Ismael Caballero
Mario Piattini *Editors*

Data Governance

From the Fundamentals to Real Cases

Springer

Data Governance

Ismael Caballero • Mario Piattini
Editors

Data Governance

From the Fundamentals to Real Cases

Springer

Editors
Ismael Caballero
Alarcos Research Group
Institute of Technologies and Information Systems, University of Castilla-La Mancha (UCLM)
Ciudad Real, Spain

Mario Piattini
Alarcos Research Group
Institute of Technologies and Information Systems, University of Castilla-La Mancha (UCLM)
Ciudad Real, Spain

ISBN 978-3-031-43772-4 ISBN 978-3-031-43773-1 (eBook)
https://doi.org/10.1007/978-3-031-43773-1

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

The editors want to dedicate this book to all the DQTeam members and partners for their great outstanding in Data Governance and Quality.

To my parents, Juan and Agustina, for their love, their example of life, and their continuous support.
—Ismael Caballero

A Danilo Caivano e Maria Teresa Baldassarre per le loro qualità di ricerca e umane.
—Mario Piattini

Foreword by Yang Lee

On a crisp and bright autumn day in 2004 at MIT, Cambridge, Massachusetts, a young, tall, and serious PhD student was presenting a paper from his dissertation research, entitled, "Getting Better Information Quality and Improving Information quality Management." Almost two decades later, I received an email from the same student, now Professor Ismael Caballero, along with Professor Mario Piattini, inviting me to write a foreword for their co-edited book on data governance. I opened the file with much excitement and expectation of reading a great book.

I have known Dr. Caballero and Dr. Piattini for about 20 years. My encounters with them have been mainly through data-related conferences and discussions. I have first read Ismael's research paper as I was reviewing his paper with a few other reviewers and organizers for the International Conference on Information Quality (ICIQ, also known previously as MIT IQ conference), which I co-founded in 1996 and organized for several decades. The ICIQ team's agenda around 2004 was to nurture and grow the next generation researchers and practitioners. The core team of the ICIQ, Dr. Richard Wang, Dr. Stuart Madnick, and myself actively encouraged and supported new PhD students, academic researchers, and industry practitioners through creating various opportunities.

The incredible duo, Professor Caballero and Professor Piattini are obviously two path finders, creating and serving research leadership, and generous service contributions to the international academic and industry communities in the backdrop of the dramatic growth in data-related industry, particularly in the areas of data quality, data governance, data analytics, and AI.

Here is one such example scene at an ICIQ conference, after a full day of discussing data quality. Imagine a group of Spanish dancers for "Sevillanas on a Tablao Flamenco," with live music at a beautiful outdoor courtyard in Spain. Another important encounter with the duo was in 2016, when Ismael and Mario, along with the Alarcos Research Group support, hosted the ICIQ conference in Ciudad Real, Spain. The Spanish hospitality and impeccable organization by Professors Piattini, Caballero, and the entire team was much beyond my and participants' expectations.

Dr. Caballero and Dr. Piattini's book, in collaboration with many international experts, is a valuable and timely guide for studying and practicing data governance comprehensively, from frameworks to technologies, in the critical era of dramatic data growth, data quality management, data technology, analytics, security/privacy, and unforeseen data use in AI.

Specifically, in Chap. 7, the co-editors succinctly introduce the various maturity models for data governance, ranging from the DAMA model to IBM's, Gartner's, EDM's, MAMD's (Alarcos' Model), DMM's, Aiken's, and the DCAM model. Readers should be able to appreciate the potpourri of pointers from all models and utilize at least one or two models that best fit their values, purpose, and organizational and industry contexts. In addition, Chap. 7, Maturity Model, summarizes the chapters from Part One, the Fundamentals of Data Governance, which introduces multiple prescriptive frameworks, models, and methodologies, and invites readers to Part Two, providing multiple descriptive chapters of how data governance models are applied and implemented in the real-world industry with cases and exemplars, including in the public sector, and the banking, insurance, healthcare, and telecommunications industries.

Lessons learned along the way from implementing data governance models in various industries and organizations in Part Two should be particularly useful for many students, researchers, and practitioners of data governance in their own journey.

As the data governance area grows to include contemporary and future use of data, data management mechanisms, and related technology, this book should be a good guide to the readers who want to learn and implement current models and who want to create and explore future models, frameworks, and technology.

As I close this foreword, I am flipping through the photos of Spanish dancing and food and am looking forward to witnessing future endeavors and reading future research and practice by Ismael and Mario.

Congratulations to Dr. Ismael Caballero Muñoz-Reja and Dr. Mario Gerardo Piattini Velthuis on producing this valuable and timely book on data governance.

Cheers!

Yang Lee PhD

Northeastern University, Boston, MA, USA

University of São Paulo, São Paulo, Brazil

Foreword by Alberto Palomo

The dizzying process of digitalizing the global economy in recent years and the growing desire of private and public organizations to better exploit their data have produced exponential growth around data. In this sense, organizations want to benefit from this exponential growth to make their processes more efficient and innovative, providing new products and services. This digital explosion has clearly revealed the need to address the challenges posed by properly and efficiently managing information. Therefore, data management and governance have become critical for organizations due to their fundamental role in planning and programming their activity and, therefore, in decision-making.

In the era of big data, and as a premise before the transformation and exploitation of large data sets, it is clear that it is necessary to establish adequate planning for its governance and management to capitalize on its maximum value. A strategy that ensures the quality and security of the information and, in turn, allows its practical use is required. This strategy must provide coherence and efficient alignment between all the procedural areas in the data value chain, from its collection to its use, distribution, and, ultimately, its destruction.

A new paradigm has recently emerged with force in this adventure of maximizing data value. The main proposal of this new paradigm lies in generating utility beyond the ecosystem where it is created. The intention is to break down the silos created by data modeling itself, both in the definition and internal semantics of a specific set of data, having been adapted for a specific purpose and in the general architecture of the information systems on which it is based. Even within the same organization, it is common to find barriers and impediments hindering a more holistic data exploitation. There is an avid desire to create horizontal structures through which data can become a shared resource that, from different perspectives, can add value to the organization's strategy to mitigate this effect. Because data, far from being a matter of only ICT interest, has a cross-cutting potential that feeds all business areas.

Data life cycles in the public and private spheres are increasingly complex; they can follow nonlinear trajectories interrelated with each other without clear points of

governance, and they often even cross different areas or types of data. This vision of the data life cycles means that uncertainties accumulate. It is essential to address data governance on solid foundations, both from the regulatory and applied knowledge spheres, to avoid an adverse effect.

Thus, the European Data Strategy seeks to make the Union a leader in an innovative and digital society, where the development of a single market for data allows its free circulation, both geographically and between sectors, to benefit entrepreneurship and innovation, researchers, and public administrations. As a critical part of this document, common European dataspaces are postulated as guarantors of data available across the economy and society based on compliance with competitive frameworks and European digital sovereignty. However, even beyond the institutional impulse, the work developed from initiatives such as the Data Spaces Business Alliance, with permeability through their respective national and regional hubs, from academic institutions, and the governments of different Member States, has allowed the configuration of a common shared space for reflection and analysis with which to generate fertile ground for the emerging data economy and, ultimately, the digital single market.

This book, therefore, represents a pertinent contribution insofar as it offers relevant contributions to constructing a solid scientific corpus to clarify and pave the way for organizations interested in capitalizing on data. The chapters in the first part significantly enrich the creation of a conceptual framework for data governance, while the second part presents advances and concrete, practical experiences. In short, this is an enriching contribution regarding both approach and content, bringing us to state-of-the-art data governance. These considerations will undoubtedly guide all those who, in one way or another, work in this incipient and exciting field. They will allow us to continue advancing in opening new lines of knowledge and consolidating existing ones.

State Secretariat for Digitalization and Artificial Intelligence, Ministry of Economic Affairs and Digital Transformation, Madrid, Spain

Alberto Palomo

Preface

Overview

Data has always been a key element for the operation of organizations' information systems. However, in the last decade aspects such as digital transformation; the spread of technologies such as big data, analytics, and artificial intelligence; the increase of uncertainty and the necessary adaptability of business models; the growing regulatory and normative frameworks; and the necessary personalization and improvement in the provision of services have made data governance acquire a capital importance for the survival and profitability of companies and organizations.

In fact, data has become one of the most important strategic assets for organizations and is increasingly becoming a source of business innovation. It has even become in itself a product that must be managed and governed like any other product so that it can then be marketed and sold (e.g., in data markets), giving rise to the emergence of data ecosystems.

All this justifies that the data economy will be worth at least 550 billion euros by 2025 and that organizations are significantly increasing their budgets for data governance, management, and quality.

This book has been conceived with the objective, on the one hand, of bringing together a set of models, methods, and techniques that allow the successful implementation of data governance in an organization. And, on the other hand, to gather real experiences of data governance in different public and private sectors.

Organization

The book is composed of two parts.

Part I: Data Governance Fundamentals

The first part of the book begins with an enjoyable introduction to the concept of data governance (DG) by Peter Aiken, who stresses that DG is not primarily focused on databases, clouds, or other technologies, but that the DG framework must be understood identically by business users, systems personnel, and the systems themselves. This expert proposes proactive versus reactive DG and discusses the role of DG frameworks.

Dominik Lis, Joshua Gelhaar, and Boris Otto address in Chap. 2 crucial topics for data governance, such as the evolution of data management in organizations, data strategy and policies, and defensive and offensive approaches to data strategy. In addition, they discuss the emergence of data ecosystems and their use as part of data strategy and give recommendations for individual organizations as well as for the design of data ecosystems.

In Chap. 3, David Plotkin details the central role that human resources play in data governance, analyzing the Executive Steering Committee, Data Governance Board, Data Stewardship Council, and the Data Governance Program Office (DGPO). Also, the key roles and responsibilities for data stewards are described.

The value and monetization of data is addressed by Douglas Laney in Chap. 4, in which he discusses data management as a real asset and the most common barriers. In addition, drawing on GAAP, he proposes the Generally Agreed-Upon Information Pronciples (GAIP), as well as a new model for the data supply chain and the adaptation of the main existing data-related frameworks and standards.

Christine Legner, Martin Fadler, and Tobias Pentek summarize, in Chap. 5, the paradigm shifts in data governance, from control to value creation, presenting a reference model as a three-step approach towards data and analytics governance, which has been developed in an industry-research collaboration and tested with companies from different industries.

Chapter 6 by Kash Mehdi explores the needs and characteristics of data governance tools. It also illustrates through real cases the key functionalities needed in the data governance tools.

This first part ends with a chapter on maturity models for data governance by Ismael Caballero, Fernando Gualo, Moisés Rodríguez, and Mario Piattini. These authors provide an overview of the main models (DAMA, Aiken, IBM, Gartner, DCAM, etc.) and discuss in more detail the Alarcos' Model for Data Maturity (MAMD) based on the ISO/IEC 33000 and 8000-6x family of standards and its practical applications.

Part II: Data Governance Applied

The second part of the book reviews the situation of data governance in different sectors and industries. In Chap. 8, Raul Cruces Rufo analyzes the situation of data governance in the banking sector. He reviews the legislation and regulations affecting this sector and describes the vision of a data-driven bank, which comprises data stewardship, Single Data Marketplace ecosystem (SDM), DM&G dashboard, and Data as a Service (DaaS).

Chapter 9 is dedicated to data governance in public administration. Carlos Alonso Peña, Alberto Palomo, and Javier Esteve address two distinct but ultimately intertwined topics. On the one hand, it sets out the concepts and constraints underpinning federated data governance as a critical element in achieving strategic digital autonomy. On the other hand, the chapter details the principles that should govern a data-oriented administration to unlock the potential of data as an internal and external transformative power.

In Chap. 10, Juan Francisco Riesco discusses data governance in the insurance industry. He discusses the heterogeneous data governance strategies in the insurance sector, the different characteristics of data governance in this sector, and insurance trends and their impact on data governance.

Data governance and its implications in the healthcare sector is discussed in Chap. 11 by Alberto Freitas, Julio Souza, and Ismael Caballero. The authors also present a case study of a hospital in Portugal including a framework denominated CODE.CLINIC.

This part ends with a chapter dedicated to Data Governance in the Telecommunications Sector, by José Luis Sanzana and Eric Ancelovici, who summarize how a telecommunications company is structured at the functional level, the type of services it provides, how it deals with the avalanche of data it has to manage, how to structure the specialized areas to organize and govern the data, and, finally, some examples of problems.

Target Readership

The target readership for this book is assumed to have previous knowledge of information systems and databases. The book is aimed at academic, researchers, and practitioners involved in data governance.

As for practitioners, it is especially indicated for Data Governors, Chief Data Officers, Data Stewards, Chief Information Officers, Chief Digital Officers, Data Administrators, and Data Managers. It may also be useful for Audit and Compliance and Risk Officers as well as for Data Protection Officers.

It can also serve as a reference book for monographic courses on data governance, as well as for the subjects to be incorporated in the curricula of bachelor's and master's degree courses in the field of information systems.

Ciudad Real, Spain
June 2023

Ismael Caballero
Mario Piattini

Acknowledgments

We would like to express our gratitude to all those individuals and parties who helped us to produce this volume. First, we would like to thank all the contributing authors and reviewers who helped to improve the final version. Special thanks to Springer-Verlag and Ralf Gerstner for believing in us once again and for giving us the opportunity to publish this work.

We would also like to express our gratitude to Natalia Pinilla of Universidad de Castilla-La Mancha for her support during the production of this book. We would also like to thank Prof. Yang Lee (from the Northeastern University in the USA) and Dr. Alberto Palomo (Chief Data Officer of the Spanish Government) for agreeing to write forewords to this work.

Finally, we wish to acknowledge the support of the "ADAGIO (Alarcos' DAta Governance framework and systems generatIOn)" project funded by JCCM, Regional Ministry of Education, Culture and Sports and ERDF Funds (SBPLY/21/180501/000061), and the "AETHER (A holistic Smart data approach for context-driven data analysis with a focus on quality and safety)" project funded by the Ministry of Science, Innovation and Universities ERDF Funds (PID2020-112540RB-C42).

Contents

Part I Data Governance Fundamentals

1 **Introduction to Data Governance: A Bespoke Program Is Required for Success** 3
Peter Aiken

2 **Data Strategy and Policies: The Role of Data Governance in Data Ecosystems** 27
Dominik Lis, Joshua Gelhaar, and Boris Otto

3 **Human Resources Management and Data Governance Roles: Executive Sponsor, Data Governors, and Data Stewards** 57
David Plotkin

4 **Data Value and Monetizing Data** 75
Douglas Laney

5 **Data Governance Methodologies: The CC CDQ Reference Model for Data and Analytics Governance** 99
Christine Legner, Martin Fadler, and Tobias Pentek

6 **Data Governance Tools** 121
Kash Mehdi

7 **Maturity Models for Data Governance** 139
Ismael Caballero, Fernando Gualo, Moisés Rodríguez, and Mario Piattini

Part II Data Governance Applied

8 **Data Governance in the Banking Sector** 165
Raúl Cruces Rufo

9 Data Has the Power to Transform Society 179
Carlos Alonso Peña, Alberto Palomo Lozano,
and Javier Esteve Pradera

10 Data Governance in the Insurance Industry 199
Juan Francisco Riesco

11 Data Governance in the Health Sector 215
Alberto Freitas, Julio Souza, and Ismael Caballero

12 Data Governance in the Telco Sector 233
José Luis Sanzana

Contributors

Peter Aiken Virginia Commonwealth University/Data Blueprint, Richmond, VA, USA

Carlos Alonso Peña State Secretariat for Digitalization and Artificial Intelligence, Ministry of Economic Affairs and Digital Transformation, Madrid, Spain

Ismael Caballero DQTeam/Alarcos Research Group, University of Castilla-La Mancha (UCLM), Ciudad Real, Spain

Raúl Cruces Rufo Santander Bank, Madrid, Spain

Javier Esteve Pradera State Secretariat for Digitalization and Artificial Intelligence, Ministry of Economic Affairs and Digital Transformation, Madrid, Spain

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

Alberto Freitas Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS) / Center for Health Technology and Services Research (CINTESIS), Faculty of Medicine, University of Porto, Porto, Portugal

Joshua Gelhaar Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Fernando Gualo DQTeam / Alarcos Research Group, University of Castilla-La Mancha (UCLM), Ciudad Real, Spain

Douglas Laney Data & Analytics Strategy, West Monroe, Chicago, IL, USA

Yang Lee Northeastern University, Boston, MA, USA

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

Dominik Lis Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Alberto Palomo Lozano State Secretariat for Digitalization and Artificial Intelligence, Ministry of Economic Affairs and Digital Transformation, Madrid, Spain

Kash Mehdi DataGalaxy, Lyon, France

Boris Otto Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Tobias Pentek CDQ AG, St. Gallen, Switzerland

Mario Piattini Alarcos Research Group, University of Castilla-La Mancha (UCLM), Ciudad Real, Spain

David Plotkin Metadata Services at MUFG Union Bank, Walnut Creek, CA, USA

Juan Francisco Riesco Mutua Madrileña, Madrid, Spain

Moisés Rodríguez Alarcos Research Group, University of Castilla-La Mancha (UCLM), Ciudad Real, Spain

José Luis Sanzana Zurich-Santander, Santiago, Chile

Julio Souza Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS) / Center for Health Technology and Services Research (CINTESIS), Faculty of Medicine, University of Porto, Porto, Portugal

List of Abbreviations

AI	Artificial Intelligence
AP	Auxiliary Process
BAU	Business as Usual
BCBS	Basel Committee on Banking Supervision
CCPA	California Consumer Privacy Act
CDAO	Chief Data and Analytics Officer
CDE	Critical Data Element
CDG	Continua Design Guideline
CDMP	Certified Data Management Professional
CDO	Chief Data Officer
CEO	Chief Executive Officer
CFO	Chief Financial Officer
CIB	Corporate & Investment Banking
CIM	Computer-Integrated Manufacturing
CIO	Chief Information Officer
CMMI	Capability Maturity Model Integration
COBIT	Control Objectives for Information and Related Technology
CRO	Chief Revenue Officer
CRUD	Creating, Reading, Updating, and Deleting
DA	Data Analytics
DaaS	Data as a Service
DAMA	Data Management Association
DCAM	Data Capability Assessment Model
DG	Data Governance
DGPO	Data Governance Program Office
DICOM	Digital Imaging and Communications in Medicine
DIMV	Data Inner Monetary Value
DIP	Data Improvement Projects
DISA	Data and Information Self-Assessment
DIV	Data Inner Value

DM&G	Data Management & Governance
DMBOK	Data Management Body of Knowledge
DMM	Data Management Maturity
DNA	Deoxyribonucleic Acid
DoD	Definition of Done
DQ	Data Quality
DQI	Data Quality Indicator
ECB	European Central Bank
ECM	Enterprise Content Management
EDMC	Enterprise Data Management Council
EHR	Electronic Health Record
EIM	Enterprise Information Management
ERP	Enterprise Resource Planning
ESG	Environment, Social, Governance
ETL	Extract, Transform, Load
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
FEPA	Foundations for Evidence-Based Policymaking Act
FHIR	Fast Health Interoperability Resources
GAAP	Generally Accepted Accounting Principles
GAIP	Generally Agreed-Upon Information Principles
GDPR	General Data Protection Regulation
GSBPM	Generic Statistic Business Process Model
G-SIB	Global Systemically Important Bank
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level Seven International
HR	Human Resources
IAM	Information Asset Management
IFLA	International Federation of Library Associations and Institutions
IFRS	International Financial Reporting Standard
IM	Information Management
IoT	Internet of Things
ISACA	Information Systems Audit and Control Association
ISC	Information Supply Chain
ISO	International Organization for Standardization
IT	Information Technology
ITAM	IT Asset Management
ITIL	Information Technology Infrastructure Library
KDE	Key Data Elements
KPI	Key Performance Indicators
LIS	Library and Information Science
MAMD	Modelo Alarcos para la Madurez de Datos
MBO	Management By Objectives
MDR	Medical Device Regulation

MIS	Management Information System
ML	Machine Learning
MP	Main Process
MRI	Magnetic Resonance Imaging
MWC	Mobile World Congress
NAO	Network Administrative Organization
OCR	Optical Character Recognition
OEM	Original Equipment Manufacturers
PA	Process Attribute
PAM	Process Assessment Model
PAS	Publicly Available Specification
PDCA	Plan, Do, Check, Act
PDS	Personal Data Services
PII	Personal Identifying Information
PO	Process Outcome
PRM	Process Reference Model
RIM	Records Information Management
ROA	Return on Assets
ROE	Return on Equity
ROI	Return on Investment
ROT	Redundant, Outdated, Trivial
RPA	Robotic Process Automation
SAM	Software Asset Management
SCOR	Supply Chain Operations Reference
SDM	Single Data Marketplace
SEI	Software Engineering Institute
SLA	Service Level Agreements
SME	Subject Matter Expert
SSOT	Single Source Of Truth
TOC	Theory Of Constraints
UCUM	Unified Code for Units of Measure
UNE	Una Norma Española
WM&I	Wealth Management & Insurance

Part I
Data Governance Fundamentals

Chapter 1
Introduction to Data Governance: A Bespoke Program Is Required for Success

Peter Aiken

This database ain't big enough for the two of us
– Bumper sticker seen on an automobile in Texas

1.1 Chapter Overview

The bumper sticker should really have stated "There is no database big enough for two bosses." Importantly, (1) this has always been true, and (2) it means absolutely nothing to most of the public or much of Information Technology (IT). Let's address each of these separately.

Just as in any situation where coordination, integration, and information are required, there must be one and only one individual implementing decisions to maintain integrity, continuity, and operational capabilities. Required minimally from a change management perspective, this can always be used to justify Data Governance (DG) in general. Ask the skeptical: "how can any complex adaptive system function with multiple Chiefs?"

The public and unfortunately too many in business and IT do not understand this sort of basic law of (data) nature. Because they are not data literate, when someone proposes having multiple chiefs for database operation, or that group X should "own" dataset Y, or that the DG group should report to the Chief Information Officer (CIO), they do not know these are not workable concepts!

