpd-passer-a-laction, accessed September 13, 2018).

DAMA International. 2009. *The DAMA Guide to the Data Management Body of Knowledge*, (M. Mosley, M. Brackett, and S. Earley, eds.), Bradley Beach, NJ: Technics Publications.

De Hert, P., and Papakonstantinou, V. 2012. "The Proposed Data Protection Regulation Replacing Directive 95/46/EC: A Sound System for the Protection of Individuals," *Computer Law & Security Review* (28:2), pp. 130–142. (https://doi.org/10.1016/j.clsr.2012.01.011).

De Hert, P., and Papakonstantinou, V. 2016. "The New General Data Protection Regulation: Still a Sound System for the Protection of Individuals?," *Computer Law & Security Review* (32:2), pp. 179–194. (https://doi.org/10.1016/j.clsr.2016.02.006).

Debet, A., Massot, J., and Métallinos, N. 2015. *Informatique et Libertés: La Protection Des Données à Caractère Personnel En Droit Français et Européen*, Les Intégrales, Issy-les-Moulineaux: Lextenso.

European Data Protection Board. 2018a. "Guidelines On Consent Under Regulation 2016/679 (WP259, Rev.01)," Regulatory Guidelines, Regulatory Guidelines, European Union, October 4.

European Data Protection Board. 2018b. "Guidelines on Transparency under Regulation 2016/679 (WP260 Rev.01)," Regulatory Guidelines, Regulatory Guidelines, European Union, November 4.

European Parliament and Council of the European Union. 2016. *Regulation on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (Data Protection Directive).*

Greenberg, J. 2003. "Metadata Generation: Processes, People and Tools," *Bulletin of the American Society for Information Science and Technology* (29:2), pp. 16–19. (https://doi.org/10.1002/bult.269).

Guadamuz, A. 2017. "Developing a Right to Be Forgotten," in *EU Internet Law: Regulation and Enforcement*, T.-E. Synodinou, P. Jougleux, C. Markou, and T. Prastitou (eds.), Cham: Springer International Publishing, pp. 59–76. (https://doi.org/10.1007/978-3-319-64955-9_3).

Hakim, S., Li, Z., Pan, Y., Zaunseil, F., Chi, H., and Zhou, W. 2018. "The Impact of General Data Protection Regulation (GDPR) on Data Management Platforms (DMP): A Policy Perspective," *Management & Data Science*, , September 9. (https://management-datascience.org/2018/09/09/4457/, accessed May 27, 2019).

Higgins, S. 2008. "The DCC Curation Lifecycle Model," *International Journal of Digital Curation* (3:1), pp. 134–140.

Hirschheim, R., and Klein, H. K. 2012. "A Glorious and Not-So-Short History of the Information Systems Field," *Journal of the Association for Information Systems* (13:4), pp. 188–235.

Kluza, K., and Honkisz, K. 2016. "From SBVR to BPMN and DMN Models. Proposal of Translation from Rules to Process and Decision Models," in *Artificial Intelligence and Soft Computing*, Lecture Notes in Computer Science, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada (eds.), Cham: Springer International Publishing, pp. 453–462. (https://doi.org/10.1007/978-3-319-39384-1_39).

Levitin, A. V., and Redman, T. C. 1993. "A Model of the Data (Life) Cycles with Application to Quality," *Information and Software Technology* (35:4), pp. 217–223. (https://doi.org/10.1016/0950-5849(93)90069-F).

McKeever, S. 2003. "Understanding Web Content Management Systems: Evolution, Lifecycle and Market," *Industrial Management & Data Systems* (103:9), pp. 686–692. (https://doi.org/10.1108/02635570310506106).

Meier, P. 2011. *Protection Des Données: Fondements, Principes Généraux et Droit Privé*, Précis de droit Stämpfli, Bern: Stämpfli.

Métille, S., and Raedler, D. 2017. "Swiss Data Protection Act Reform Set in Motion," *Data Protection Leader* (14:2), pp. 14–16.

Mickeviciute, E., Butleris, R., Gudas, S., and Karciauskas, E. 2017. "Transforming BPMN 2.0 Business Process Model into SBVR Business Vocabulary and Rules," *Information Technology And Control* (46:3), pp. 360–371. (https://doi.org/10.5755/j01.itc.46.3.18520).

Mitrou, L. 2017. "The General Data Protection Regulation: A Law for the Digital Age?," in *EU Internet Law: Regulation and Enforcement*, T.-E. Synodinou, P. Jougleux, C. Markou, and T. Prastitou (eds.), Cham: Springer International Publishing, pp. 19–57. (https://doi.org/10.1007/978-3-319-64955-9_2).