DG is not focused primarily on databases, clouds, or other technological ephemera. Instead, the DG framework must be understood identically by business users, systems personnel, and the systems themselves (as shown to the right; see Fig. 1.1). This essential, metadata-based communication is at the heart of any enterprise operation. DG removes barriers to data efficiencies, allowing organizations to

P. Aiken (✉)
Virginia Commonwealth University, Richmond, VA, USA
e-mail: peter.aiken@vcu.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_1

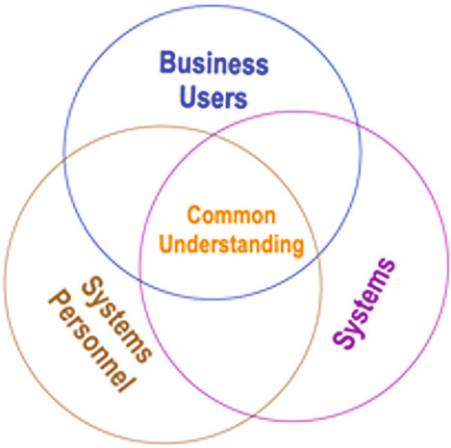

Fig. 1.1 Essential metadata-based communication about data in DG

function more effectively and efficiently. Resources consumed by bad data practices can now be used to support the mission.

Increasingly organizations are attempting to do "more" with data. This "more" represents the other strategic dimension, "innovation." By definition, most attempts to innovate will fail, so the lessons learned by becoming more effective and efficient will also help in this "innovation" dimension. Innovating with data requires programmatic support for the efforts – well supported by data infrastructure and mature organizational data practices.

It is the responsibility of DG programs to manage this and other delicate balancing acts required to successfully contribute to better organizational use of data. DG is a comparatively new, certainly unstandardized, and under-studied topic. While some excellent DG programs are maturing, the majority have not. This leaves individuals and organizations the sequential tasks of:

1. Learning about data
2. And then learning about their data
3. Next, developing plans to increase the data literacy of their executive leadership and their knowledge worker population before expecting to make progress faster and further with data

This chapter takes you through the who, what, where, when, why, and how of DG. It provides a common basis for building individual and organizational knowledge of this topic – starting with the *why* (the motivation for DG) followed by the *who*, *when*, and *where*. The how section is a bit longer and the bulk of the remaining material concentrates on the *what* – a way to successfully start to govern subsets of your data.

Most organizations should not attempt to govern *all* of their data. Successful DG program goals include subsetting their data into essential and nonessential data. Governing the essential subset and ignoring (or better still removing) the rest reduce the size of the challenge. Since the definition of an organization's essential data will

differ from organization to organization, the governed data will also differ among organizations.

One quick word about the use of the term *bespoke* in the title, it is of course deliberate. The only way that *your* organization can use data to better support organizational strategy is to use *your* data in support of *your* strategy using the capabilities that *you* currently have. Cookie cutter methods will not help *your* organization learn about your data!

1.2 Why Does Data Need to Be Governed?

A friend was speaking with an organization on data matters and noticed that the urinals in the restrooms all had unique numbers (see Fig. 1.2). Presumably this was in case of malfunction so that the specific instance could be more rapidly identified. Of course, my friend used a suitable-for-work (as opposed to not-suitable-for-work) photograph to make a point to leadership that (at least for this organization) it was worthwhile to keep maintenance histories of this equipment type. Ironically, it was noted that the substance of the discussion for which my friend had been invited was whether the organization should maintain similar information about their

Fig. 1.2 Urinals with unique numbers

organizational data assets. The photo provoked a nice motivational discussion with a decision to proceed with DG as the outcome. After all, if we are going to govern our restroom facilities, shouldn't we also govern our data assets?

Writing as a deeply, industry-immersed university professor, I can say that the academic community has failed its customers with respect to integrated data knowledge. For generations we have graduated students who have become leaders in business and IT. The only class taught about data was really a class about database development. Smart students who placed their trust in the educational system were educated that the only concept they needed to learn about data was how to build new relational databases! No one should be surprised that one of the major DG challenges is that far too many poorly designed databases clutter most organizations or (more increasingly) their clouds. As Abraham Maslow stated, "If the only tool you know is a hammer, every problem looks like a nail."

When considering the asset itself, data has a unique collection of properties including the following from Doug Laney. Data:

- Does not obey all of the laws of physics
- Is not really visible
- Is non-rivalrous (many can use it at once)
- Has zero costs in providing an additional copy
- Is nondepleting
- Does not require replenishment
- Is regenerative
- Has low inventory and transportation/transmission costs
- Is more difficult to control and own than other assets
- Can be eco-friendly
- Is impossible to clean up if you spill it [1]

When considering career fields and learning experiences, not all data professionals take similar paths. For example, data scientists often discover useful data maintenance utilities instead of learning that various classes of tools exist and when to apply each as part of their educational programs. For many, data is like the story of the blind men and the elephant, and collectively it is DG responsibility to shape this understanding into an organization-wide perspective.

For these and other reasons, there continue to be questions as to whether data processing should continue to be part of IT or of the business or of special operations such as finance and risk? While the Federal Government resolved this issue correctly with new FEPA legislation, the jury is still out on the rest of the world. Currently it is comprised of 1/3 of each type: 1/3 reporting to CIOs; 1/3 reporting to CEOs; and 1/3 reporting to CFOs/CROs.

[1] See Datanomics by Doug Laney, Routledge Publishing 2017 ISBN 1138090387.

1.2.1 Long-Lasting Consequences of Poor Data Decisions?

Unfortunately, short-term application-centric thinking [2] has dominated, relegating development of data products to subsets of ERPs, digitization initiatives, or cloud-hosted projects (to name just a few types). Virtually none of the popular software integration packages from the major vendors have escaped the long-term consequences of inadequate data Design (big "D" is used to emphasize the *entire* life cycle). These well-documented imperfections are locked in for life – wrapped as they are, in a dense set of application constructs interwoven with the imperfect data model. Worse still, the corrections to the organization's data and processing are layered on as additional code – complicating the apps still further. The vast majority of database functionality is not used beyond table-handling. In this manner, developers restrict any subsequent data investment benefits and decrease data leverage potentials. At the very least, DG must illustrate and resolve the 20–40% of IT budgets that are devoted to data evolution:

- Data migration (changing the data location)
- Data conversion (changing data form, state, or product)
- Data improving (inspecting and manipulating, or rekeying data to prepare it for subsequent use)

None of these are accounted for in the usual (and very important) *data storage costs – measure*. DG must also articulate these various costs and trade-offs associated with increased data rigor (or the risks of not doing so) to the rest of the organization.

1.2.2 Mounting Data Debt

The failure to do any of this has caused organizations to pay to accumulate large amounts of data debt. (Yes, the indignity that your own organization is creating data pollution that is directly harmful to its operation should be professionally embarrassing!) It is not easy to visualize the cost of data debt, but the phrase *many many many unnecessary paper cuts* [3] describes the situation well. Data debt slows DG efforts making everything slower, of lower quality, cost more, or present increased risks.

Data debt is like quicksand that mires down all efforts. Defined simply, data debt is the time and effort it will take to return your data to a governed state from its likely current ungoverned state. A quick back of envelope calculation of data debit can be

[2] See The Data-Centric Revolution: Restoring Sanity to Enterprise Information Systems by Dave McComb, Technics Publications ISBN 1634625404.

[3] https://en.wikipedia.org/wiki/Paper_cut

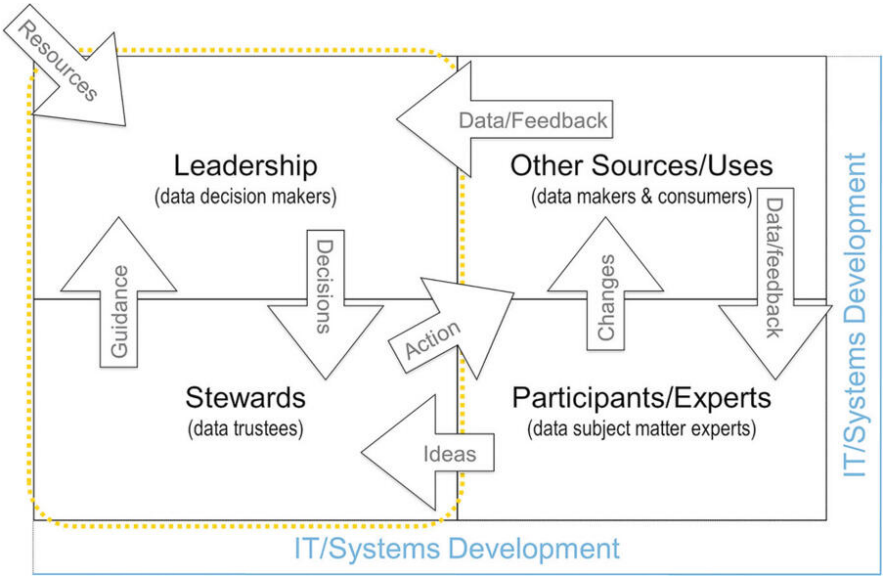

Fig. 1.3 Relations between leadership, stewardship, and other users and participants

done using the data storage costs that are perhaps the most tangible and objective data measure. At least 20% of that data is redundant, obsolete, or trivial (or ROT).

The good news about finding and eliminating data debt is that things can get faster, better, or cheaper. The bad news is that new skill sets are required of the DG team and that diagnostic and analytical systems thinking still requires annual proof of value. The knowledge base of graybeards who know how to apply these skills is shrinking as these individuals are judged expensive and encouraged to retire.

In summary, *data needs to be governed because society was not taught that it required specific treatment until it was too late*. Because individuals do not know that they do not know, it has been difficult to educate them to the need. By focusing on concrete results, organizations have better success making the case that an investment in DG will benefit the organization in specific measurable ways.

1.3 Who Needs to Be Involved in DG?

Unfortunately, at many organizations, everyone has been responsible for data quality, and this approach has produced the current unsatisfactory state. It is critical to start DG educational efforts with executives because (1) they are willing to invest in learning and (2) their data decisions have the greatest impact on the organizational data practices. The next goal for all DG programs is to also increase the data literacy of all organizational knowledge workers.

As illustrated in Fig. 1.3, DG efforts are generally built on an IT-provided support/foundation/infrastructure. A leadership component provides resources and clears barriers for the effort. Primary functions are (ideally full-time) data stewards who provide guidance and design/implement decisions. Typically, these two groups form the basis for DG organizations. Also, highly involved (and incorporated) are various SME or subject matter experts who know the required data and processing details. Then of course there is everyone else. As noted, DG efforts need to be integrated with both organizational and IT governance.

1.4 When Is It Appropriate for Organizations to Invest in DG?

By now I hope that you agree this is a silly question. The 20–40% of IT costs (referenced previously) are easily gauged. As the DG practice matures, processes can be optimized for key operations. By keeping disciplined measures, organizations have developed expertise in these practices. Keeping the focus on an integrated full-time team permits the case to more easily be made when timing investment in a second or third DG team.

Digital and data are dependent on high-speed automation/data processing that requires significant amounts of organizational data literacy, data standards use, and quality data supplies. Continue to evaluate and evolve DG frameworks to refine the organizational focus. Over time this approach should evolve into the standard Deming Plan-Do-Check-Act (PDCA) cycle.[4] An incomplete list of potentially useful standards that can be created with the required measurable controls is listed below.

• Access standards	• Classifications
• Change management	• Secure
• Security	• PII
• Storage	• Competitive advantage
• Reporting	• Public

1.5 Where Should Organizations Get Started with DG?

DG is a rare triple benefit capability that helps refine data strategy, improve the quality of the players, and improve data used to support the mission. However, getting started with DG can be and has been accomplished by moras of ill-defined and vendor-specific methodologies – most of which have no reported research results.

[4] https://en.wikipedia.org/wiki/W._Edwards_Deming#PDCA_myth

An easily understood model (the theory of constraints [5] or TOC) views programmatic data support as a manageable system. The system is limited in achieving more of its goals by a small number of constraints. There is always at least one constraint, and TOC uses a focusing process to identify the greatest constraint and restructure the rest of the organization to address it. TOC adopts the idiom that "a chain is no stronger than its weakest link," and processes, organizations, etc. are vulnerable because the weakest component can damage or break them and adversely affect the outcome.

Key is to visualize the various data flows through the organization and understand the value of controls in relation to various processes, risks, outcomes, and performance. The costs of various blockages can be ranked and estimated. What changes made at the data level could most help the organization achieve its strategic goals?

Iterative problem-solving provides additional benefits beyond challenge solutions. Team problem-solving enables increased organizational data literacy and some go as far as considering these capabilities their "secret sauce." It just makes sense to support a group of individuals who possess knowledge of your data and its uses.

Focus first on organizational strategy. Understand intricately the data flow supporting increasing performance, decreasing costs, impacting times, and better managing risks. Identify the various types of organizational challenges sharing the same data or (better still) data errors. These become the focus of the first iteration of a data strategy cycle. It is overseen by the DG program and coordinated to be most collectively helpful to organizational as well as IT strategy. Ensure you complete a full cycle to include feedback/improvement/lessons learned/organizational memory/change cycle components. Heavily incorporate the use of "branded" data checklists and standard control development.

And then (as it says on the shower bottle), lather, rinse, and repeat. This is really the only way to escape the bad data cycle. IT and business decision-makers are not knowledgeable about data and good data practices. They make poor decisions about data that result in poor treatment of organizational data assets and poor-quality data. Both of these lead to poor organizational outcomes (see Fig. 1.4).

1.6 How Should Organizations Apportion Their DG Efforts Over Time?

1.6.1 Data Debt's Impact

Over time, organizational data debt clogs value-adding pathways in a manner similar to the 40% of the internet that is now clogged with malware. Data debt is responsible

[5] https://en.wikipedia.org/wiki/Theory_of_constraints

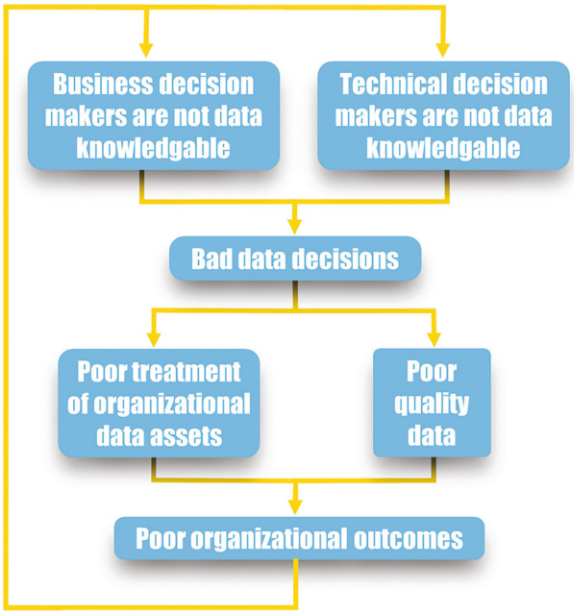

Fig. 1.4 From business and technical decisions to poor organizational outcomes

for inflicting uncounted tiny hidden data factories[6] on organizational performance – making everything cost more, take longer, deliver less, and at increased risk. Eliminating data debt requires a team with specialized skills deployed to create a repeatable process and develop sustained organizational skill sets.

A major motivation for increasing the data literacy of all knowledge workers comes from the fact that most organizational challenges come filtered through various IT and business practice combinations. The reason for multitude of paper cuts is that the DG challenges are filtered through various business processes and IT systems. As a result, common challenges go unrecognized with each instance requiring treatment instead of correcting the underlying data challenge (see Fig. 1.5).

A key aspect is to evaluate your architectural abilities to build/evolve toward organizational data capabilities in a three-step process. First, you need to improve the quality of existing organizational data. Too many organizations do not have enough information about the quality of their existing data. These data quality challenges fall into two categories: practice-related data quality challenges and structure-related data quality challenges. Second, the framework must support your efforts to increase the data literacy of literally your entire executive team and knowledge worker population and especially those who already practice data. Finally, only when you have improved your data and your organization's ability to work with data can you hope to improve the way that data supports your organizational strategy.

[6] https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

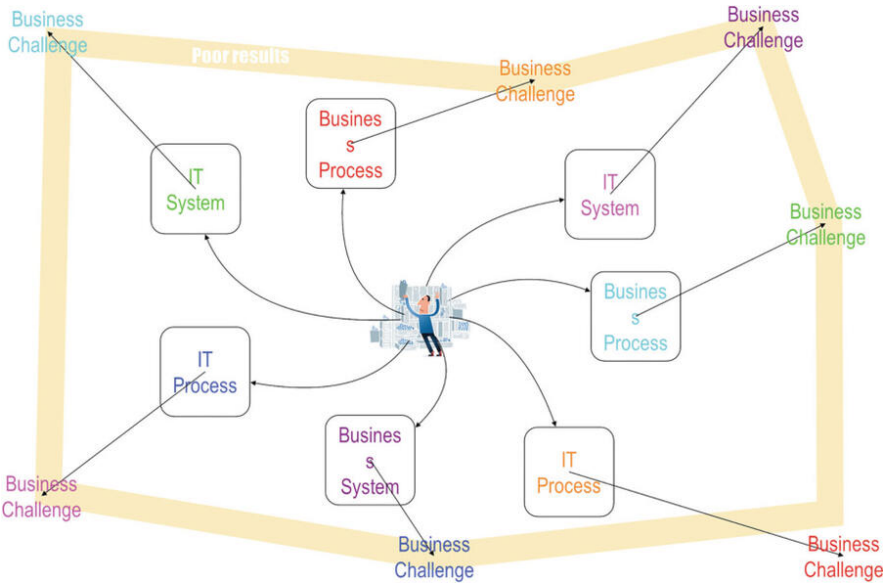

Fig. 1.5 Common DG challenges result in business challenge

1.6.2 Proactive Versus Reactive DG

One rather traditional realization (almost a rite of passage) is that whatever changes are made to the organizational data practices might take literally years to be able to exploit it. In CIO terms, it can often be a successor that will benefit from DG initiatives. As this realization sets in (that time equals years), DG initiatives come under pressure to "do something more quickly." As illustrated, a secondary capability is established to more effectively produce results as a result of direct intervention or data improvement projects (DIPs) (see Fig. 1.6).

1.6.3 MacGyver Abilities

While perhaps not widely acclaimed, the 1980s TV series MacGyver became shorthand for a nontraditional and innovative problem solver who always carried a Swiss army knife.[7] In the same manner, the DG program must imagine itself as the "help desk" for organizational data. Literally all data challenge solutions should be minimally coordinated and, in many instances, led by DG. The key is to develop new data capabilities within a dedicated group focused on organizational data

[7] https://en.wikipedia.org/wiki/MacGyver

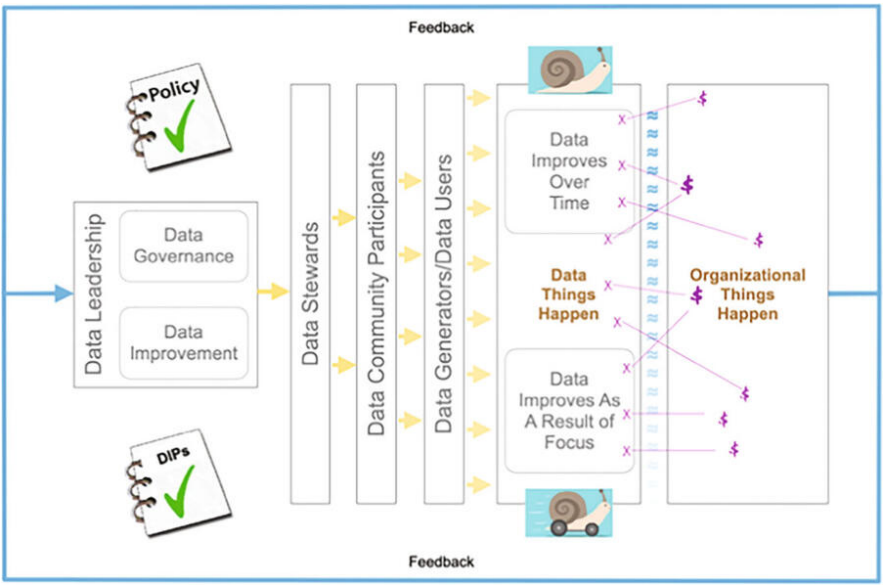

Fig. 1.6 Results of data improvement projects

governance. Have this group focus on and conquer a series of DG challenges, producing positive ROI numbers.

1.7 What Organizational Needs Does DG Fill?

It is useful to describe the organizational needs that DG fills. These include:

- Improving the way that data is treated as an asset
- Available but not widely known research results
- Using data to better support the organizational mission
- Using data strategically

1.7.1 Improving the Ways That Data Is Treated as an Asset?

One of the primary challenges for organizations is to learn how data requires specific considerations. If you consider data as an asset (and currently most business leaders do not yet do so), then one should expect that it would be treated as other organizational assets. I use a series of questions developed by my colleague Dr. Christopher Bradley to help organizations determine whether their data is maintained as an asset. They are as follows:

1. Do you have executive positions to support data as an asset?
2. Does the organization track usage of this asset?
3. Are organizational or fiscal controls put in place to manage this asset?
4. By and large, are these controls actually executed?
5. Is there general acceptance of the need to manage this asset? That is, do people "get it"?
6. Do serious discussions about this asset feature on the agenda of senior management meetings?

Using this rather obvious set of criteria, it is easy to determine that most organizations are not treating data as an asset, but so far, we do not survey results on this particular measurement.

1.7.2 Available but Not Widely Known Research Results

As referenced above, there is a dearth of knowledge about data much less data governance. On that note however, we do have access to two solid lines of research to which I will refer throughout this chapter. The first is in the form of the annual (2013–today) data practice surveys conducted by NewVantage Partnerships and are reference able at: https://www.newvantage.com/thoughtleadership. Annually several thousand of the same or similar organizations have been asked the same questions repeatedly providing pictures of how issues are considered over time. Results reproduced here will be referred to as *New Vantage*. A second set of research results come from the collaboration (called the Data Literacy Project) between Accenture and Clique. These results will be referred to as *Data Literacy Project* and are referenced at https://thedataliteracyproject.org/. These two efforts have provided a good framework that can be used to dive further into research in this area.

One of the New Vantage results has been the following: what percentage of your data challenges are people-/process-related versus technology-related? The consistent answer (see Fig. 1.7) continues to surprise: not once since 2018 has the percentage of technology challenges risen to above 20%. This means that for more than 6 years, *everyone* should have known that the people/process dimension of DG represents the largest challenge. Yet very little organized research beyond surveys has been conducted into this area.

Consider the following please: what group in your organization is in charge with decreasing the number and impact of people- and process-oriented data challenges? This is precisely the role that your DG organization must address in your organization. *If not DG, then who* in your organization is responsible for improving the people and process aspects of your data operations?

It is crucial that DGs provide a holistic view of minimally the above detail but also include data's role in the organization, how individuals can assist, and where to go for more information.

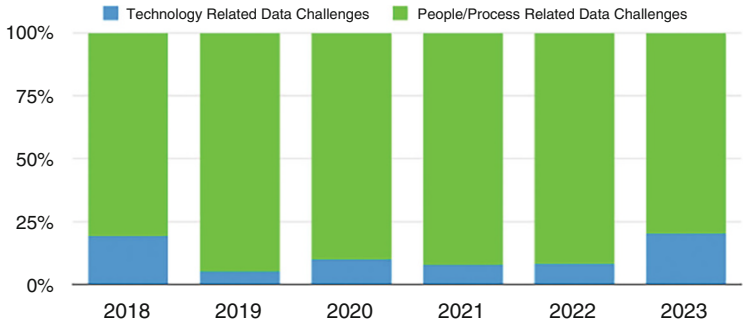

Fig. 1.7 Percentage of technology-related data challenge vs people-/process-related data challenges

1.7.3 Using Data to Better Support the Organizational Mission

This section's title "Using Data to Better Support the Organizational Mission" must be the mission of any DG program. But first a specific word about data ownership (bad concept) and data requirements ownership (good concept).

Avoid a first (and always a major) misstep: trying to assign data "ownership." While it is tempting to "establish data owners" as a goal of data governance, it is usually a bad idea. However, many are familiar with the process architecture practice. It correctly embraces and leverages the term "process owner" as the single individual responsible for the integrity of the process design, implementation, and improvement.

While it makes intuitive sense, the concept of data ownership has caused more DG effort to fail than any other. As soon as you allow an underinformed individual (or group) to "own" any data items, they begin to make decisions about the data that optimize it from their local perspective. If your organization does not formally manage a process architect, skip to the next paragraph. If it does, careful analysis will yield maintainable, high-level process/data interaction matrix called a CRUD matrix – showing data/process interaction by access type (see Fig. 1.8). (CRUD matrices such as the one illustrated show business processes and their activity type **c**reating, **r**eading, **u**pdating, and **d**eleting various data items.)

If nothing else, these maintainable metadata collections show the interdependencies: data exist only to be consumed by various business processes, and only purpose for a business process to exist is to produce data to be consumed by another business process. If you do not have an organization CRUD matrix at hand and need to shut down any data ownership conversations, ask the question: "To whom does the data that accounting stewards belong?" Since accounting processes data from across the organization, a case could be made that accounting "owns" much organizational data.

	Business Process 1	Business Process 2	Business Process 3	Business Process 4	Business Process 5
Data Item A		Create	Read		Delete
Data Item B	Read	Create		Update	
Data Item C			Read	Update	
Data Item D	Create	Update	Delete	Update	
Data Item E		Create			

Fig. 1.8 CRUD matrix for organizational business processes

The reason data ownership is such problematic concept is that data persists across business functions. Ownership would only apply to a specific data processing stage. Instead of asking the question, "who are the data owners?", the statement should be that all data belongs to the organization! At best, ownership could only be limited to specific life cycle phases.

If the organizational culture requires use of the word ownership, then allow ownership of the *data requirements!* Local expertise should be used to specify the size and shape of the specific data items required to perform organizational functions at various stages of data at it is processed.

1.7.4 The Role of DG Frameworks

All evidence to date points to frameworks being useful:

- As system of ideas for guiding subsequent analyses
- As means of organizing measures and project data and then assessing progress
- For evaluating priorities for data decision-making
- For assessing overall functionality
- For moving toward a determination of ROI [8]

For example, a building construction conceptual framework would incorporate bits of wisdom such as the following:

- Don't put up walls until foundation inspection is passed.
- Put the roof on ASAP so that work can proceed in inclement weather.

[8] Interestingly, ROI means risk of incarceration to most DG professionals.

- Make it each construction phase dependent upon continued funding by passing a series of checkpoints.

Much has been written about data governance frameworks. I have seen research proposals that anticipate evaluating one type of framework against another. It is far too early to start to "type" DG frameworks. Nonstandard understanding of terms and data concepts leads to "results" of the sort that were popular at the start of the CDO movement. (Note: Researchers have tried and failed to establish correlations between having a CDO and organizational financial performance – similar specious results can be expected until the entire DG profession matures.)

Use the existing DG frameworks to envision what your program should look like given your organizational needs. "Try each of them on" conceptually and discuss the suitability of each for your organization. Since no two organizations are alike, each organizational DG program must be custom fitted to the organization rather like getting fitted for a suit. The word "bespoke" well describes the design of DG programs that provide good returns on organizational DG investments.

It is quite useful to view representations of various approaches to DG in the same manner that an architect presents sketches of a future building to prospective funders. The utility of DG frameworks generally stops at this point. There are essentially few types of DG frameworks in popular use. (Note: You can see representations of many of these at https://anythingawesome.com/DataGovernanceFrameworksCollection.html.)

All subsequent are theme and variations on these. Pay no attention to "proprietary" methods. The goal is to give you something to compare, contrast, and consider when designing the first version of your DG organization. (Note: This first version will evolve to a second and third as the organization; DG practices should mature and evolve over time.)

This is where the concepts of stewardship and fiduciary responsibilities come into play. *Stewardship* in this concept is derived from the definition "a person employed to manage another's property." *Fiduciary* is used to describe the nature of the relationship as involving trust, especially with regard to the relationship between a trustee and a beneficiary. This is accompanied by specific duties.

1.7.4.1 Related Term Definitions

It is now time to introduce a few terms to show both the evolution/etymology of the term DG and the most useful definition of DG.

Let's start with the term *governance*: "Governance is the process of interactions through the laws, norms, power or language of an organized society over a social system (family, tribe, formal or informal organization, a territory or across territories). It is done by the government of a state, by a market, or by a network. It is the decision-making among the actors involved in a collective problem that leads to the creation, reinforcement, or reproduction of social norms and institutions" (https://en.wikipedia.org/wiki/Governance).

Corporate governance is next. Below are three good definitions highlighting different aspects of this evolving concept.

- "Corporate governance - can be defined narrowly as the relationship of a company to its shareholders or, more broadly, as its relationship to society…" (*Financial Times*, 1997).
- "Corporate governance deals with the ways in which suppliers of finance to corporations assure themselves of getting a return on their investment" (*The Journal of Finance*, Shleifer, 1997).
- "Corporate governance is about promoting corporate fairness, transparency and accountability" (James Wolfensohn, World Bank President, *Financial Times*, June 1999).

Note that the concept of corporate governance is evolving. Just before the pandemic, Jamie Dimon (then head of Chase) led a group of CEOs to proclaim "Maximizing shareholder value can no longer be a company's main purpose." [9] Similarly, the concept of DG continues to evolve.

Well, if corporate governance exists, then certainly *IT governance* should be a useful concept. It is and is defined as "Putting structure around how organizations align IT strategy with business strategy, ensuring that companies stay on track to achieve their strategies and goals, and implementing good ways to measure IT's performance. It makes sure that all stakeholders' interests are taken into account and that processes provide measurable results" (https://en.wikipedia.org/wiki/Corporate_governance_of_information_technology).

IT governance frameworks should answer some key questions, such as "How is the IT department functioning overall?", "What key metrics does the management need?", and "What return IT is giving back to the business from the investment it's making?". Included are typically foci on:

- Strategic alignment
- Value delivery
- Resource management
- Risk management
- Performance measures

IT governance is an established discipline with common vocabulary and understanding among those who participate. [10] Of note is the fact that data practices are not typically included as a topic under IT governance or are lightly treated. This may account for or reflect the current slowly maturing state of DG practices.

Data governance has suffered from both too many definitions and inaccessible (by the business) terminology. However, auditors easily get the concepts. Below are some standard definitions of DG.

[9] https://www.marketwatch.com/story/maximizing-shareholder-value-can-no-longer-be-a-companys-main-purpose-business-roundtable-2019-08-19

[10] https://en.wikipedia.org/wiki/Corporate_governance_of_information_technology

- "The formal orchestration of people, process, and technology to enable an organization to leverage data as an enterprise asset." – The MDM Institute
- "A convergence of data quality, data management, business process management, and risk management surrounding the handling of data in an organization." – Wikipedia
- "A system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods." – Data Governance Institute
- "The execution and enforcement of authority over the management of data assets and the performance of data functions." – KiK Consulting
- "A quality control discipline for assessing, managing, using, improving, monitoring, maintaining, and protecting organizational information." – IBM Data Governance Council
- "Data governance is the formulation of policy to optimize, secure, and leverage information as an enterprise asset by aligning the objectives of multiple functions." – Sunil Soares
- "The exercise of authority and control over the management of data assets." – DMBoK

Technically they are all correct but imagine the following scenario. Stepping onto an elevator for a minute-long ride and an executive enters the car. As the doors close, the executive turns and says, "I've heard you are working on DG. Can you tell me what it is? I'm confused." Imagine responding with "DG is the exercise of authority and control over the management of data assets." Do you think the executive would (1) find the answer useful and (2) think well of your ability to communicate this concept?

I think the answer is no to both questions. A better response to the executive is "DG is about managing data with guidance." Short and to the point, this definition incorporates self-explanatory motivation. When I provide this information (the definition of DG) to most executives, their first question to me is "So we have not been managing our data with guidance?". The answer usually is "Only recently have we been managing our data with guidance." Of course, the eternal hope is that the executive will be curious to learn more and present an opportunity to become more data literate. Subsequent conversation topics could include the following:

- Why it is generally not a good idea to govern all of your data.
- Why DG will never be complete at our organization.
- Why some decisions that involve data are not considered as such.

The Data Literacy Project reports that four out of five executives surveyed were willing to invest time resources in improving data skill sets. This represents a once in a generation opportunity to reach these executives with good DG education. (Note that anyone offering to improve your organization with DG *training* should be ignored – the process requires *education*, not training.)

1.7.4.2 A Small Concentrated Team Is Preferred Over Distributed (Dissipated) Knowledge

The next item to consider is what format DG should take. Remember, asking *everyone* to be responsible (for data, data quality, data governance, etc.) has produced the current state of affairs. Organizations assigning new DG duties to existing personnel have two options: (1) incorporate the new duties along with existing duties and (2) assign these DG duties to full-time individuals.

When considering this, it is useful to ask: how long will the need to manage data with guidance exist? The answer turns out the be: *you will need your data program as long as your organization needs to have its finance, HR, and planning operations*. Think about it in the future: Will more or less data exist? Will data collection modes increase or decrease? Will data be found in fewer or more formats? A solid recommendation is to staff with full-time team members dedicated fully to DG. Data literacy and organizational data practice maturity are generally low. Dedicated personnel interacting with each other more – greatly stimulate their individual learning curves. It also makes tracking DG program costs clearer. It is critical to begin to build organizational DG capabilities. This can best be started with dedicated teams with a clear ROI. Against these, results can be evaluated.

1.7.5 Using Data Strategically

The next question is *on what* do we focus these DG efforts? In regulated environments, these efforts are often compliance driven. Key is to approach these efforts in the same manner. Do we think that regulations will increase or decrease in the future? If increasing, then it seems useful to "get good" at implementing compliance-driven changes. If nothing else, you may gain an implementation advantage over the competition subject to the same data regulations but perhaps not able to implement as efficiently or effectively. Data regulation compliance can become a valued organizational capability with an easily determined ROI.

Outside of compliance, organizations strive to use data strategically with either efficiency/effectiveness or innovation goals. Personal interaction with more than 1000 organizations indicates that about ½ have clearly articulated strategic goals and objective measures supporting goal achievement at the organizational level. Absent these, it is not possible to improve the manner in which data supports this Jell-O strategy. I also find universal distain for 3–5-year plans, most of which fell apart rapidly with the onset of the Covid-19 pandemic. So just a word of caution, check your organizational strategy to ensure it has clear objective and measures before attempting to improve how data can support it.

1.7.5.1 Strategy Is About Why

. . .it's not *what* you do, it's *why* you do it. . .

Among many great TED Talks, Simon Sinek's "How Great Leaders Inspire Action" is a favorite. Recorded in 2009, Sinek's talk has enjoyed more than 25 million views. His point is quite simple: most of us are very good at describing what we do, and some of us are good at describing how we do things. Not as many of us are good at describing why we do things.

Strategy is the highest-level guidance available to an organization, focusing activities on articulated goal achievement and providing direction and specific guidance when faced with a stream of decisions or uncertainties. More succinctly, strategy is a pattern in a stream of decisions. This pattern must be supported by data or it will not be possible to determine if the strategy is correct or working.

1.7.5.2 What Is Data Strategy?

Data strategy is the highest-level guidance available to an organization, focusing data-related activities on articulated data program goal achievements and providing directional and specific guidance when faced with a stream of decisions or uncertainties about organizational data assets and their application toward business objectives. The data strategy must be understood and supported at the organizational level. Only with this level of scrutiny and involvement can a true systems view be applied to the challenge of improving how data can support strategy.

1.7.5.3 Working Together: Data and Organizational Strategy?

Figure 1.9 indicates the close relationship among organizational strategy, data strategy, and data governance. Two key aspects of the interaction are as follows: (1) express the data strategy in terms of specific business goals, and (2) ensure that the language of DG is metadata.

1.7.5.4 Strategic Commitment: Program Versus Project Focus

A commonly asked question is "When will you be done?". This is a warning that the individual considers DG a *project*. Organizations failing to implement DG at the program level (as a program) are unable to view the totality of their data challenges holistically, and the solutions fail. Many organizations require a second or increasingly a third DG "reset."

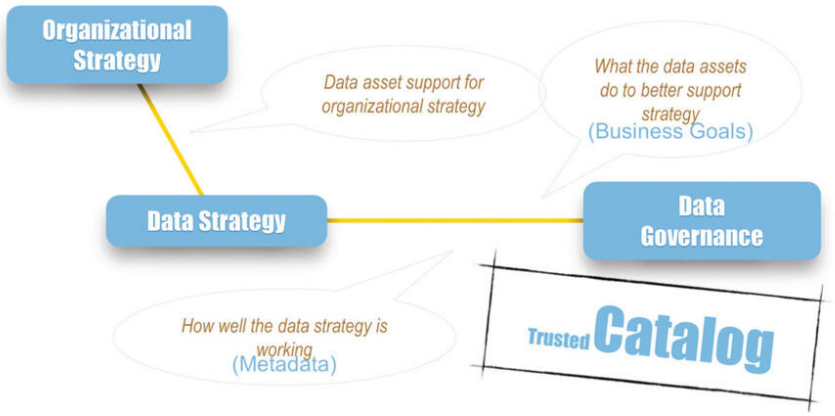

Fig. 1.9 Relationship among organizational strategy, data strategy, and data governance

Fig. 1.10 Garbage data results in garbage digital results

1.7.5.5 Digitizationing

One of the more important areas that DG can be focused to support is "going digital." Once again, many vendors have offerings and expertise in these areas. DG sets the standards required to support digitization because you cannot "digitize" without a good data capabilities foundation. Garbage in, garbage out is always true. At this point, effective DG is a requirement for digitization; otherwise you will be unable to trust any digital system outputs (see Fig. 1.10).

1.7.5.6 A Watchful Eye Toward the US Federal Government (FEPA)

Finally on the *what* question (yes – we are still in *what*), it will be useful to observe the progress being made in the US Federal Government. As part of my service as a DoD employee, our group is often sent to "learn from the private sector." Now the situation has been reversed. In 2019 the **F**oundations for **E**vidence-Based **P**olicymaking **A**ct was signed into law. Three specific aspects of the law make this especially interesting for DG to follow. They are the following:

- Explicitly nonpolitical CDOs must be established separate from CIO roles. From a DG perspective, organizations have been slower to adopt CDOs with non-CIO reporting role.
- Government data is now open by default and must be maintained using open standards. In just a few years, the Federal agencies will have developed a great deal of expertise in these areas.
- Use of open data and open models is required in policy evolution. Policy changes are only permitted with both models and datasets specified prior to the analyses and decisions.

Collectively these efforts, if fully implemented, will improve governmental decision-making and overall effectiveness. More importantly, all impacted Federal organizations are also rapidly developing and implementing DG as compliance activities still further increasing the pool of DG professionals worldwide.

1.7.6 Breaking Through the Barriers of Data Governance

There are a host of barriers to implementing DG. This includes the usual failures to include change management and cultural refocusing as key dependencies. While the accounting profession has had literally millennia to develop GAAP, no such guidance exists for DG. There is a vast tendency to depend on technologies that are incapable of acting as silver bullets.

An example of these difficulties was illustrated in 2020 when Forbes ran an article on airline valuations. [11] It purported to show how the airlines were monetizing the data in their frequent flyer programs. However, the buried lede was that in 2020, both United and American Airlines were valued at tens of billions of dollars less than the anticipated value of the data in these programs. You better believe that if airline leadership could have unlocked that value during the time when most were avoiding flying (the pandemic), they would have unlocked it ASAP! The fact that they were unable to do so highlights the uphill climb that poorly fitting DG efforts face.

Some basic DG execution principles follow:

[11] https://www.forbes.com/sites/advisor/2020/07/15/how-airlines-make-billions-from-monetizing-frequent-flyer-programs/?sh=66da87a614e9

- Ensure that the organization's data strategy is properly aligned with the business strategy. Implement regular processes with key stakeholders to ensure proper alignment.
- Ensure that data debt is properly being managed and the process is under statistical control.
- Perform a capability maturity assessment or "reassessment" to determine the required maturity. If the maturity levels are not meeting expectations, ensure that there is a remediation plan with a properly monitored work-arounds.
- Consider refresher training for your knowledge workers and data professionals, e.g., data stewards, architects, and engineers, as a feedback mechanism for determining needed improvements and remediations.

Based on the organization's strategy, the DG group must determine if they are to initially follow a model primarily focused as a:

- Utility – back office, efficiency goal
- Steward – more asset focused, quality goal
- Enabler – strategic partner, innovation goal

This should be determined through the building of the data strategy. If an organization is striving toward a modernization transformation, DG should trend to an "enabler." To measure the effectiveness of an enabler, DG standards should be repeatable and statistically stable. The focus can be changed at a later stage, but the focus can be on effort and discussions during initial phases.

Hopefully your organization will be spared major data catastrophes. It is more likely you will experience one or more in the future. In this event, attempt to learn as much as possible from the event. Take, for example, the story of two major banks in the process of consummating an arranged marriage. The deal came down to a single spreadsheet containing many rows, each representing an asset. If an asset on the spreadsheet was to not be transferred, that row was hidden with agreement by both parties. After final agreement was reached, the spreadsheet was handed to a junior associate who was told to "make it look nice for the Judge tomorrow." Unfortunately, late in the evening, the junior accidentally unhid hundreds of rows and did not notice! Presented to the judge as the golden copy, the judge would not reverse – even on appeal. [12] As you might imagine, DG practices around the use of spreadsheets are quite extensive. I assisted 1 organization with the elimination of more than 400,000 legacy systems of a certain type. The list of preventable spending continues.

Unfortunately, the conversations have been generally unsatisfactory. Key to getting started with *data valuation* is to add up "at least" instead of attempting to master the entire costs. I justified an investment into an organizational repository at one organization with a business case built on the premise of saving everyone in IT *1 hour annually*. The organization conducted surveys asking if the 1-hour saving was achieved. It was!

[12] https://www.businessinsider.com/2008/10/barclays-excel-error-results-in-lehman-chaos

When determining the internal and external value of data, two prerequisites exist: first, business and data strategies must support data monetization, and second, DG must be effective and properly measured. Components of data value can include:

Internal	External
• Properly managed data debt • Efficient usage of cataloging and master data management • High trust in supplier and customer data integration • Measured positive ROI	• Organizational data monetized in a public market or exchange • Organizational data becomes a profit center • Organizational data becomes a Band-Aid of adhesive strips

Sometimes it is easier to highlight the value with unfortunate examples with clear costs to society. Early Covid-19 monitoring was inhibited because health care workers did not know how to save MS Excel data sheet and workbooks as .xlsx instead of .xls files. The difference, unknown to the users, was that the older .xls files dropped all rows beyond the 16,000th or so row without warning. We will likely never know how much better performing the early monitoring systems were because all the errors are in one direction.

On a cheerier note, an agency charged with home evaluation/intervention discovered that 40 questions on its evaluation assessment were immaterial. This shortened each interview by half and ultimately shifted more than $1 million from overhead to service delivery.

In terms of execution, DG should be viewed as an iterative process that the organization is striving to *get better at!* Each cycle focuses on aspects of the various data challenges with a goal of eliminating or reducing the impact of a specific constraint. To understand the importance of this shift in thinking about DG, consider the circumstances where a plan was the goal. It was former President and General Eisenhower who said:

> In preparing for battle I have always found that plans are useless, but planning is indispensable. [13]

Mike Tyson's version is that everyone has a plan until they get punched in the face. A team knows how to react to unforeseen challenges and efficiently address the ones they have planned for. The PDCA cycle provides operational context.

[13] https://quoteinvestigator.com/2017/11/18/planning/

1.8 Chapter Summary

The word **bespoke** has evolved from a verb meaning 'to speak for something', to its contemporary usage as an adjective. Originally, the adjective bespoke described tailor-made suits and shoes. Later, it described anything commissioned to a particular specification. Wikipedia [14]

The difference between data analysis capabilities and data requiring analysis is increasing. DG will continue as a maturing and growing field and can only be assisted by increased research into the various challenges outlined. Practice standardization and improvement are clearly the next steps on this industry's maturity curve. As a new discipline, DG works best directly addressing the manner in which data is used to support achievement of organizations' strategy. There is no other best way and right now there is no agreement on terminology, hence on anything. Consequently, the only way to obtain a positive ROI on investments in DG is to ensure that *your* data is successfully leveraged using methods (*your* data strategy) that *your* knowledge workers and *your* executives understand.

The goal is to improve DG effectiveness and efficiencies (and the data itself) over time. The more data literate the organization, the easier the transformation. Perhaps now the phrase quoted at the beginning of the chapter is more understood (see Fig. 1.11).

Fig. 1.11 This database ain't big enough for the two of us – Bumper sticker seen on an automobile in Texas

Acknowledgments My colleague Rob Greaves made many helpful suggestions that were incorporated into this chapter.