Modritscher, F. 2009. "Semantic Lifecycles: Modelling, Application, Authoring, Mining, and Evaluation of Meaningful Data," *International Journal of Knowledge and Web Intelligence* (1:1/2), pp. 110–124. (https://doi.org/10.1504/IJKWI.2009.027928).

Möller, K. 2013. "Lifecycle Models of Data-Centric Systems and Domains: The Abstract Data Lifecycle Model," *Semant. Web* (4:1), pp. 67–88.

Nicolaidou, I. L., and Georgiades, C. 2017. "The GDPR: New Horizons," in *EU Internet Law: Regulation and Enforcement*, T.-E. Synodinou, P. Jougleux, C. Markou, and T. Prastitou (eds.), Cham: Springer International Publishing, pp. 3–18. (https://doi.org/10.1007/978-3-319-64955-9_1).

Ofner, M., Otto, B., Oesterle, H., and Straub, K. 2013. "Management of the Master Data Lifecycle: A Framework for Analysis," *Journal of Enterprise Information Management* (26:4), pp. 472–491. (https://doi.org/10.1108/JEIM-05-2013-0026).

Palanisamy, M., and Nandle, R. 2018. "Understanding India's Draft Data Protection Bill," , September 13. (https://iapp.org/news/a/understanding-indias-draft-data-protection-bill/, accessed January 31, 2019).

Parliament of the Republic of India. 2018. *The Personal Data Protection Bill*.

Payne, A., and Frow, P. 2005. "A Strategic Framework for Customer Relationship Management," *Journal of Marketing* (69:4), pp. 167–176. (https://doi.org/10.1509/jmkg.2005.69.4.167).

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).

Peyret, H., Cullen, A., McKinnon, C., Blissent, J., Iannopollo, E., Kramer, A., and Lynch, D. 2017. "Enhance Your Data Governance to Meet New Privacy Mandates," Consortium Report, Consortium Report, Forrester Research.

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).

Rubio, M. 2019. "To Impose Privacy Requirements on Providers of Internet Services Similar to the Requirements Imposed on Federal Agencies under the Privacy Act of 1974," No. S.142, Congress of the United States, January 16.

Saarijärvi, H., Karjaluoto, H., and Kuusela, H. 2015. "Customer Relationship Management: The Evolving Role of Customer Data," in *Marketing Dynamism & Sustainability: Things Change, Things Stay the Same…*, Developments in Marketing Science: Proceedings of the Academy of Marketing Science, Jr. Robinson Leroy (ed.), Springer International Publishing, pp. 505–515.

Sinaeepourfard, A., Garcia, J., Masip-Bruin, X., and Marín-Tordera, E. 2016. "A Comprehensive Scenario Agnostic Data LifeCycle Model for an Efficient Data Complexity Management," in *2016 IEEE 12th International Conference on E-Science (e-Science)*, , October, pp. 276–281. (https://doi.org/10.1109/eScience.2016.7870909).

Sinaeepourfard, A., Masip-Bruin, X., Garcia, J., and Marín-Tordera, E. 2016. "A Survey on Data Lifecycle Models: Discussions toward the 6Vs Challenges," UPC BarcelonaTech.

Skersys, T., Tutkute, L., and Butleris, R. 2012. "The Enrichment of BPMN Business Process Model with SBVR Business Vocabulary and Rules," in *Proceedings of the 34th International Conference* *on Information Technology Interfaces (ITI 2012)*, , June, pp. 65–72. (https://doi.org/10.2498/iti.2012.0366).

Skersys, T., Tutkute, L., Butleris, R., and Butkiene, R. 2012. "Extending BPMN Business Process Model with SBVR Business Vocabulary and Rules," *Information Technology And Control* (41:4), pp. 356–367. (https://doi.org/10.5755/j01.itc.41.4.2013).

Staab, S., Studer, R., Schnurr, H.-, and Sure, Y. 2001. "Knowledge Processes and Ontologies," *IEEE Intelligent Systems* (16:1), pp. 26–34. (https://doi.org/10.1109/5254.912382).

Tapsell, J., Akram, R. N., and Markantonakis, K. 2018. "Consumer Centric Data Control, Tracking and Transparency – A Position Paper," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, , August, pp. 1380–1385. (https://doi.org/10.1109/TrustCom/BigDataSE.2018.00191).

Tikkinen-Piri, C., Rohunen, A., and Markkula, J. 2018. "EU General Data Protection Regulation: Changes and Implications for Personal Data Collecting Companies," *Computer Law & Security Review* (34:1), pp. 134–153. (https://doi.org/10.1016/j.clsr.2017.05.015).

Voigt, P., and Von Dem Bussche, A. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*, Cham: Springer International Publishing.

Warren, S., and Brandeis, L. 1890. "The Right to Privacy," *Harvard Law Review*, pp. 193–220.