[14] https://en.wikipedia.org/wiki/Bespoke

Chapter 2
Data Strategy and Policies: The Role of Data Governance in Data Ecosystems

Dominik Lis, Joshua Gelhaar, and Boris Otto

2.1 Introduction

The importance of data in the digital age is undisputed. The potential of creating value with data is evident from a multitude of success stories in all domains. The perception of data as an enabler of novel business models and data-driven innovations has changed fundamentally as a result, which is why the significance of data for companies as a strategic asset has grown strongly. During this development, data governance has a prioritized role within the formulation of data strategies, as it provides a mandate to organize data and information in a targeted manner [1].

In order to operate successfully and sustainably in the market and use data to create value, companies need to define and design a data strategy with a clear vision along with the internal capabilities required to successfully implement the data strategy. To implement and operationalize this data strategy within an organization, a data governance framework is needed that defines, implements, and monitors data policies, for example, in the form of processes and standards. This triad of data strategy, data policies, and data governance is a continuous process that must be regularly reviewed and adapted.

A data governance framework includes norms and data standards, which may result from legal or organizational requirements, methods, and standards to ensure the ongoing evaluation and further development of the data strategy, concrete policies for managing the data life cycle, and the structure of the data organization in the form of responsibilities within the organization [2, 3]. Integrating data governance principles within the data strategy ensures consistent management of data across the organization. At the same time, data governance provides the

D. Lis (✉) · J. Gelhaar · B. Otto
Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany
e-mail: dominik.lis@isst.fraunhofer.de; joshua.gelhaar@isst.fraunhofer.de; boris.otto@isst.fraunhofer.de

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_2

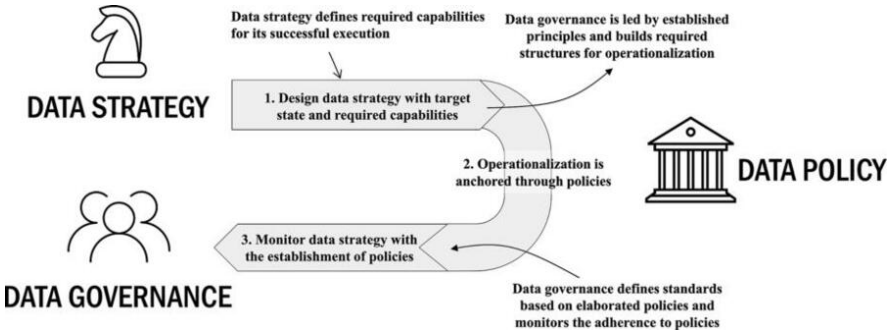

Fig. 2.1 The interplay of data strategy, data policies, and data governance

necessary rigor when changes result from the context of the data strategy for the organization [3]. Figure 2.1 gives an overview of interplay between the three activities.

In addition to the internal organizational challenges of implementing data governance, the range of new external challenges is growing, which in turn increases the radius of data governance. For example, there is a great need for data from industrial and production-related environments for data-driven optimization of production processes in the context of Industry 4.0. Another factor is the consideration of data governance in the inter-organizational environment, e.g., for sharing data with third parties in ecosystems, which today is conducted highly static due to restrictions or other uncertainties. In both scenarios, internal and, more recently, external influencing factors must be taken into account when designing a data strategy. The latter represents a new and relatively unexplored scenario for the data governance body of knowledge.

Therefore, the objective of this chapter is to bridge the traditional perception of data strategy and policy with a novel perspective on data governance due to the emergence of data ecosystems. This chapter provides insights into practical issues and describes the growing amount of external contextual factors, which affect existing data governance frameworks. This chapter ends with recommendations on how organizations can position themselves to utilize data ecosystems beneficially as part of their strategic directive for data.

2.2 Data Strategy and Policies

The management of data has been a subject of scholarly research and practical application since the advent of databases and application systems in the early 1980s. The significance of data in organizations has undergone significant changes over time, resulting in the development of a substantial body of knowledge in this

Table 2.1 The evolution of data management in organizations based on [16]

	Phase 1: administration-centric	Phase 2: quality-centric	Phase 3: value-centric	Phase 4: collaboration-centric
Main focus	• Application development • Automation in business functions	• Organization-wide business processes • Decision support • Reporting	• Advanced analytics • Data-driven business models	• Data sharing • Inter-organizational data life cycle transparency • Data-driven innovation
Data resources	• Structured data • Databases for automated data processing in organization functions	• Integrated information systems • Enterprise resource planning systems • Computer-integrated manufacturing • Data warehouses • Business intelligence	• Unstructured big data • Data lake architectures • Data analytics pipelines • Connected information systems	• Data products • IoT data • Digital twin • Digital platforms • Open data
Data-related concerns	• Data model quality • Data availability and reuse	• Organization-wide data integration • Data quality • Process management • Compliance	• Business value • Data privacy and security • Data architecture	• Trust • Data ethics • Data sovereignty • Data ownership
Management approach	• Database management	• Resource management, quality management	• Strategic management	• Ecosystem management
Data governance	• Governance of data conducted implicitly as part of IT governance and database administration	• Increase of approaches and adaptions for data governance • Data governance as mechanism to comply with regulations or support business processes	• Incorporating of data governance structure in organizations • Adaptions from data governance to the concept of digital platforms	• Data ecosystem governance • Enforcement of sustainable data-driven collaborations

field. The strategic utilization of data anchored in the form of a formal data strategy is becoming pivotal to digitalization.

The evolution toward the level of strategic utilization of data has occurred in distinct phases as every phase entails characteristic technological advancements and changes that impacted the role of how data was perceived and managed. Table 2.1 provides an overview of these phases.

The first phase is mainly characterized by the management of data through administering database systems. The focal area of operations has been data

processing in centrally managed enterprise systems [4]. The next phase of data management has experienced a fundamental shift with advancement in the development of databases and database software. In the 1990s, the focus has moved increasingly from a pure functional domain perspective to end-to-end business processes covering multiple functions. Computer-integrated manufacturing (CIM) and enterprise resource planning (ERP) systems exemplified this concept, supporting the integration and shared use of data across operational and administrative processes. It was increasingly recognized that the traditional understanding of data administration and focus on single databases must move toward reflecting data as a resource at organizational level, which has led to the emergence of data resource management. The field of data resource management further promoted the improvement of data management as an organization-wide instrument for data planning, enforcement of policies, as well as technical functions.

Gradually, a more strategic approach has been adopted for data by incorporating established practices from the management of tangible resources from the discipline of total quality management [5]. Data quality became a primary concern and effective way to leverage data for the improvement of business processes, supply chains, customer relationships, operations, and reporting. The body of data management-related knowledge further evolved from a database-centric perspective to encompass organizational and technical capabilities, particularly pertaining to organization-wide data integration, data architecture, and data governance [6, 7].

A third phase of data management in organizations began in the 2010s with the use of larger volumes of internal and external data (big data) and the emergence of digital business models and data-driven services [8–10]. These developments emphasize the business value and impacts of data [11, 12]. The strategic role of data is reflected in additions to the data management-related knowledge base: the technological and organizational capabilities to acquire, store, and process the increasing variety and volume of data, based on data lakes and advanced analytics platforms [13–15]. Data management is also increasingly associated with strategic capabilities to enable data monetization by improving business processes and decision-making or by innovating business models [10, 11].

In sum, the role of data has evolved from an enabling resource to a strategic one. In response, data management has developed from a technological capability focused on single databases to an enterprise-wide organizational and strategic capability. This development is mirrored in the accumulation of data management-related knowledge, which required substantial adaptation and extension to cope with the evolving roles of data in businesses over time.

This chapter focuses on the aspect of the latter phases and provides relevant development strands emerging from data ecosystems, which need to be considered in the design and implementation of data strategies and data governance.

2.2.1 Data Strategy Fundamentals

For companies to have the ability to use data to their advantage and remain competitive in the long term, they need a comprehensive data strategy that forms the basis for the optimal use of their data as strategic assets [17].

For data strategies to materialize, the cultivation of three fundamental capabilities must be prioritized. First, relevant data assets must be identified and prioritized which must be organized and managed accordingly. In a second step, this data must be examined analytically. Last, the organization is able to make data-based decisions [18].

There is no consensus definition of the term data strategy in the research community. Table 2.2 provides a short overview of selected definitions.

One approach to defining a data strategy involves a detailed specification of its distinctive components. This can be achieved by aligning it with the five elements of strategy, namely, plan, ploy, pattern, position, and perspective [19], which can be applied to the notion of data assets. A data strategy can be understood as a reference of methods, services, architectures, usage patterns, and procedures along the data life cycle. It forms the basis for the digital transformation of organizations by setting a target vision and defining action steps to achieve it [2, 18].

In this regard, a data strategy promotes the governance and management of data as a corporate asset, which is applied to business decisions at all levels and thus enables a significantly higher state of digital maturity for an organization. A data strategy includes key performance indicators and success criteria to ensure measurability of the defined goals. Furthermore, strong sponsorship and governance by the organization's management are required to maximize the potential of the data strategy. Ideally, a data strategy forms an overarching umbrella for individual data management initiatives within companies, including a framework for data sharing with external parties. The definition of a data strategy should include a road map that aligns individual initiatives to achieve the most value from data [3].

Table 2.2 Definitions of data strategy

Definition	Source
"A data strategy is a common reference of methods, services, architectures, usage patterns and procedures for acquiring, integrating, storing, securing, managing, monitoring, analyzing, consuming, and operationalizing data. It is, in effect, a checklist for developing a roadmap toward the transformation journey that companies are actively pursuing as part of their modernization efforts."	[17]
A data strategy "defines the scope and objectives of data management and specifies the roadmap for providing the data management capabilities required."	[18]
"A data strategy establishes common methods, practices and processes to manage, manipulate and share data across the enterprise in a repeatable manner."	[19]
"A modern data strategy is a roadmap to enable data-driven decision-making and applications that helps an enterprise achieve its strategic imperatives. An effective data strategy helps an enterprise make technology choices, grounded in business priorities, to get the most value from their data."	[20]

In sum, six central characteristics can be consolidated from literature that sum up core activities for a data strategy [20]. A data strategy should include extracted core elements of a data strategy from existing elaborations:

- Clear vision, mission, and business objective alignment
- Long-term benefits and competitive advantage
- Constitution of a road map and objectives
- Organizational and technological assessment and change management
- Long-term and organization-wide data strategy establishment
- Set boundaries and objectives for data management

2.2.2 From Defensive to Offensive Data Strategy

A common approach for distinguishing data strategy approaches is through a defensive and an offensive perspective [1]. Accordingly, companies can target a more controlled or more flexible use of corporate data against the background of their business environment. In this context, the two strategy approaches differ in terms of their objectives and activities. The focus on a defensive strategy includes activities for compliance with regulations and the security and protection of data. It also addresses the management of sensitive and business-relevant data in a single source of truth (SSOT). A stronger commitment to an offense data strategy, on the other hand, seeks to support the achievement of business goals and accordingly includes activities such as analytics on customers and market data, as applied predominantly in sales and marketing [1].

A sound data strategy ensures that the data available in a single source of truth (SSOT) is standardized and of high quality and that variations of this data in the form of multiple versions of truth are transparently derived from the SSOT and adequately controlled, which is why data governance must be comprehensively considered. In this respect, companies must consider both defensive and offensive aspects, but the focus can vary substantially. In this instance, the stronger focus on one aspect often results from the business environment. At the same time, the regulatory treatment of structured and standardized data is easier to manage, whereas flexible, easily transformable data is particularly useful in offensive applications [1] (Table 2.3).

2.2.3 Data Policies

According to the Data Management Body of Knowledge (DAMA-DMBOK), a policy is a "statement of a selected course of action and high-level description of desired behavior to achieve a set of goals" [21]. Policies are a consolidation of principles that reflect as processes, standards, or controls in business operations.

Table 2.3 Characteristics of a defensive and offensive data strategy approach [1]

	Defense	Offense
Key objectives	Ensure data security, privacy, integrity, quality, regulatory compliance, and governance	Improve competitive position and profitability
Core activities	Optimize data extraction, standardization, storage, and access	Optimize data analytics, modeling, visualization, transformation, and enrichment
Data management orientation	Control	Flexibility
Enabling architecture	Single source of truth	Multiple versions of the truth

Data policies are essential instruments for ensuring commitment to an overall data strategy and for shaping an organization's overarching self-perception for data [21].

Data policies play a crucial role in data governance programs for establishing consistency and structure and for enabling a sophisticated management of data. They make a significant contribution in anchoring a formal and strategic approach for the management of data. The definition of standards and guidelines promotes the improvement of the accuracy and reliability of data, resulting in more trust in data and a better foundation for decision-making [22].

A data policy serves as a strategic signal to all stakeholders as it assists in driving the communication in change management initiatives. Besides its purpose as a means of communication, a data policy can act as leverage for the allocation of resources required for the transformation toward becoming a data-driven organization. The main purpose is to emphasize the importance of data as a strategic asset and provide transparency about the value data has for an organization. Having a clear data policy in place can also facilitate data sharing provide incentives for collaboration between departments.

The focal areas of data policies may differ depending on the maturity and prioritized strategic directive of organizations. The most persistent building blocks of data policies include the protection of sensitive data, improvement of data quality, complying with regulatory demands, maintaining data security, or managing the data life cycle. It is common to establish multiple function or domain-specific policies such as policies for data quality management or distinct data security policies where procedures and standards have matured. Additionally, policies contain the logic of the organizational structures applied to the governance data, e.g., through the allocation of authority, description of roles and responsibilities, and establishment of data committees or working groups.

For many years, adhering to regulatory compliance that impacted the management of data has been a dominant factor in establishing some form of data governance in the organization.

Despite the long-term and strategic purpose of data policies, they are subject to an audit process for continuous improvement and fine-tuning. As digitalization

progresses and new challenges evolve, data policies must address strategic alignments in the scope of data management. Policies are increasingly being adapted because their scope can no longer keep up with the new development strands of the data economy such as data monetization, inter-organizational data sharing, or artificial intelligence. The consideration and governance of analytical and highly dynamic data pipelines or data sharing across organizational boundaries are application scenarios growing in frequency but have not been deliberately elaborated in the context of data policies. In this regard, future data policies can simplify the facilitation of data sharing and act as a seal of approval between parties to certify the adequate management of data.

2.3 New Development Trajectories for Data Governance

2.3.1 Data as Strategic Asset for Organizations

The function, perception, and characteristics of data for companies have been constantly changing over the last decades and have led to changing factors influencing the data governance of companies [23]. The success of digital platforms and the increasing end-customer orientation of many business sectors are just two examples of developments that require companies to rethink how they handle data. This concerns both internal data management and the cooperation with external partners. When it comes to the relevance of data for companies, a distinction can be made between four different types of functions (Table 2.4). First, data is still, and has been for the last decades, a source of business process improvement. The integration and automation of business processes requires effective and efficient data governance and management. Second, data is increasingly a source of business innovation [24]. Data-based services in different industries require access to and combination of data from various sources. These data sources can be both internal and external to the organization, e.g., from suppliers or customers. For example, original equipment manufacturers (OEMs) are increasingly cooperating with their business partners, component manufacturers, or service providers to provide better end-to-end services to their customers. Third, data itself has become a product that needs to be managed and governed like any other product so that it can then be traded and sold on, e.g., data marketplaces. For example, mobile network operators sell anonymized data about the behavior and movements of their customers. Traffic authorities, for example, can analyze this data and use the information obtained to maintain and improve the traffic infrastructure. And fourth, data is increasingly seen as a strategic resource for the long-term sustainability of the economy. For example, the European Union estimates that the data economy will be worth at least €550 billion by 2025 [25].

However, this value can only be achieved if data is shared and used [26]. Against this background, politics, science, and the private sector have a great interest in increasing the sharing and joint use of data. Industrial companies are sitting on a

Table 2.4 The different roles of data for businesses

Function of data	Description
Data as a source of business process optimization	Data quality as a prerequisite for automated and integrated business processes Management of digital twins along the entire value chain and over the entire life cycle (e.g., of products and plants) Integration of digital factory concepts with supply chain management
Data to enable digital business models	Necessity of combining own data (e.g., on products, plants, customers, etc.), data from business partners, and contextual data End-to-end support of customer processes based on shared databases or data models
Monetization of data in ecosystems	Ecosystems as a new multilateral organizational form for creating customer innovation Data as a platform resource in ecosystems Revenue and benefit potentials for data providers and data users
Data as an economic resource	Strategic resource in the platform economy Data as basis for (data-driven) innovation Data sovereignty and fair handling of data as the core of the European and German data strategy Demand for national or European data infrastructures (cf. Gaia-X)

"hidden treasure" of data, which is created, for example, by manufacturing processes or through the use of products by customers [27]. However, data holders also have an interest in ensuring that the data they share is not misused and that they are paid appropriately. After all, offering and sharing data, especially high-quality data, generates costs. Therefore, appropriate governance mechanisms need to be defined, which, for example, incentivize data holders to offer their data in the ecosystem and ensure sovereignty over their shared data [28].

2.3.2 The Emergence of Data Ecosystems

In addition to the shift in the importance of data for businesses and the shift from tangible to smart products described above, there is another fundamental change in the digitalized economy. Innovation is increasingly taking place in so-called **ecosystems**, in which different actors such as companies, research institutions, intermediaries, government institutions, customers, and competitors join forces to create innovative value propositions [29]. Ecosystems are characterized, among other things, by the fact that no single member can create innovations on its own, but that the ecosystem must work together as a whole [30].

Originally, the term ecosystem comes from the field of biology, where it is used to describe interactions between organisms of different species and their environment

interrelated system. Since then, there have been various research areas that have applied the characteristics and properties of the ecosystem concept to their field of interest. One of the well-known areas of application comes from the field of business administration, where [31] introduced the concept of business ecosystems [31]. A business ecosystem is defined as a community consisting of companies, producers, suppliers, and other actors that cooperate to achieve a common goal, such as the creation of an innovative product or service. Building on this preliminary work, further fields of application of the ecosystem concept have been identified in the context of the data economy, describing interactions between a wide variety of actors cooperating in the construction or manipulation of a shared resource (e.g., service, software, or platform). A special form of these digital ecosystems are data ecosystems, in which data is the strategic resource of the ecosystem, which is exchanged, shared, (re)used, and monetized between the actors [32]. Consequently, a data ecosystem can consist of various actors, such as companies, research institutions, or private individuals, who perform different data-specific functions in the ecosystem, for instance, data provision, data exchange, data processing, or data use [33]. The various activities of the individual members in a data ecosystem essentially lead to a complete coverage of the data value chain. Each individual member must contribute in order to benefit, as ecosystems only function in the long term if they can create a state of equilibrium of mutual benefit for all members [34]. Participation in data ecosystems offers new growth opportunities for the participating actors through networking with other participants and acts as a driver for innovative services and customer experiences. The sharing of data opens new opportunities for progress and the formation of cooperations with other companies or actors, from which every participant in the data ecosystem benefits. Through the sustainable exchange of data, the participating actors can develop further and engage in value creation cooperation that leads to new digital value propositions.

2.4 Widening the Scope of Data Governance Operations

Despite an increased awareness of the relevance of data for data-driven value creation and the motivation to fully utilize the potential data, the necessary structures and corresponding competences are often not available in companies or have only recently been developed [35, 36]. This lack of consideration for data governance within the organization can manifest in various ways, such as:

- The implementation of hasty data initiatives to improve data quality without a sustainable approach
- The lack of initiatives elaborating opportunities for exploitation of/with data in the sense of data-driven products and services
- The prolonged search for necessary data/information and the appropriate contacts
- The waste incurred from duplication of work and repetition of tedious data maintenance actions

- The lack of communication and discontinuities of information throughout the data life cycle
- The emergence of uncoordinated silos and different semantic understandings of data between departments and/or business units

It is important for organizations to recognize the significance of data governance as part of the data strategy and policies to effectively leverage data for value creation and to avoid the aforementioned symptoms that impede data-driven innovation.

In practice, it can be observed that the transformation to a data-driven organization is progressing only slowly as a reactive approach for the management of data often prevails throughout many industries. Positive effects from implemented projects to combat the challenges, for example, through targeted projects for improving data quality of master data, are often only short-lived because data-centric responsibilities are not anchored in the organization. Therefore, during the development toward a data-driven organization, data governance is ideally accompanied by effective communication measures as depicted in the following chapters.

Fortunately, a clear course can be seen in practice. In numerous companies, data is increasingly being placed on the strategic path with clear visibility. Additionally, initiatives are being launched to target the strategic utilization of data with the required structural foundations of data governance and policies. The linking of data governance as a lever for data-driven innovation is also triggering a new paradigm shift, whereby the image of data governance as a pure compliance and master data topic is being slowly overtaken by the reality seen in practice [37].

2.4.1 Consideration of Challenging External Influencing Factors

The data economy entails novel development trajectories that need to be considered in the governance of data, e.g., diversity and velocity of data, data monetization, or inter-organizational data sharing. Additionally, companies must cope with a highly dynamic and growing regulatory landscape. In the last few years, the European Commission has adopted several new regulations that have an immediate impact on the implementation of companies' digital business models. In addition to the already established General Data Protection Regulation, the recently developed and adopted regulations such as Data Governance Act, Data Act, Artificial Intelligence Act, Digital Services Act, or Digital Markets Act will have to be considered and in line with business operations soon as they trigger the implementation of further measures for the management of data in the private and public sector [38].

This is just one of many novel development trajectories, which require companies to continuously improve their data-related capabilities to reach a maturity level that allows them to realize innovative value creation opportunities with data.

In this context, the role of data governance as an instrument for establishing and monitoring a data strategy is becoming increasingly vital. The strategic constituents

Table 2.5 Challenges arising from the data economy affecting the governance of data

Perspective	Influencing factors and challenges
Data	• Complex and dynamic data landscapes consisting of static master data and dynamic streaming data from IoT applications • Previously only internal data must be processed and shared with ecosystem partners • Data shared by external partners must be included in the internal systems
Technology	• Variety of tool options • Advanced analytics capabilities • Data lake architectures • Complex data pipelines to capture data from the field • Emerging new technologies for sovereign and secure data sharing
People	• Raise awareness among employees about the importance of data to create a data mindset • Cultural shift toward considering data as a resource • Management support to invest in new technologies needed for successful participation in data ecosystems • Enabling employees to handle data properly
Processes	• Increasing requirements from the business or from the shop floor • The implementation of data governance in complex organizational structures • Business and IT processes must increasingly be aligned and optimized together
Market	• The transformation from traditional engineering-driven value creation to data-driven services • Managing dominant cloud data platforms • The increasing need for networking with external partners in so-called data ecosystems • Grand challenges such as circular economy and sustainability cannot be solved by one organization alone
Service	• The operationalization of hybrid data-driven business models • To create data-driven services, data from various internal and external sources must be combined
Regulatory	• Increase in regulatory demands with impact on the governance and management of data

of companies must take far-reaching influencing factors and application scenarios in the context of data governance into consideration, which evolve from cross-organization data sharing or using data in digital or hybrid business models.

For an organization to realize new opportunities for value creation based on data, it is becoming increasingly essential to develop awareness about the relevance of data, to achieve the required maturity in managing data, and to look beyond the internal data landscape for value creation opportunities.

To fully capitalize on the opportunities presented by data-driven value creation, companies must not only address common requirements but also the new presented conditions arising from digitalization. This means that data governance will continue to be a crucial instrument for companies to comply with regulatory guidelines for managing business processes in the administrative and planning environment. As depicted in Table 2.5, the new trends and developments come with challenging tasks, which transcend the traditionally perceived remit of data governance. In

addition to existing challenges in managing business operations and data, the range of novel challenges is expanding, thereby increasing the scope and authority of data governance.

2.4.2 *Bridging the Intra-organizational Perspective on Data Governance with the Inter-organizational Perspective*

Another factor is the consideration of data governance in the inter-organizational environment, such as in the exchange of data with third parties (data sharing), which today is only possible under very strict restrictions or is not pursued due to further uncertainties. This external view represents a novel and relatively unexplored scenario for the topic of data governance because this development breaks organizational boundaries as internal data sources are increasingly utilized externally and vice versa. Organizations must identify an equilibrium between the opposing interests of maintaining control over their data assets and willingness to share data for the development of common value propositions [39].

To understand the implications arising from inter-organizational data sharing, an initial distinction between an internal and external perspective on data governance is essential. Most of the body of knowledge on data governance explores data governance practices from within a single organization, focusing on topics associated with organizational structures, data quality, processes, guidelines, or tools [40–42].

The intra-organizational perspective on data governance constitutes a significant portion of the current academic and practical discourse on data governance. However, the link between this perspective and the external intra-organizational perspective remains insufficiently investigated [36, 43]. It is widely established that from an internal viewpoint, data governance manifests itself within organizational structures and hierarchies, ensuring that principles, decision rights, and guidelines related to data assets are effectively implemented and monitored [44].

However, the use of these traditional instruments for data access and use is often limited to the bounds of a single organization, and thus, the influence of authority in inter-organizational constellations may be limited [45]. To bridge these two perspectives, the inter-organizational perspective on data governance includes novel factors that must be considered in the formulation of data strategies and policies. In data ecosystems, where the provision of data from multiple actors is critical, it is imperative to examine the governance mechanisms that foster a collaborative and trustful environment for all actors involved [45].

Initial ideas in advancing the knowledge toward an external perspective have recently begun to examine data governance in the case of digital platforms as data sharing between organizations often revolves around platform-based technical infrastructure [45–47].

In this regard, the focus lies on the combination of data governance and platform success. Despite the growing interest in inter-organizational data sharing, the

Table 2.6 Differentiation between intra- and inter-organizational data governance characteristics [49]

Characteristic	Intra-organizational data governance	Inter-organizational data governance
Scope	• Internal (within an organization, e.g., departments and business areas)	• External, between organizations or ecosystem (e.g., platform, business partner, customer)
Purpose	• Ensure the provision of decision rights and accountabilities for the management and use of data • Set up organizational structures and use governance mechanisms to improve data quality, manage resources across a single organization, and formalize guidelines for data resources	• Establishment of governance mechanisms that foster collaboration between multiple entities • Facilitate data sharing under consideration of data ownership, access, integration, and usage • Ensuring that each participant contributes to pursuing common goals and value propositions
Goals	• Establish strategic importance of data as an asset on corporate level • Maximize the value of data for the organization by improving the quality of decision-making • Establishment of clearly designated roles for data elements	• Creation of an ecosystem with aligned balance of control and authority to incentive data sharing and value creation among actors • Adherence to fair overarching rules that protect the interests of ecosystem partners while overcoming conflicts
Roles and organization	• Designated data roles, councils, or committees within the organization, e.g., data owner, data steward, chief data officer • Organization anchored within hierarchal structures of the organization	• Depending on the activities, an organization can embrace different roles, e.g., data provider, data broker, infrastructure provider • Different modes of organization are possible depending on the conceptualization of the ecosystem in technical or sociotechnical aspects
Governance instruments	• Structural, procedural, relational mechanisms manifested within the organization	• Regulatory instruments, licenses, formal contract-based agreements, technical measures for data integration and usage policies, data sharing agreements

implications for organizations engaging in data ecosystems have yet to be fully analyzed. The differentiation between the internal and external perspectives on data governance is a crucial factor in improving the understanding of the challenges associated in inter-organizational data sharing, as the range of authority for traditional (internal) governance instruments may be limited in the context of data ecosystems [48]. Table 2.6 provides a comparison of the main characteristics between intra-organizational data governance and inter-organizational data governance.

2.5 Utilizing Data Ecosystems as Part of Data Strategy

Despite the competitive nature of organizational relations, there has been a growing trend toward data-centric collaborations, in which organizations utilize and provide access to distributed data sources. Over time, these relationships have evolved from simple dyadic interactions to the emergence of complex ecosystem structures. These ecosystems are comprised of multiple autonomous organizations that engage in data sharing to leverage data more effectively. For value propositions based on data to be realized, the configuration of data governance can play a crucial role in influencing the design, dynamics, and success of these collaborations. However, in the context of data ecosystems, the conceptual understanding of data governance is not fully explored and integrated as part of data strategies. The paradigm shift toward considering the significance of data as a strategic resource and the external view that considers inter-organizational data sharing are phenomena that just begin to gain practical and research attention in the context of data governance.

2.5.1 The Role of Ecosystem Data Governance

Most research and practical contributions in the field of data governance have primarily focused on the analysis of single entities, specifically the design and implementation of organizational structures to enhance data quality and manage data-related resources across the organization [36]. The body of knowledge on the internal reflection of data governance is extensive and provides valuable materials in the form of practical frameworks and data governance tools, which promote desirable behavior and conduct through policies. However, when it comes to the utilization and sharing of external data with third parties, data governance enters a gray zone with many unresolved issues. For instance, the dynamics within ecosystems are more complicated and diverse because value creation processes, governance, and ownership structures over data become less transparent [39]. The lack of consensus regarding data governance in intra-organizational settings can therefore lead to uncertainties about who can use which data for what purpose. Hence, in the context of data ecosystems, the allocation of decision-making rights and responsibilities that promote desirable behaviors in relation to intangible assets becomes increasingly ambiguous [50].

Today, much of the arrangements take place in digital platform or cloud infrastructures [40, 46, 47], where data governance is associated with a focal key actor and mechanisms enforcing governance to its ecosystem [47, 51]. While data governance from an intra-organizational perspective typically implies hierarchical structures and a controllable organizational environment, structural arrangements regarding data in ecosystems can result in conflicts of interest between participating organizations [52, 53]. In this context, the role of ecosystem data governance is to establish a collaborative environment that facilitates data sharing among

organizations by implementing coordination mechanisms to align the interests and collective goals of participants. **Ecosystem data governance** can be defined as an arrangement of institutions and structures with the objective of assuring that individual organizations behave in coherence with collective intentions by establishing common set of rules that allow for an effective and fair utilization of data within the inter-organizational collaboration [39, 52].

Data ecosystems underline the necessity to bridge internal policies for data with mechanisms that can be transferred beyond organizational borders to provide clarity in cultivating novel forms of collaboration. The possibilities arising from external data in the form of data monetization opportunities, e.g., through the development of data-driven business models, often require an extensive scope of data collection, which makes it imperative to engage in data-centric collaborations with mutually agreed terms. We therefore emphasize the necessity to extend the body of knowledge of data governance beyond the organizational sphere.

2.5.2 Inter-organizational Data Governance Modes

The literature identifies distinctive patterns that can be applied to practical scenarios. The configuration of a data ecosystem can determine how collaborations function and evolve and to which degree decision-making authority over data can be executed. Dominant actors such as platform owners possess the ability to control access and interactions within their technical infrastructure, constituting to the concept known as *lead governance*, where a single organization acts as a centralized entity that coordinates essential network maintenance and decision-making processes. In contrast, the more decentral approach, known as *shared governance*, exists in settings where all organizations govern the ecosystem equally without formal governance structures. A further distinction can be made between ecosystems governed by participants themselves and those governed by a separate entity, serving only as a coordinator. This form of governance is referred to as network administrative organization (NAO) and has a purely administrative function requiring a neutral stance, in which the factors trust, size, goal consensus, and competencies serve as critical attributes for the effectiveness of the collaboration [51].

The concept of data governance in ecosystems can be applied to the established understanding of generic governance modes of market, hierarchy, network, and bazaar, which encompass various overarching arrangements and incentives for control. These regimes can be adapted to interpret inter-organizational data collaborations in ecosystems, each exhibiting distinct characteristics and coordination mechanisms [52, 54, 55]. The governance mode market is characterized by strict compliance through contractual terms for property rights with a low level of trust as every interaction (data sharing) can be managed through contractual agreements. A central coordination mechanism in the market mode is pricing [52, 56]. In the context of data ecosystems, market-based arrangements are associated with data

marketplaces, where relationships between buyers (data consumers) and sellers (data providers) are based on market forces [52].

The hierarchy governance mode, on the other hand, enforces control through the administrative authority of a dominant actor, who orchestrates formal procedures and decisions for the coordination of individual actors [56, 57]. This mode is visible in supply chain networks where data exchange is managed by dominant actors or in platform settings, where owners of the technical platform infrastructure have control over the partnership hierarchy of complementors [58, 59].

The network mode of governance represents a hybrid arrangement, characterized by interdependent capabilities and collaboration based on reciprocity, collective goals and benefits, and trust. Networks evolve through the establishment of relationships and trust naturally over time, which, if required, provides a solid basis for the facilitation and transition to more formal structures [57]. Decision-making and coordination in this mode are conducted jointly to reach consensus. This mode is the closest to the underlying idea of data ecosystems with multilateral data sharing and alliance-driven data collaborations. They are conducted jointly to reach consensus. The network governance mode shares similarities with multilateral data sharing in data ecosystems or alliance-driven engagements to enable data collaborations [49].

The bazaar governance was introduced with the emergence of the open source movement, characterized by open licenses and engagements driven by the willingness to distribute information or by intrinsic motivation for better reputation [54] (Table 2.7).

This mode has been successfully established in various settings of open data initiatives in the public sector, which are aimed at fostering innovation through the provision of free access to data [60].

The presented types of engagement and occurring regimes demonstrate that organizations lay the foundations internally for successfully engaging in inter-organizational data sharing. This includes knowing which data is existent and relevant within the organization; who is responsible or can provide information related to these data assets; how the data is used (both internally and externally); and under which conditions data can be shared with whom and where. These new external aspects exceed and challenge traditional tasks and responsibilities of dedicated data roles within the intra-organizational sphere because data can also be in control of external entities. Figure 2.2 provides an example of an organization that targets a central positioning in a data ecosystem by engaging in a mode that constitutes to the characteristics of the mode *hierarchy*. In this example, the organization is an original equipment manufacturer in the automotive industry. The strategic decision regarding the ecosystem of the organization includes an active management of the ecosystem and relevant data for a seamless production process. To achieve this, the OEM provides the IT infrastructure in the form of a data platform for all actors involved to share data and information. The OEM also considers the option of providing the technical infrastructure and acts as intermediary between the provider and consumer of data. Regarding the data governance options in this exemplary case, different mechanisms for the design and control of

Table 2.7 Attributes of governance regimes adapted to inter-organizational use of data [52, 54, 55]

Attributes	Market	Hierarchy	Network	Bazaar
Nature of data sharing	Data sharing on a contractual basis	Data sharing through dominant actors	Data sharing for collective targets with trusted actors	Open and unrestricted data sharing
Equivalent within data economy	Data marketplaces or data intermediaries	Data platform with platform owner who retains full control of the technical infrastructure	Multilateral data sharing in data ecosystems	Open data portals
Normative basis	Contracts	Formal hierarchy	Social contracts	Open license
Incentives for engagement	Competition	Market share, status	Trust, common objectives	Reputation, data access
Control over incentives	Moderate due to contracts	High through administrative power	Moderate through reciprocity and social contracts	Low based on reputation in the community
Reasons for adaption	High flexibility for participants; decreasing coordination costs	Negotiation position; strategic differentiation	Low-cost access to resources; common value propositions	Innovation and low coordination costs
Flexibility of the collaboration	High	Low	Moderate	High
Duration of the collaboration	Short term	Unlimited	Long term	Unlimited
Relation between network members	Independent	Dependent	Independent	Independent

the platform can be exercised as the technical infrastructure is provided by the OEM itself. From an internal perspective, the organization considers the development of data-driven services from the data utilized in the field. This requires changes in the internal role structures as teams increasingly work across functional domains to ensure standards in the logic and semantics persistent to the whole organization.

2.5.3 Adequate Positioning for Engaging in Data Ecosystems

The previous section demonstrates that an organization has different design and utilization options for data and type of engagement in data ecosystems. The following section emphasizes which specific role and function a single organization can execute based on existing capabilities.

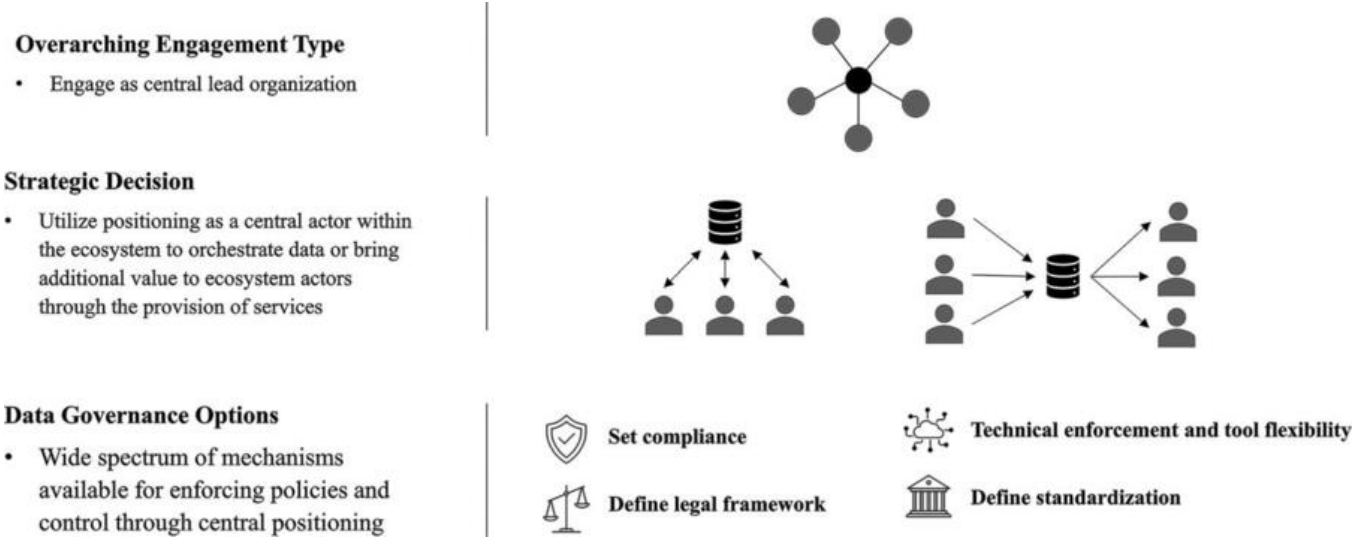

Fig. 2.2 Exemplary design choices for the role of a lead organization in data ecosystems

A data ecosystem involves a variety of different actors who differ in their capabilities and their contribution to the ecosystem. Depending on the level of engagement, an ecosystem participant may take on a specific role or function [61]. To achieve common values, these roles and functionalities are linked to tasks and activities. The naming and occupation of roles differ in scientific publications and practice-oriented initiatives. However, a general overview can be gained from the derivation of roles that are necessary for the creation of an ecosystem. In general, a data ecosystem needs actors who make data available. These actors are called data providers. A data provider publishes data that can be used by various other participants (data consumers). Data consumers use data, for example, to extend and enrich existing services with additional data, e.g., for analyses. In addition to data providers and data consumers, depending on the scenario, other actors are often named that may be necessary for a holistic and decentralized network organization. We summarize these here as data intermediaries. For example, data intermediaries may be necessary to support the establishment of the connection between data providers and data users, a so-called catalogue or broker, providers of the technical infrastructure such as platforms, or providers of services to carry out data-related activities such as preparation, analysis, and visualization of data. Other examples are services for the exchange and monetary settlement of data. In this context, all actors are involved with different activities such as the provision of data or services in the data ecosystem, depending on their role and perspective. At the core is the end customer, who benefits from new and innovative data-driven products and services enabled by the exchange of data from different partners [62].

For organizations to understand how to position and organize themselves in data ecosystems, it is essential to know which role to embrace and the implications arising from certain collaborations [39]. The following selection provides a generic overview of engagements that can be pursued by organizations. The engagement is not limited to one role.

- **Data Provider:** A data provider makes data available for sharing among participants in a data ecosystem. Data providers lay the foundation for successful participation in data ecosystems within the organization. To act as a data provider, an organization needs a precise overview of the existing data assets and which business models can be realized with these assets. Ideally, data providers can specify their data resources and apply valuation methods in terms of their value proposition. The development of pricing models for data requires a high level of maturity in the management and maintenance of data. The entire life cycle of data, from generation to provision on data marketplaces, depends on the support of adequate governance structures that allow having transparency over the relevant data assets. From an organizational perspective, these prerequisites should be considered in a data strategy. In addition, a data provider should analyze from a market perspective which platforms are suitable for its data products. It is possible that data providers will divide their resources and use different platforms to meet the needs of specific data consumers. As more and more platforms will enter the market, data providers need to select trusted platforms with the appropriate

technical infrastructure for the relevant data domain. Initiatives such as those of the International Data Spaces and Gaia-X support the necessary measures to allow individuals and legal entities to determine the use of their own data resources.

- **Data Consumer:** In the area of data ecosystems, the transparency of existing datasets in platforms can be limited. It is important to build a technical infrastructure that allows potential data consumers to make queries to search existing datasets. Therefore, a data consumer needs to be able to search the datasets provided by different data providers. Once the data consumer has identified the data suitable for his or her purpose, a connection must be established between the data provider and the consumer. Metadata brokers and (federated) catalogues are examples that enable this data transaction on a secure basis. There are multiple scenarios in which data consumers can benefit from data sharing in data ecosystems. Companies need to re-evaluate their existing business models in terms of their digital capabilities. This includes, on the one hand, knowing what data is available and, on the other hand, understanding what data is required to extend and increase the value of products or services. However, all stakeholders need to overcome the trust barrier by building on a trusted and agreed technical infrastructure where a data consumer respects the terms of use set by the data owner.
- **Data Intermediary:** Data intermediaries may foster data reuse, thus facilitating efficiency and innovation. Providers of data sharing services (data intermediaries) are expected to play a key role in the data economy, as a tool to facilitate the aggregation and exchange of substantial amounts of relevant data. Data intermediaries offer services that connect the different actors having the potential to contribute to the efficient pooling of data as well as to the facilitation of bilateral data sharing. Specialized data intermediaries that are independent from both data holders and data users can have a facilitating role in the emergence of new data-driven ecosystems independent from any player with a significant degree of market power. In addition, organizations can strategically decide to position themselves in the market as providers of digital trusted platforms. When designing the platform, the right governance mechanisms should be established to manage the complexity, control, and growth that come from having multiple parties from different business units involved in the platform. Consequently, it is necessary to find a suitable platform architecture that regulates the governance issues between all parties involved. On the one hand, platform providers need to be able to motivate data providers to share data, and on the other hand, data consumers need to find the right and high-quality data on the platform. All these aspects are reflected in the design and functionality of the platform. The goal is to create value-added connections between all stakeholders within the platform (Table 2.8).

In the future, it will be essential for organizations to understand which function they can engage in within data ecosystems to utilize data effectively. In practice, a clear trend can be seen in today's market activity: the rise of digital platforms, e.g., by original equipment manufacturers or providers of other essential technical

Table 2.8 Recommendations for actions for data ecosystem roles [29]

Data provider	Data consumer	Data intermediary
• Build up data capabilities • Identify business-relevant data resources • Elaborate on data-driven business models • Establish data governance on organizational and technical level • Find trustworthy platforms for providing data • Identify relevant partners in your own ecosystem for data sharing	• Identify relevant data resources to enhance existing business models • Combine various data sources to enrich data-driven services • Identify suitable providers of qualitative data • Find trustworthy platforms for acquiring data • Identify relevant partners in your own ecosystem for data sharing	• Establish trusted services and technologies to enable engagement of multiple actors • Know your governance mechanisms to manage engagements • Find balance between openness and control in the ecosystem design • Build trust by respecting security standards and sovereign data exchange

infrastructure (cloud), is increasing rapidly as means of consolidating as much data as possible from customers or partners. Organizations that are in control of the technical platform infrastructure usually direct the course of action and possess more control mechanisms for influencing user interaction and data management.

2.6 Recommendations for Action

This section will conclude the presented chapter with recommendations for individual organizations as well as for the design of data ecosystems as a whole. First, recommendations are described for individual organizations and how they can use the potential of cross-organizational data cooperation. This is followed by recommendations on the design of data ecosystems and which components need to be considered in cross-organizational data cooperation with third parties.

2.6.1 Recommendations for Actions for Single Organizations

For each organizational entity, it is imperative to understand the control mechanisms that can be implemented within these data-centric collaborative arrangements. Organizations can leverage both formal and relational elements to enforce governance mechanisms that impact the behavior and dynamics of the collaboration, such as through the utilization of incentives, rewards, or penalties [59].

Formal instruments are based on regulations and guidelines, which must be adhered to by participating organizations when sharing data within the ecosystem [63, 64]. Both formal and relational control strategies can be employed in platform-based ecosystems, where platform owners can enforce formal mechanisms such as contracts or certification, standards, and policies to encourage desirable behavior

among complementors within the ecosystem. Relation mechanisms are rooted in social norms and can be used to support and encourage appropriate behavior among ecosystem participants. Relational mechanisms are characterized by a collective and interdependent commitment of the ecosystem, e.g., for business models, and are anchored in trust and stable relationships [65].

Both instruments can be utilized to harmonize a set of formal and relational regulations, thereby allowing the two approaches to coexist and complement each other. The aspect of trust appears to be critical to the success of functional relationships within data ecosystems because the technological developments for enforcing trust through technical means have not experienced widespread exposure across industries and organizations are still reluctant about sharing a strategically relevant asset [66].

Figure 2.3 illustrates a conceptual model that visualizes the interdependency between established internal data governance structures and an inter-organizational data governance perspective to engage in data ecosystems. Depending on the role and type of engagement setting (data governance mode), organizations have the ability to exercise various governance mechanisms or must adhere to governance mechanisms to successfully engage in a data-driven collaboration and comply to data sovereignty demands. Within these constellations, organizations must be cognizant of the implications arising from these interactions as in some cases, the influence of authority and control within the data ecosystem might be limited. This is the case when one organization must adhere to the standards and regulations of external platforms through enforcement of entry barriers such as fees. The exemplary elements of internal data governance structures within each organization emphasize that the means of assigning decision-making rights and accountability for the governance of data lie within the purview of the intra-organizational perspective. Conventional instruments for this purpose include data strategy, data policy, or organizational frameworks defining roles, tasks, and responsibilities for data. This type of organizational setup is embedded in the hierarchy of the organization and mainly focuses on internal data. With enabling a gateway through technical means to the data ecosystem, e.g., through connectors or other infrastructures, the organization executes a role or function in the data ecosystem, and this development must be reflected internally to consider the new influencing factors in processes, roles, and policies. Therefore, organizations initially must lay the foundations internally to be able to successfully engage in inter-organizational data sharing. This involves considering the utilization of internal and external data by:

- Understanding which data is available and relevant within and outside the organization
- Clarifying the ownership status and usage rights of data assets
- Assigning responsibilities that transcend organizational scope through cross-organizational alignment and coordination of data activities
- Creating transparency and lineage of how the data is used (both internally and externally)
- Defining under which conditions data can be shared with whom

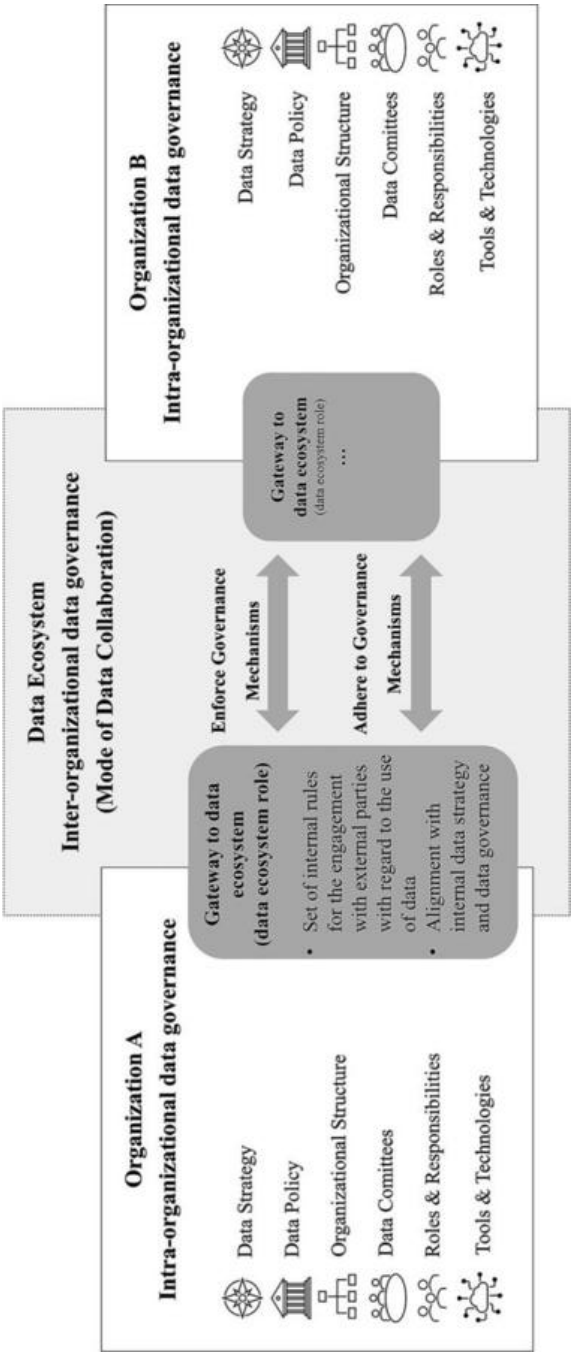

Fig. 2.3 Conceptual model of the transcending intra-organizational data governance perspective in data sharing

2.6.2 *Recommendations for Actions for Data Ecosystem Design*

While the emergence of data ecosystems offers new business opportunities for the various participants in the ecosystem, many social, environmental, and business challenges must be overcome to pave the way for the realization of these innovative potentials. Some of the biggest challenges are:

- **Interoperability:** Data ecosystems need to create a trustworthy environment that provides user-friendly data protection mechanisms and solutions that ensure that citizens and businesses can share data while ensuring privacy and sovereignty [39]. The challenge is to create an appropriate overall technical architecture that considers the main reference platforms and technologies supporting data sharing, enhances existing solutions and architectures, defines the overall reference architecture, and develops platform-independent building blocks for trusted data sharing and interoperability.
- **Trust:** New technologies and approaches are needed to increase trust in data sharing so that more data holders make their data available for new applications [66]. A framework is needed that includes building blocks for data management, data sharing, data protection techniques, and processing of data while maintaining data sovereignty and traceability. This framework should not only include technologies but also incentive and business model tools for developers and strategists of companies that want to use data for new collaborations and business opportunities.
- **Data sovereignty:** A data ecosystem should support compatibility with the latest and emerging legislation, such as the EU General Data Protection Regulation (GDPR), and the free flow of nonpersonal data, as well as ethical principles. This will increase trust in industrial and personal data platforms, enabling larger data markets that connect currently isolated data silos and increase the number of data providers and users in the markets. The outcome should be platform-independent so that it can be applied in different domains with platforms based on different technologies.
- **Compliance:** When building data ecosystems, attention must be paid to compliance with antitrust regulations. To avoid the risk of data monopolies, efforts should be made to improve the cross-border mobility of nonpersonal data in the internal market, which is currently restricted in many Member States by localization restrictions or legal uncertainty in the market. Furthermore, it should be ensured that the powers of competent authorities to request and obtain access to data for control purposes, e.g., for inspections and audits, remain unaffected. Finally, switching of service providers and data transfers should be facilitated for business users of data storage or other processing services without creating excessive burdens on service providers or market distortions.
- **Data economics:** Data is at the center of data ecosystems as a strategic resource. Against this background, data ecosystems should motivate data providers and

owners to open their data for various applications [67]. Personal data is becoming a new economic asset class, a valuable resource for the twenty-first century that will touch all aspects of society. The rapid development of the personal data services (PDS) market will greatly change the way individuals, companies, and organizations interact with each other, as individuals gain more control over their data or service providers process personal data.

References

1. DalleMulle, L., Davenport, T.H.: What's your data strategy? Harv. Bus. Rev. **95**, 112–121 (2017)
2. Dey, S.: Defining a data strategy. https://dxc.com/us/en/insights/perspectives/paper/defining-a-data-strategy (2021). Accessed March 2023
3. SAS. The 5 Essential Components of a Data Strategy (2016)
4. Aiken, P., Gillenson, M., Zhang, X., Rafner, D.: Data management and data administration: assessing 25 years of practice. In: Innovations in Database Design, Web Applications, and Information Systems Management, vol. 22, pp. 289–309. IGI Global (2013)
5. Wang, R.Y.: A product perspective on total data quality management. Commun. ACM. **41**, 58–65 (1998)
6. Ballou, D., Wang, R., Pazer, H., Tayi, G.K.: Modeling information manufacturing systems to determine information product quality. Manag. Sci. **44**, 462–484 (1998)
7. Goodhue, D.L., Kirsch, L.J., Quillard, J.A., Wybo, M.D.: Strategic data planning: lessons from the field. MIS Q. **16**, 11–34 (1992)
8. Buhl, H.U., Röglinger, M., Moser, F., Heidemann, J.: Big data. Bus. Inf. Syst. Eng. **5**, 65–69 (2013)
9. Provost, F., Fawcett, T.: Data science and its relationship to big data and data-driven decision making. Big Data. **1**, 51–59 (2013)
10. Wixom, B.H., Ross, J.W.: How to monetize your data. MIT Sloan Manag. Rev. **58**, 10–13 (2017)
11. Chen, H., Chiang, R.H.L., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**, 1165–1188 (2012)
12. Clarke, R.: Big data, big risks. Inf. Syst. J. **26**, 77–90 (2016)
13. Abbasi, A., Sarker, S., Chiang, R.: Big data research in information systems: toward an inclusive research agenda. JAIS. **17**, I–XXXII (2016)
14. Chen, K., Li, X., Wang, H.: On the model design of integrated intelligent big data analytics systems. Ind. Manag. Data Syst. **115**, 1666–1682 (2015)
15. O'Leary, D.E.: Embedding AI and crowdsourcing in the big data lake. IEEE Intell. Syst. **29**, 70–73 (2014)
16. Legner, C., Pentek, T., Otto, B.: Accumulating design knowledge with reference models: insights from 12 years' research into data management. JAIS. **21**, 735–770 (2020)
17. Loth, A.: Die Notwendigkeit einer modernen Datenstrategie im Zuge der digitalen Transformation. Inf. Wiss. Prax. **68**, 75–77 (2017)
18. Barton, D., Court, D.: Three Keys to Building a Data Driven Strategy. McKinsey & Company Quarterly (2013)
19. Mintzberg, H.: The strategy concept I: five Ps for strategy. Calif. Manag. Rev. **30**, 11–24 (1987)
20. Gür, I., Spiekermann, M.: Data Strategy Praxis Report: Tools and Approaches in the Current Data Economy. Fraunhofer ISST (2020)
21. Henderson, D. (ed.): DAMA-DMBOK: Data Management Body of Knowledge, 2nd edn. Technics Publications, Basking Ridge (2017)

22. Ladley, J.: Definitions and concepts. In: Data Governance, pp. 7–20. Elsevier (2012)
23. Otto, B., ten Hompel, M., Wrobel, S. (eds.): Designing Data Spaces: The Ecosystem Approach to Competitive Advantage. Springer, Cham (2022)
24. Otto, B., Österle, H.: Corporate Data Quality. Springer, Berlin (2016)
25. European Commission. The European data market monitoring tool: key facts & figures, first policy conclusions, data landscape and quantified stories: d2.9 final study report (2020)
26. Otto, B.: Quality and value of the data resource in large enterprises. Inf. Syst. Manag. **32**, 234–251 (2015)
27. Azkan, C., Strobel, G., Iggena, L., Gelhaar, J., Kreyenborg, A.: Barriers to the development of data-driven services: an ISM approach for SMEs. In: Proceedings of the 56th Hawaii International Conference on System Sciences. University of Hawaii at Manoa (2023)
28. Gelhaar, J., Gürpinar, T., Henke, M., Otto, B.: Towards a taxonomy of incentive mechanisms for data sharing in data ecosystems. In: Proceedings of the Twenty-Fifth Pacific Asia Conference on Information Systems. AISeL, Dubai, UAE (2021)
29. Otto, B., Lis, D., Jürjens, J., Cirullies, J., Opriel, S., Howar, F., et al.: Data Ecosystems: Conceptual Foundations, Constituents and Recommendations for Action. Fraunhofer ISST (2019)
30. Gelhaar, J., Becker, F., Groß, T.: Characterization of relationships in data ecosystems. In: Proceedings of the Conference on Production Systems and Logistics: CPSL 2022, vol. 2022. CPSL
31. Moore, J.F.: Predators and prey: a new ecology of competition. Harv. Bus. Rev. **71**, 75–86 (1993)
32. Oliveira, M.I., Barros Lima, G.F., Lóscio, B.F.: Investigations into data ecosystems: a systematic mapping study. Knowl. Inf. Syst. **61**, 589 (2019)
33. Gelhaar, J., Groß, T., Otto, B.: A taxonomy for data ecosystems. In: Proceedings of the 54th Hawaii International Conference on System Sciences 2021. University of Hawaii at Manoa (2021)
34. Cappiello, C., Gal, A., Jarke, M., Rehof, J.: Data ecosystems: sovereign data exchange among organizations: report from Dagstuhl seminar 19391. Dagstuhl Reports. **9**, 66–134 (2019)
35. Bean, R.: Why is it so hard to become a data driven company? Harv. Bus. Rev. (2021)
36. Abraham, R., Schneider, J., vom Brocke, J.: Data governance: a conceptual framework, structured review, and research agenda. Int. J. Inf. Manag. **49**, 424–438 (2019)
37. Lis, D., Arbter, M.: Data Governance als Hebel für datengetriebene Wertschöpfung: Der Weg zu einer datengetriebenen Organisation. ERP. Management. (2022)
38. European Commission. Shaping Europe's digital future: a European approach to artificial intelligence. 02.02.2023. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence
39. Otto, B., Jarke, M.: Designing a multi-sided data platform: findings from the international data spaces case. Electron. Mark. **29**, 561–580 (2019)
40. Al-Ruithe, M., Benkhelifa, E., Hameed, K.: Data governance taxonomy: cloud versus non-cloud. Sustainability. **10**, 1–26 (2018)
41. de Haes, S., van Grembergen, W.: IT governance and its mechanisms. Inf. Syst. Control J. **2004**, 27–33 (2004)
42. Otto, B.: Organizing data governance: findings from the telecommunications industry and consequences for large service providers. Commun. Assoc. Inf. Syst. **29**, 45–66 (2011)
43. Alhassan, I., Sammon, D., Daly, M.: Data governance activities: an analysis of the literature. J. Decis. Syst. **25**, 64–75 (2016)
44. Weber, K., Otto, B., Österle, H.: One size does not fit all: a contingency approach to data governance. J. Data Inf. Qual. **1**, 1–27 (2009)
45. de Prieelle, F., de Reuver, M., Rezaei, J.: The role of ecosystem data governance in adoption of data platforms by internet-of-things data providers: case of Dutch horticulture industry. IEEE Trans. Eng. Manag. **69**, 940–950 (2020)

46. Lee, S.U., Zhu, L., Jeffery, R.: Data governance for platform ecosystems: critical factors and the state of practice. In: Twenty First Pacific Asia Conference on Information Systems. PACIS, Langkawi, Malaysia (2017)
47. Hein, A., Schreieck, M., Wiesche, M., Krcmar, H.: Multiple-case analysis on governance mechanisms of multi-sided platforms digitale. In: Multikonferenz Wirtschaftsinformatik. Technische Universität Ilmenau, Ilmenau, Germany (2016)
48. Lis, D., Otto, B.: Towards a taxonomy of ecosystem data governance. In: Hawaii International Conference on System Sciences, pp. 6067–6076. HICSS (2021)
49. Lis, D., Otto, B.: Data governance in data ecosystems – insights from organizations. In: Americas Conference on Information Systems (AMCIS). AISeL (2020)
50. Winkler, T.J., Wessel, M.: A primer on decision rights in information systems: review and recommendations. In: ICIS 2018, San Francisco, CA (2018)
51. Provan, K.G., Kenis, P.: Modes of network governance: structure, management, and effectiveness. J. Public Adm. Res. Theory. **18**, 229–252 (2007)
52. van den Broek, T., van Veenstra, A.F.: Modes of governance in inter-organizational data collaborations. In: ECIS 2015. AIS Electronic Library, Münster, Germany (2015)
53. Selander, L., Henfridsson, O., Svahn, F.: Capability search and redeem across digital ecosystems. J. Inf. Technol. **28**, 183–197 (2013)
54. Demil, B., Lecocq, X.: Neither market nor hierarchy nor network: the emergence of bazaar governance. Organ. Stud. **27**, 1447–1466 (2006)
55. Powell, W.M.: Neither market nor hierarchy: network forms of organization. In: Cummings, L. L., Staw, B.M. (eds.) Research in Organizational Behavior, pp. 295–336. JAI Press, Greenwich, CT (1990)
56. Williamson, O.E.: The institutions of governance. Am. Econ. Rev. **88**, 75–79 (1998)
57. Lowndes, V., Skelcher, C.: The dynamics of multi-organizational partnerships: an analysis of changing modes of governance. Public Adm. **76**, 313–333 (1998)
58. Halckenhaeusser, A., Foerderer, J., Heinzl, A.: Platform governance mechanisms: an integrated literature review and research directions. In: Proceedings of the 28th European Conference on Information Systems (ECIS), pp. 15–17. ECIS (2020)
59. Dekker, H.C.: Control of inter-organizational relationships: evidence on appropriation concerns and coordination requirements. Acc. Organ. Soc. **29**, 27–49 (2004)
60. Enders, T., Wolff, C., Satzger, G.: Knowing what to share: selective revealing in open data. In: Proceedings of the 28th European Conference on Information Systems (ECIS). ECIS (2020)
61. Oliveira, M.I., Lóscio, B.F.: What is a data ecosystem? In: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, pp. 1–9. ACM, Delft, Netherlands (2018). https://doi.org/10.1145/3209281.3209335
62. Otto, B., Korte, T., Azkan, C., Spiekermann, M., Lis, D., Gelhaar, J., et al.: Data Economy: Status quo der deutschen Wirtschaft & Handlungsfelder in der Data Economy. Institut der deutschen Wirtschaft (2019)
63. Manner, J., Nienaber, D., Schermann, M., Krcmar, H.: Governance for mobile service platforms: a literature review and research agenda. In: 2012 International Conference on Mobile Business. AIS (2012)
64. Jagals, M.: Expanding data governance across company boundaries – an inter-organizational perspective of roles and responsibilities. In: Serral, E., Stirna, J., Ralyté, J., Grabis, J. (eds.) The Practice of Enterprise Modeling, pp. 245–254. Springer, Cham (2021)
65. D'Hauwers, R., Walravens, N.: Do you trust me? Value and governance in data sharing business models. In: Yang, X.-S., Sherratt, S., Dey, N., Joshi, A. (eds.) Proceedings of Sixth International Congress on Information and Communication Technology, pp. 217–225. Springer Singapore, Singapore (2022)

66. Gelhaar, J., Otto, B.: Challenges in the emergence of data ecosystems. In: Pacific Asia Conference on Information Systems (PACIS) 2020, p. 175. AIS (2020)
67. Gelhaar, J., Müller, P., Bergmann, N., Dogan, R.: Motives and incentives for data sharing in industrial data ecosystems: an explorative single case study. In: Proceedings of the 56th Hawaii International Conference on System Sciences, pp. 3705–3714. University of Hawai'i at Mānoa (2023)

Chapter 3
Human Resources Management and Data Governance Roles: Executive Sponsor, Data Governors, and Data Stewards

David Plotkin

3.1 Introduction

Data Governance involves a lot of people and a lot of roles. These include roles directly engaged in Data Governance (Executive sponsor, data governors, and various types of data stewards) as well as expertise and support from the Data Governance Program Office, which includes roles such as the Data Governance Manager and the Enterprise Data Steward. To staff such an organization, it will likely be necessary to hire some expertise, as well as recruit and train within the organization. Thus, it is important to know the duties and responsibilities of each role – not only to recruit from outside the organization but also to pick the right people from inside the organization – and to ensure that any bonus program (sometimes called "Management by objectives," or MBO) is measuring and rewarding the appropriate goals. This chapter describes the role of Human Resources in coordinating the filling of these roles as well as describing the responsibilities of each role.

3.2 The Role of Human Resources in Data Governance

The implementation of Data Governance requires people with appropriate skills who can take on roles and responsibilities that are not common in most organizations. As we shall see later in this chapter, these roles and responsibilities focus on the rigorous management of data, working well in groups to reach consensus on ways to achieve the goals of organization in the data space, and being willing and able to take the "enterprise view" for managing data, that is, the willingness to think about and

D. Plotkin (✉)
Metadata Services at MUFG Union Bank, Walnut Creek, CA, USA

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_3

implement strategies and tactics that are best for the organization, and not just for the business function to which the individuals belong. It also requires having people who can make decisions about data and metadata both in the interests of the business function they represent and in the best interests of the organization as a whole.

Some of the work and the decisions are undertaken as a group in specific committees, such as the Data Stewardship Council, or the Data Governance Board. These committees and the work they do will be explained later in this chapter. On the other hand, some of the work is undertaken by individuals in various roles, independent of the committees. For example, while the standards for defining business terms properly are decided on by the Data Stewardship Council, individual Business Data Stewards (who form the Data Stewardship Council) educate their business functions and make decisions about the data their business function owns. Similarly, Data Governors (who make up the Data Governance Board) have both committee and individual responsibilities.

As we shall see, the list of responsibilities for both the committees and the individuals is relatively long and involved – and the job definition needs to be carefully crafted by Human Resources. Furthermore, Human Resources must work with the business functions to appropriately set the goals/objectives/MBO (management by objective) so that evaluations and compensation (including bonuses) take into account the special requirements of the role(s) that individuals fill in the Data Governance effort. It should come as no surprise that paying participants for achieving these goals is an effective way to incentivize them to do the job well!

Beyond properly defining the roles and responsibilities for participants in the Data Governance effort, new positions must be defined for the organization (*Data Governance Program Office*, or DGPO) that will guide the entire effort forward. The members of the DGPO are highly specialized professionals, and it is imperative that the people hired into these roles can perform the necessary duties. Defining these positions is critical to a successful Data Governance effort, especially since the very first step in establishing Data Governance is usually to hire the leader of the DGPO (often called the Data Governance Manager).

3.3 Understanding the Structure of the Data Governance Organization

Although there can be slight variations, the Data Governance organization is usually made up of three levels, as shown in Fig. 3.1.

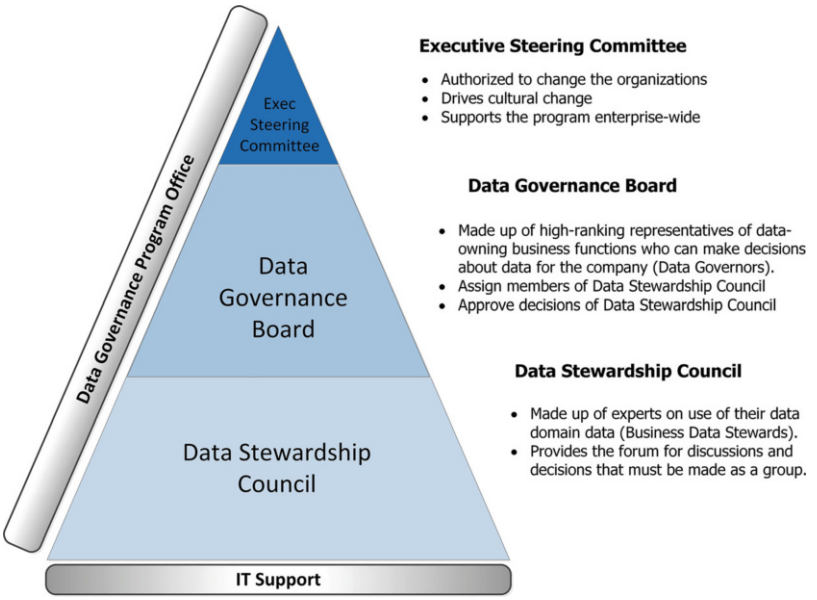

Fig. 3.1 The multiple levels that are usually present in a Data Governance Organization (© by David Plotkin)

3.3.1 Executive Steering Committee

Not surprisingly, the Executive Steering Committee is made up of a selection of executives. These highly placed officers provide the support needed for a data governance initiative. They are necessary because any effort that requires significant change in the way the organization operates requires support that is both extremely visible and has the authority to implement the changes. Without this level of support, other members of the organization may not take it seriously or be willing to put in the extra effort to make it successful.

The Executive Steering Committee has several specific responsibilities. First of all, it drives the cultural changes needed to manage data across business functions and drives a decision-making process that takes the overall enterprise into account. Cultural changes to the functioning of the organization may also be necessary, and the executives must drive that change. For example, it may be that all decisions are normally made by consensus. In Data Governance, however, it is often true that the responsible and accountable groups must decide, and consensus is not reached. Executives are the ones who must communicate that this is expected.

Organizational changes (including hiring) are often necessary for the Data Governance effort to succeed. Further, new tools (and support personnel) are also often needed. Executives can drive the funding and required changes to the organizational structure.

Executives usually have the widest understanding of the business and its objectives. They can balance business priorities with operational needs across the enterprise. Further, they can ensure that decisions regarding the data support the strategic direction of the organization and ensure that the appropriate policies and practices are adopted to drive Data Governance.

Executives work closely with the Data Governors (members of the Data Governance Board), appointing the individuals that represent their business function(s), resolving issues escalated by the Data Governance Board, making sure that business functions and IT are participating in the Data Governance effort, and providing advice, direction, and feedback to the board.

3.3.2 Data Governance Board

The Data Governance Board resides at the middle level of the Data Governance pyramid. The board members (referred to as Data Governors or Data Owners) are ultimately accountable for business data use, data quality, and resolution of issues. They make the decisions about the data owned and used by their business function. They ensure that the most relevant needs of the organization are being addressed, establish priorities of issues to be worked on, and provide the funding and personnel to make a change or remediate an issue.

Other duties of the members of the Data Governance Board include the following:

- Ensure that annual performance measures are set up and used that align with Data Governance and business objectives. Most participants in the Data Governance effort are normally chosen from the ranks of existing employees, and their existing performance measure does not include Data Governance goals. These measures must be added as part of the Data Governance effort.
- Review and approve Data Governance policies and goals.
- Assign Business Data Stewards from their business function. Since Data Governors represent their business function, they need to pick and assign the best individuals to represent their function as Business Data Stewards. The need this authority since the Business Data Stewards may not work for the Data Governor, and the supervisor for the chosen people may not be happy about allowing their people to have extra duties!
- Represent all data stakeholders in the Data Governance process, including owners of business processes that produce the data, and report owners that track metrics based on the data and everyone who uses the data.
- Identify and provide data requirements that meet both the business objectives of their business function and those of the enterprise.
- Define data strategies that support the business strategy and requirements of the enterprise.

- Communicate concerns and issues about the data to the Data Stewardship Council and the Data Governance Program Office.

3.3.3 Data Stewardship Council

The *Data Stewardship Council* is the group of (mostly) Business Data Stewards (discussed later in this chapter) who work together to get the day-to-day work of Data Governance done. In a sense, the Data Stewardship Council is the operational aspect of Data Governance. It is where policies and strategies get turned into procedures, processes, and tactics, and these processes and procedures are worked on daily to get the desired results.

The responsibilities of the Data Stewardship Council include the following:

- Focus on ways to improve how an organization obtains, manages, leverages, and gets value out of its data. The Data Stewardship Council should be looking for ways to improve the data quality in support of projects as well as key business processes.
- Be the advisory body for enterprise-level data standards and processes. The standards and processes establish HOW data-related work gets done, as well as the goals of the work. Recommending what standards and processes are needed to meet business objectives is an important task because the Business Data Stewards are on the front lines and thus in a great position to see how well the standards and processes are working to make Data Governance a success.
- Resolve issues. The Data Stewardship Council must work together as a team to settle any data issues that arise. These may include disagreements over meaning or rules, differing requirements for data quality, modifications to how data is used, and which business functions should own key data elements.
- Communicate decisions made by the Data Stewardship Council and Data Governance Board. Decisions about the data need to be communicated to the data analyst community and others who use the data. Data Governance should not be run in a silo – its power comes from sharing the decisions with the people who need to know.
- Align Data Governance to the business. The rules, processes, and procedures used to govern the data must align to the business. Data Governance must not be perceived as a roadblock or out of synch with the priorities of the business. If Data Governance does not prove its value, the resources will be put elsewhere.
- Provide feedback and participate in data governance processes. The Data Stewardship Council (as a group) needs to define and design the processes since they are mostly the people (or represent the people) who will be expected to follow them. They will also be expected to provide feedback on the processes to determine which ones are working and which ones need to be changed or discarded.

- Communicate data governance processes, procedures, and objectives across the organization. The Data Stewardship Council must communicate the processes and objectives as well as the reasons for following the processes and achieving the objectives. The business functions represented by the Data Stewards expect to receive these communications on a regular basis, and members of the business functions are expected to follow the rules and procedures.
- Review and evaluate Data Governance performance and effectiveness. As performance objectives are defined for the Data Stewards, they need to accept those objectives, agree with how the results are measured, and work toward achieving them. This is easier if the Data Stewards want to participate and agree with performance objectives.
- Provide input into Data Governance goals and scorecard development. The Data Governance goals must align with performance objective measurement, so the Data Stewards need input into the Data Governance goals and the scorecards that present the progress.
- Collaborate on Processes and Procedures. Policies drive what must be accomplished; procedures say HOW the accomplishments will be met. Since the Data Stewards must execute on the processes and procedures, they must have input into them. In addition, the Data Stewards (who are knowledgeable about the data and care about the data) are the very people who are best able to suggest what processes and procedures are needed as well as what is reasonable for them to achieve.
- Collaborate with other interested parties in the management of definitions and data issues. The Data Stewardship Council provides a forum for the Data Stewards to discuss and reach agreement (or at least consensus) on definitions of business data elements and data quality issues.
  - Definitions: Many people (commonly known as stakeholders) have an interest in how terms are defined, and it is especially important that the stakeholders have a common understanding of the data names and definitions. Managing definitions requires soliciting input from stakeholders during both the initial definition phase and for any changes to the definitions that are proposed.
  - Data Quality Issues: The Data Stewards both individually and as part of the Council manage issues with the data and data quality. The impacts of the issues and any proposed remediations must be assessed and the impact determined.
- Enforce use of agreed-upon Business terminology. Different terminology should not be used to represent the same concept. Business terms are named and given a robust definition, and business rules are defined. These terms must be used consistently across the organization and synonyms should be actively discouraged by the Business Data Stewards because that leads to confusion. This is especially true when the incorrectly used term name has been defined to mean something different.

3.3.4 Data Governance Program Office (DGPO)

The Data Governance effort is run by the Data Governance Program Office (DGPO). The purview of the DGPO includes documentation, coordination of the program, communication, and enforcement of policies, procedures, and decisions, including escalation of issues to the Data Governance Board or the Executive Steering Committee. Ample resources are required, including appropriately skilled staffing. Failing to create and staff a DGPO may well doom the Data Governance effort to ineffectiveness or even failure.

Members of the DGPO have many responsibilities. The responsibilities can be broken down into three areas: the responsibilities of the overall Data Governance Program Office, the responsibilities of the Data Governance Manager, and the responsibilities of the Enterprise Data Steward.

3.3.4.1 Data Governance Program Office (DGPO) Responsibilities

The DGPO responsibilities include the following:

- Schedule meetings, set agendas, and document the activities of the Executive Steering Committee, Data Governance Board, and Data Stewardship Council.
- Provide best practices in Data Governance as goals for the organization to strive for.
- Provide and make available educational materials as well as practical training for the various roles needed.
- Enforce (and escalate where necessary) policies and procedures related to Data Governance.
- Manage and publish working documents (such as the issue log) in a document repository.
- Maintain and publish Data Governance-related processes, procedures, and standards.
- Create and measure Data Governance metrics.
- Responsible for disseminating material related to Data Governance, including the strategy and vision statement.

3.3.4.2 Data Governance Manager Responsibilities

The Data Governance Manager oversees the DGPO. This individual must have a strong working knowledge of how to implement Data Governance and Data Stewardship.

The first responsibility of the Data Governance Manager is to create the DGPO, specify the job requirements for the staff (most importantly the Enterprise Data Steward), and hire the staff. This hiring often requires adding headcount, creating

new job classifications, and other tasks that require cooperation from Human Resources.

The Data Governance Manager must also start out by creating a task list for the early stages of the Data Governance, an initial timeline for implementation, and the introductory material needed to work with the executives to begin recruiting the Data Governance Board members. A training plan and material to train the Data Governance Board is an important deliverable as well, since the Data Governors need to understand their responsibilities, including picking the right people to be Business Data Stewards.

Once the DGPO is up and running (and the staff hired and trained), the Data Governance Manager has day-to-day responsibilities for running the DGPO. These include the following:

- Manage the DGPO, including making sure there are adequate staffing levels.
- Track which business functions should be participating in the Data Governance effort and make sure they are represented in both the Data Governance Council and the Data Stewardship Council.
- Recruit involvement from support organizations, including Enterprise Architecture, Program Management, IT applications, and Human Resources.
- Implement Data Governance and Data Stewardship capabilities in alignment with the needs of the business.
- Ensure that the Executive Steering Committee, Data Governance Board, and Data Stewardship Council have representation from all business functions that own data.
- Help build the Data Governance strategy, necessary policies, and a consensus for acceptance by the Data Governors.
- Various support organizations need to participate in supporting Data Governance, so the Data Governance Manager needs to obtain that support or escalate if there is a lack of support as necessary.
- Identify the business needs for Data Governance capabilities by collaborating with the organization's leadership.
- Ensure that annual performance measures align with Data Governance and business objectives by working with the Executive Steering Committee and Data Governance Board.
- Integrate the Data Governance processes into the enterprise processes, including project management and software development.
- Report Data Governance performance to the Executive Steering Committee.
- Work with IT to develop or license appropriate tools for documenting procedures, capturing business metadata, building the communication plan and issue log, and documenting other deliverables.
- Meet with the Business Data Stewards and stakeholders to understand their needs and the feasibility of proposed issue resolutions.
- Ultimately be responsible for providing the vision and important Data Governance messages to the enterprise.

- Manage the Enterprise Data Steward, who coordinates and manages the activities of the Data Stewards.

3.3.4.3 Enterprise Data Steward Responsibilities

The Enterprise Data Steward reports to the Data Governance Manager and is the key member of the DGPO charged with managing the day-to-day efforts of the Data Stewards and the Data Stewardship Council. Although the Data Governance Manager can fill this role temporarily, in the long term, that is not a good idea. This is because starting up and running a Data Governance effort is a BIG job. In addition, the skills of the Enterprise Data Steward lean much more heavily toward managing a group of independent (and knowledgeable) individuals to solve the ongoing issues that arise and work effectively as a group.

The responsibilities of the Enterprise Data Steward can be broken down into three major categories: Leadership, Program Management, and Measurement. The Leadership responsibilities include the following:

- Provide leadership for the community of Data Stewards and run the Data Stewardship Council. The Business Data Stewards don't report *functionally* to the Enterprise Data Steward, but do have a "dotted line" relationship, and the Enterprise Data Steward will be responsible for providing the evaluation on how effectively the Business Data Stewards fulfill that role.
- Alongside the Data Governance Manager, help to develop the Data Governance framework, objectives, road map, and timeline.
- Propose and initiate projects that drive forward the vision and objectives of Data Governance. Project may include building workflows for critical data processes, incorporating data governance deliverables into project plans, and choosing and implementing supporting tool sets.
- Focus the efforts of the Business Data Stewards and DGPO on projects and efforts that are of highest importance to the organization.
- Define the standardized criteria for prioritizing projects and efforts that need Data Governance resources. The Business Data Stewards are then responsible for using these criteria to establish the actual priorities.
- Be the single point of contact for Data Stewardship for anyone who needs to get or provide information.
- Lead the Data Stewardship organization. The initial setup of the Data Stewardship Council, as well as leading the council, is the purview of the Enterprise Data Steward.

The Program Management responsibilities include the following:

- Design the procedures for Data Stewardship. This includes collecting specifications on how data should be managed and formulating the processes and procedures that will be followed by the Business Data Stewards. Publishing

documentation on the procedures – and updating that documentation when it changes – is also part of the responsibilities.

- Create, manage, and lead the agenda for Data Stewardship Council meetings. The agenda would include issues, status updates, and anything else worthy of discussion. Any meeting notes would also be published by the Enterprise Data Steward.
- Create and maintain a repository for documents and other Data Governance deliverables. Documentation about processes and procedures, presentations, and training should all be stored in the repository.
- Help the Business Data Stewards participate in enterprise efforts, including data quality improvement, creation of data and metadata life cycles, various aspects of Master Data Management, risk assessment, data warehouse/data lake engagements, and reference data management.
- Review and manage issues, and work with the Business Data Stewards to prioritize issues and find solutions.
- Provide counseling to projects to ensure that the project is developed in line with Data Governance principles. Guidance from Data Governance ensures that business terms are defined and used properly, and Data Governance is involved in the appropriate tasks undertaken by the project. Project managers need to be trained on what is needed, the importance of finding the right data with the necessary quality, and milestones that must be added to the project plan. The Enterprise Data Steward must also provide guidance on the types of resources necessary and where those resources can be found.

The Measurement responsibilities include the following:

- Work with the Business Data Stewards to define, build, and measure Data Governance metrics. The Enterprise Data Steward ensures that the measures are done and the metrics are created and published.
- Create and publish Data Governance scorecards. These can be generated on a periodic basis and provide information on the progress that Data Governance is making in achieving its goals.

3.4 Key Roles and Responsibilities for Data Stewards

There are three main types of Data Stewards that take on the responsibilities needed to achieve a successful and robust Data Stewardship effort. While they may be called slightly different names, in this book they are called Business Data Stewards, Technical Data Stewards, and Operational Data Stewards. Some organizations may also use another type of Data Steward – the "Project Data Steward" – to help fill in and support the Business Data Stewards on projects.

Although we will go into far more detail on each type of Data Steward, in brief the Data Stewards are classified as follows:

Table 3.1 Summary of the responsibilities of Data Stewards

Type of steward	Responsibilities
Business Data Steward	Primarily responsible for their business function's data Supports Project and Operational Data Stewards Member of the Data Stewardship Council Works with business stakeholders on resolving issues Manages their metadata Promotes Data Stewardship to their business function
Technical Data Steward	Provides expertise on the "information chain": applications, databases, and ETL (extract, transform, and load) Assigned by IT leadership to this role
Operational Data Steward	Provides support to Business Data Stewards Makes recommendations to improve the quality of the data Helps to enforce business rules for their data and data they use Often located on the "front line" for data entry
Project Data Steward	Represents the Data Stewardship effort on projects Provides deliverables to the project that are the responsibility of Data Stewards Serves as a point of contact between Business Data Stewards and the project Ensures that project issues which require the attention of the Business Data Steward are brought to their attention and solutions/resolutions brought back to the project team

- Business Data Stewards represent their business function and are responsible for understanding the data owned by that business function.
- Technical Data Stewards typically come from IT and have knowledge on how applications, data stores, transformations, and other technologies work. They often know the reasons why data is the way it is.
- Operational Data Stewards provide help to the Business Data Stewards and are usually people who work directly with data and can provide more immediate feedback when they note issues with the data.
- Project Data Stewards represent Data Governance on projects, reporting to the appropriate Business Data Stewards when questions or issues about the data arise on the project.

Each of these types of Data Stewards will be discussed in more detail, but Table 3.1 provides a summary of the Data Steward's responsibilities.

3.4.1 Business Data Stewards

A *Business Data Steward* is the primary representative for the data owned by their specific business function. The responsibility extends to the quality, usage, meaning, and rules about the data. They are people who know the data well and work with it frequently. Since no one can know everything about a wide range of data, the

Business Data Steward must know the subject matter experts about the data that they can consult with. Even after consulting with the Subject Matter Expert, the Business Data Steward (and not the subject matter expert) takes responsibility for the data and metadata. These individuals have the authority to require that subject matter experts participate in providing the necessary information.

The responsibilities of the Business Data Steward can be broken down into three categories: Business Alignment, Data Life cycle Management, and Data quality and reduced risk. The Business Alignment responsibilities include the following:

- Work closely with other Business Data Stewards, mostly through the Data Stewardship Council. Small groups may also collaborate through "working groups" of targeted Business Data Stewards who have an interest in a particular data set or topic.
- Align with a business function. Business Data Stewards represent the needs of their business function in the Data Governance effort. They are responsible for speaking up if they are facing data issues, as well as when proposed changes or problem solutions will not work for them. They also help to drive (along with the Data Governor for that business function) the Data Governance effort in their part of the business. Finally, they are the single point of contact for members of their business to engage with Data Governance.
- Identify and own key business terms that are important to their business function. Business Data Stewards need to prioritize the business terms that are important to their business and provide the important metadata about those terms. The metadata must include the definition, a unique name that meets the naming standards, business rules (including those that define quality), and key systems where the business terms have a physical counterpart.
- Participate in efforts to define Data Stewardship metrics, processes, and standards. They are in a good position to define the metrics that they must meet and to ensure that processes and standards are practical and can be executed on.
- Support the Data Governors by reviewing items such as issues or concerns about the data and, where appropriate, making recommendations.
- Help the members of their business to have a practical understanding of the data. Data Analysts within the business must understand what the data means and the business rules it must follow. This will allow them to use the data properly and spot potential issues early so that the issues can be brought to the attention of the Business Data Steward. The data analysts may find out critical information that should be brought to the attention of Business Data Steward.
- Communicate data decisions and the impact of those decisions to their business function.
- Provide business requirements about data usage and quality on behalf of their business function. They must also evaluate stated business requirements for Data Governance and projects that might conflict with the needs of their business.

Data Life cycle Management responsibilities include the following:

- Provide input and guidance to the Data Governors to engage in the change control process. This process is used when approving recommendations made by the Business Data Stewards.
- Collect business requirements and priorities from their business function to identify where the requirements can be combined into a single workstream, potentially along with business requirements from other business functions.
- Work with stakeholders (including other Business Data Stewards) to resolve conflicts or manage issues through the resolution and escalation process. The conflicts may include definitions, appropriate data usage, and required data quality.
- Assess the impacts of proposed changes to their business function, stakeholders, and the enterprise. The Business Data Steward should know the needs of their business function, who their stakeholders are, and areas of the enterprise that would be affected. A diagram of how the data flows through enterprise – sometimes known as an "information chain" – can help to assess the impacts. See Fig. 3.2 for an illustration of an information chain and some sample impacts that can occur when changes are made.
- Participate in "working groups" that consist of a subset of Business Data Stewards who need to cooperate to achieve a common result focused on limited data set – and thus which does not require all the Business Data Stewards.
- Ensure that data in their business function is used in a consistent way and only for approved usages. Proposals for new ways of using the data need to be reviewed with the Business Data Steward because the data may not support the new usage.
- Define and publish the business rules relating to their data. These rules can include capture, usage, derivation, and data quality business rules. Having a set of well-defined and understood rules that everyone is aware of ensures the data is not used in ways it was never intended for.

Data Quality and reduced risk responsibilities include the following:

- Work with the Technical Data Stewards to define the data quality rules based on the requirements of all the stakeholders. These rules serve as the basis for programming the data quality tool.
- Define the acceptable levels of data quality based on business needs and the data usage. The results of examining the data (a process called "profiling") against the data quality rules establish the quality of the data, which can then be monitored against the required quality.
- When the quality of the data falls below acceptable levels, the Business Data Stewards need to participate in the effort to evaluate the issue, find the root cause for the deterioration, and help to determine whether there is sufficient business benefit to correct the cause of the declining data quality. Once again, Technical Data Stewards play a role in providing the data to be examined as well as interpreting the results of the data profiling.
- Manage the business function's reference data. Many systems that the business depends on use a set of valid values and descriptions/meanings for those values. These code/description lists must be managed to ensure that the codes and their

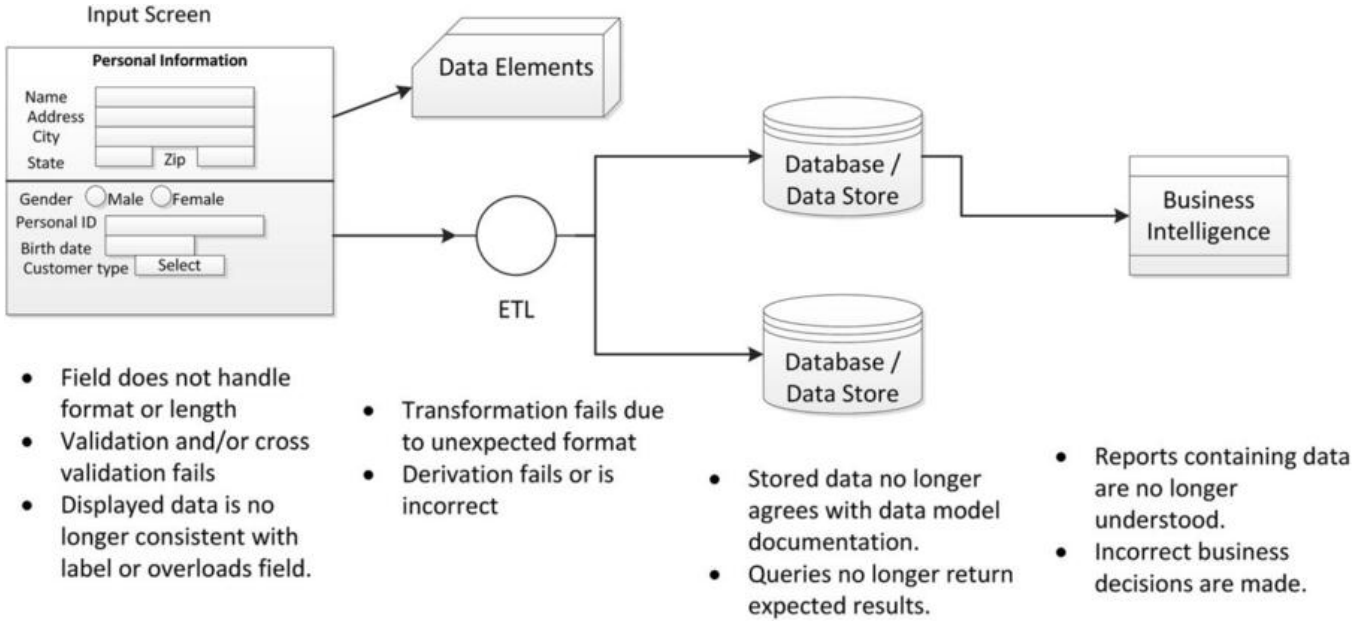

Fig. 3.2 Decisions about data have impacts across the entire information chain

descriptions are understood, used correctly, and only updated (new or removed values, change descriptions) when appropriate. Business Data Stewards also participate in "harmonizing" the values when codes and descriptions must be brought together into a system (such as a data warehouse or data lake) that gathers data from multiple sources. "Harmonizing" refers to ensuring that only values that mean the same thing are combined.

3.4.2 Technical Data Stewards

Technical Data Stewards are IT personnel who support the Data Governance effort. They are associated with specific systems, applications, data stores, ETL (extract, transform, and load) jobs, and other portions of the technical information chain. Technical Data Stewards can provide information on how the data is created, transformed, and moved, as well as how data came to be in the state currently observed. Technical Data Stewards are usually drawn from the application specialists and may change since IT departments often rotate these people to increase their range of knowledge.

The role of a Technical Data Steward is different from the various IT subject matter experts the business may be used to working with in three ways. Firstly, they are assigned the role by IT management, and working with Data Governance is an "official" part of their job. Secondly, they are responsible for providing answers in a timely manner, and providing those answers is part of their job. That is, the data management tasks are central to their role. Lastly, they are also part of the Data Stewardship team, and it is important to keep them up to date on Data Stewardship activities, goals, and tasks.

Technical Data Stewards have the following responsibilities:

- Provide technical expertise for systems, and extract, transform, and load processes, data stores, and reporting/business intelligence tools.
- Be able to clearly explain how a system or process functions.
- Be able to explain the historical reasons for the condition of the data.
- Check code, database structures (tables, views, columns, foreign keys, etc.), and other programming constructs to understand how the data is created, stored, and transformed.
- Assist in finding where business terms are physically implemented in databases and other structures.

3.4.3 Operational Data Stewards

Operational Data Stewards are basically helpers for the Business Data Stewards. They are often involved in the day-to-day maintenance of data, including the

collection and input. They are thus in a great position to notice when data is not being maintained properly, data collection rules are being violated, or the quality is in danger of being degraded due to data collection processes. Although the Business Data Stewards remain responsible for the data, the Operational Data Steward can report all these issues and help to minimize the impact.

The responsibilities of the Operational Data Steward include the following:

- Following data creation and update policies, procedures, and processes entering or modifying data. As mentioned, Operational Data Stewards are often directly involved in entering data or supervising people who do the data entry. Their duties may also include resolving mismatches and merging errors in Master Data Management.
- Help to collect data-related metrics by examining the data (including by running queries).
- Perform data analysis to assist Business Data Stewards to research and resolve data issues. The Operational Data Stewards often know which systems contain suspect data and where that data is stored as well as how the data is used. This help can make a substantial difference to the Business Data Steward's workload.
- Assist project teams that need to make changes to data. Project teams often require direct and knowledgeable help in making these changes because they are not familiar with the data.
- Identify and communicate opportunities to improve the data quality. Operational Data Stewards tend to be very close to the data because they use it every day and may even be part of the process to input or change the data. Thus, they see issues with the data quality long before it gets noticed in a database or other data store. This ability to warn the Business Data Steward about these issues can be invaluable in preventing major impacts of insufficient quality.

3.4.4 Project Data Stewards

The role of Project Data Steward helps to fill the requirement that there should be Data Stewardship representation on all major projects. It is not, however, practical to have every Business Data Steward that has project-affected data attend all the meetings and workshops just in case they might be needed. The Project Data Steward *represents* the Business Data Stewards on the projects to note when Business Data Steward participation is needed or questions need to be answered and involve the appropriate Business Data Steward(s) at that point. That is, the Project Data Steward is trained to recognize where input is needed, bring the issues and questions to the Business Data Steward to make decisions and provide information, and then bring those decisions and information back to the project team. It is important to realize that Project Data Stewards are *not* Business Data Stewards, and they do not make the decisions. The Business Data Stewards remain responsible for this work.

The responsibilities of the Project Data Stewards can be broken down into three areas: metadata, data quality, and project alignment. The metadata responsibilities include the following:

- Work with the Data Stewardship Council to identify the business function that should own new business terms identified by a project. Once the data is identified by the project, the Project Data Steward needs to work with the Business Data Stewards and the Enterprise Data Steward to identify who should own it and be responsible for identifying the metadata for the term. A proposed name and description should be identified by the project SMEs to enable the Business Data Stewards to correctly identify the owner.
- Review the business term name and sample description with the Business Data Steward to get a business definition that meets the standards for such definitions.
- For derived quantities, collect proposed calculations from the project SMEs and review with the Business Data Steward. Where there are differences, bring the corrected derivations back to the project.
- Bring Business Data Steward decisions back to the project for incorporation in the project.

Data Quality responsibilities include the following:

- Document proposed data quality rules and known data quality issues from the project SMEs. Any questions that arise about whether the quality of the data will support the project's intended usage should be documented as well. Review the rules and issues with the Business Data Steward to validate them, and when there are differences, bring those back to the project for review.
- Consult with the Business Data Steward to evaluate the impact of the data quality issues on the project data, and discuss whether the perceived issues are real, how difficult they would be to fix, and whether there is higher-quality data that the project can use instead.
- Assist in any data profiling efforts, including initial analysis of the results prior to reviewing with the Business Data Steward, and assist others to ensure that standards are followed and the results are properly documented.

The Project Alignment responsibilities include the following:

- Collaborate with the project manager and project members during the project.
- Ensure that the deliverables and concerns from Data Governance are addressed.
- Coordinate with the Business Data Stewards to collect definitions, data quality rules, and other metadata about the project's business terms.

3.5 Summary

Human Resources plays a central role in setting up a Data Governance practice – including writing job requisitions for hiring knowledgeable Data Governance professionals to staff the Data Governance Program Office and setting up bonus/management by objective plans for the new roles needed to govern the data. These roles include Business Data Stewards as well as several other types of data stewards – Technical Data Stewards, Operational Data Stewards, and Project Data Stewards. Each of these types of data stewards has a specific set of responsibilities – both individually and for groups such as the Data Stewardship Council. Other roles participate in the Executive Steering Committee and Data Governance Board. A robust Data Governance effort requires people named to the roles to effectively execute on their responsibilities.

Chapter 4
Data Value and Monetizing Data

Douglas Laney

4.1 Managing Data as an Actual Asset

In today's digital age, data has emerged as one of the most important assets for businesses. Many leaders and executives recognize this fact, and research from Gartner and other sources has shown that investors and financial analysts favor data-savvy and data-centric companies. Despite this recognition, many organizations struggle to manage their data assets with the same rigor and discipline as their traditional balance sheet assets.

This lack of formal accounting recognition is a significant problem. Many organizations collect, manage, deploy, and value their data with far less discipline than they manage their traditional balance sheet assets. This results in an unfortunate lack of inventory about what data assets exist throughout the organization.

If we consider the example of a retail manager with no record of his or her store's inventory, it is clear how ridiculous and impossible such a situation would be. Similarly, a CFO who has no general ledger that records his/her company's financial assets or an HR executive with no company directory, employee ratings, or compensation data would be operating in a completely dysfunctional environment. Yet, this is often the state of data management in most organizations today.

4.1.1 The Emergence of the Chief Data Officer

To address the need for better data management, we have seen the emergence of an executive role specifically for tending to data: the chief data officer (CDO). This

D. Laney (✉)
West Monroe, Chicago, IL, USA
e-mail: dblaney@illinois.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_4

position is a relatively new addition to the C-suite, with its rise over the past few years being an indication that organizations are getting serious about data management.

While chief data officers have been in place for decades, their focus has been on managing enterprise technologies. In contrast, the CDO's primary responsibility is to ensure that the organization's data and data assets are properly managed and leveraged to create value.

4.1.2 Approaches to Data Asset Management

Within the framework of enterprise data management (data management), there is a need for approaches to data asset management (IAM, Information Asset Management). This involves managing data assets throughout their entire life cycle, from creation to archiving, and leveraging them to drive business value.

The lack of a formalized practice of data accounting is at the core of the problem, and until senior executives and boards go beyond merely talking about data as a key corporate asset, data will continue to be a second-class business resource. Data and analytics leaders such as the CDO struggle to improve the organization's data management maturity because of the lack of a diligent program of measurable long-term improvement.

4.1.3 Data's Emergence as a Real Economic Asset

In today's digital age, data has emerged as a real economic asset. The need for effective data management initiatives is intensifying, and this demands that business leaders and IT executives recognize the importance of managing data assets as a legitimate economic asset.

However, data is not recognizable as a balance sheet asset, and therefore never managed like one. This lack of formal accounting recognition manifests in most organizations that collect, manage, deploy, and value their data with far less discipline than they manage traditional balance sheet assets.

Valuation experts and even accountants lament the challenges in valuing a company today without any data on its data. The head of data strategy for a major government military institution has proclaimed, "We have a better accounting of the toilets throughout [this building] than our data assets. And for the 'business' we're in, that's a really, really sad state of affairs."

4.1.4 The Need for Senior Executive Understanding

The fatuousness and ignorance of some executives seem to be rooted in a refusal to recognize the importance of data as a legitimate economic asset. Until senior executives go beyond merely talking about data as a key corporate asset, data will continue to be a second-class business resource.

Ultimately, this is a fundamental enterprise governance issue, and until the return on data is measured and rewarded or punished, nobody in the organization will be motivated to improve it. The lack of senior executive understanding is a significant barrier to effective data management.

Some executives fail to recognize the importance of valuing and managing their data assets, leading to a lack of discipline in caring for and leveraging these resources. For example, one executive once argued, "We don't need to know the value of our data. We don't need to concentrate on the data. It's just data." This refusal to recognize the importance of data as a legitimate economic asset results in a lack of a diligent program of measurable long-term improvement.

As a result, many companies have data management practices that pale in comparison to the rigor, process, and discipline with which they manage traditional balance sheet assets. This lack of discipline is problematic and can lead to a failure to effectively leverage data assets.

4.2 Impediments to Maturity in Enterprise Data Management

In the field of data management, executives face a number of challenges in advancing their organizations' enterprise data management (data management) capabilities. In workshops and discussions, many leaders have expressed concerns about leadership, priorities, resources, and corporate cultures that impede data management progress. These issues are particularly frustrating for executives who are acutely aware of their organization's current data management capabilities, which are often described as either "aware" or "reactive."

4.2.1 Leadership Issues

One major obstacle for many organizations is the lack of leadership in establishing data management initiatives and strategies. IT and business leaders may have different priorities, goals, and strategies, and there may be no clear consensus on the importance of data management metrics and effectiveness. The CDO role, which brings data into the heart of business planning and processes, is often absent or actively opposed. Workshop participants link this challenge to other stakeholders'

lack of business vision, cultural resistance, competing priorities, clarity and definition of strategy, and high-level support for the CDO role.

4.2.2 IM Priorities Over Which You Have Control or Influence

Effective data management for the digital business requires clear priorities that have the backing of an array of stakeholders, not just one business function. Priorities derive from a data management strategy, which in turn derives from the data management vision. However, competing priorities, lack of business vision, differing and unresolved business unit opinions, fear of losing control, disagreement over approaches, and knowing where to start are common challenges that span all seven of the data management maturity dimensions. These challenges rob decisions and actions of purpose, direction, and effectiveness, thereby reinforcing a reactive mode of operation where one's environment seems subject to forces over which one has little control.

4.2.3 Resources Needed to Advance Data Management Capabilities

Data and analytics leaders express frustration over a lack of experienced or knowledgeable staff resources, funding, domain-specific know-how, a dedicated CDO (seen as a data management resource), influence of data architects, and life cycle processes. These resources are either inadequate or totally nonexistent. Data leaders often cite lack of knowledge as a common challenge, sometimes for these leaders or their immediate organization but also for the larger organization. They claim knowledge was a scarce resource regarding what data is available, metrics, the cost of data quality issues, the role of data governance, the importance of IM, when and how to centralize or decentralize key roles, the data life cycle, and keeping current with technologies.

4.2.4 Negative Cultural Attitudes About Data Management

Negative cultural attitudes can have a significant impact on data management progress, creating inertia that is difficult to identify and overcome. Data and analytics leaders often identify culture as a serious obstacle in many of the data management maturity dimensions. Lack of cultural acceptance is an explicit problem for both the data management vision and strategy dimensions and for governance. Cultural

attitudes are implicit in other dimensions, such as data management metrics and the data management life cycle. In data management metrics, for example, basic concepts such as relating metrics to business processes and tying actions to metrics are proposed as "remedies" precisely because there is no "culture of measurement." Stuart Hamilton, senior hydrologist with Aquatic Informatics in Vancouver, British Columbia, believes the problem is deeper than just attitudinal: "data neglect is one of those things that you see every day but you don't see it because it is so much like bland wallpaper that covers everything. Once it is explained, so that you can see it as a business pathology, it resonates in many ways."

4.2.5 Overcoming the Barriers to Data Asset Management

Effective data management and governance is critical for organizations to successfully navigate the rapidly evolving digital landscape. However, according to research by James Price and Dr. Nina Evans, many executives still struggle to put in place effective mechanisms for the management or governance of data assets as an asset.

Price and Evans categorize the challenges to managing data as an asset into five broad categories:

- Awareness: Organizations lack recognition of the problem, have limited on-the-job training, and are organizationally immature.
- Leadership and Management: There is a lack of executive support, mistake intolerance, tolerance for work-arounds, no system of rewards or punishments, a lack of vision, and resistance to change.
- Business Governance: There is a lack of accountability and responsibility, responsibilities assigned at the wrong level of the organization, technology-focused IT leadership, and a lack of measurements.
- Enabling Systems and Practices: Organizations have imprecise language about data, insufficient accounting practices, technology shortcomings, and poor IT reputation.
- Justification: Organizations lack a catalyst for change; find compliance and risk burdensome; prioritize other initiatives over data governance; struggle to determine the cost, value, and benefits of data assets; and view data management as a strict process.

4.2.6 Moving Forward

While data management professionals have been aware of these challenges for decades, organizations must take concrete steps to address them. Leaders must support the development of formal training programs, establish accountability and

responsibility for data governance, and prioritize data management initiatives alongside other strategic priorities.

By overcoming these barriers, organizations can unlock the true potential of their data assets and use them as a critical tool to drive innovation and business success.

4.3 Generally Agreed-Upon Data Principles (GAIP)

As data governance and management executives, we have the opportunity to learn from other disciplines and apply their best practices to our own. In this chapter, we will explore asset management standards, principles, and methods from various domains such as physical asset management, supply chain management, IT, and software asset management. We will also examine principles from records management, intellectual property management, and library science, among others.

To frame a set of data asset management doctrine, we can take inspiration from Generally Accepted Accounting Principles (GAAP). The framework comprises a set of principles based on fundamental assumptions and tempered by a set of constraints. While GAAP provides guidance for preparing financial statements, it can also provide a useful way to express a concise set of GAIP (see Table 4.1).

For data governance and management executives, it's crucial to have a set of principles that guide the organization's strategy, operations, and decision-making. That's where GAIP come in. These principles are not specific to any industry or organization, making them adaptable to virtually any company.

Adopting GAIP as a foundation for data governance and management can help to establish a concise, clear, and widely accepted set of principles that can be used as a reference point for the organization's data management practices. These principles can help ensure that the organization manages its data assets in a way that maximizes their value, promotes accountability, and aligns with regulatory and legal requirements.

4.4 Data Supply Chains and Ecosystems

The concept of a "data supply chain" (ISC) was introduced in the early days of data warehousing, as professionals began to see the value of treating the production, flow, enhancement, and availability of data as a type of supply chain. The ISC is a useful metaphor for visualizing, defining, refining, and assessing the processes and resources involved in the data life cycle. The supply chain is designed with the customer in mind, so it can help data management professionals keep the business outcomes of deployed data assets in mind.

A supply chain is a system of activities and resources involved in moving a product or service from the point where it is manufactured to where it is consumed. In an ISC, *data is the raw material, and data is the product*. However, adding value

Table 4.1 Generally Accepted data Principle (GAIP)

Assumptions	Assumptions are agreed-upon basic beliefs about data. They guide our understanding of how data assets can and should be perceived, managed, and deployed	1. **Asset assumption**: Data is an asset because it meets each of the criteria of an asset 2. **Proprietorship assumption**: An organization's data assets include all forms of data and content of discernible identifiability for which it can claim ownership and/or exclusive control 3. **Appraisal assumption**: Data has realized, probable, and potential cost and value 4. **Dominion assumption**: The practice of internal data "ownership" limits its potential value to the organization and thereby the performance of the organization itself 5. **Benefit assumption**: Data has uses well beyond its original purpose, does not deplete when used, and can be used simultaneously for different purposes
Constraints	Constraints are generally agreed-upon data regulations, confinements, or bounds. They acknowledge the limits of how well or precisely data assets can be monetized, managed, and measured and therefore restrict how absolutely the principles which follow can be applied	6. **Specificity constraint**: The groupings of data or content that comprise a "data asset" will vary from one organization or use case to the next 7. **Recognition constraint**: Data cannot be represented in auditable financial statements, nor be capitalized as other assets (per current accounting standards) 8. **Jurisdiction constraint**: The provenance, lineage, ownership, and sovereignty of a data asset may be difficult to determine or legally establish 9. **Valuation constraint**: Valuation and other measurements of a data asset will be inexact but useful, just as are valuations of other kinds of assets 10. **Resource constraint**: Trade-offs among data asset quality, availability, and accessibility are inevitable
Principles	Principles are generally agreed-upon axioms that dictate how data assets should be managed and should lead to more detailed guidelines, policies, procedures, and standards specific to the organization	11. **Relevance principle**: Data assets should be managed with at least the same discipline as other recognized assets 12. **Inventory principle**: Data assets should be cataloged, described, classified, related, and tracked 13. **Ownership principle**: By default, data assets belong to the organization, not any application, department, or individual

(continued)

Table 4.1 (continued)

		14. **Authorization principle**: The quality requirements, access, use, protection, and other rights and responsibilities for any data asset, even within the organization, should be contractually established by or with a sanctioned and empowered trustee 15. **Assessment principle**: The quality characteristics, cost, value, and risks of any data asset should be knowable at any point in time and used for prioritizing and budgeting data-related initiatives 16. **Possession principle**: A data asset should be acquired or retained only if its actual or planned value is greater than its cumulative cost, or as required by laws or other regulations 17. **Replicability principle**: A data asset should be duplicated or derived only to improve its utility or availability and only if doing so also increases its net value 18. **Optimization principle**: (a) The business is responsible for optimizing the usage and understanding of data. (b) The data management organization is responsible for optimizing data's availability and utility. (c) The technology organization is responsible for optimizing data's accessibility and protection

to data to turn it into data is rarely a simple process. Data in the ISC context can be original transactions, text files, emails, images, or other similar items that often only have value in the context of the process that created or captured them.

4.4.1 Adapting the SCOR Model

As data management and governance executives, it is important to understand and apply supply chain best practices to the data supply chain (ISC) in order to ensure a seamless flow of data from acquisition to delivery. The *Supply Chain Operations Reference (SCOR)* model provides a framework for ISC planning, which includes processes for planning, sourcing, making, delivering, returning, and enabling. By adapting these processes to the ISC, organizations can plan for costs, manage inventory, handle payments and revenues, and transmit and receive data securely.

The SCOR model also provides a few levels of detail for scoping, configuring, and process/performance attributes, which can enable handling of specific supply chain scenarios such as "make-to-stock" versus "make-to-order" supply chain configurations for general and custom goods and services, respectively. Differentiating these two configurations for the supply of data can be helpful in designing for generalized data uses such as a data warehouse or specified data purposes such as an architected data mart or report.

As ISC grows more sophisticated, it behaves more as networks with complex flows of goods and services among suppliers, distributors, payment processors, and customers. Metrics for ISC can include costs, cycle times, return on assets and working capital, demand planning and management, inventory recording practices, and dozens of other procedures and considerations. ISC metrics can be used to manage and monitor the flow of data and to identify areas for improvement in the ISC process.

Overall, the application of SCOR model to the ISC is crucial for ensuring the efficient flow of data and the achievement of business outcomes. By planning for costs, managing inventory, and monitoring ISC metrics, data management and governance executives can ensure that data is acquired, managed, and transmitted in a secure and efficient manner.

4.4.2 Metrics for the Data Supply Chain

As data governance and data management professionals, it is essential to measure and optimize the performance of the data supply chain. The SCOR model provides a useful framework for defining performance attributes and metrics.

Table 4.2 presents a summary of the performance attributes, classic supply chain attribute definitions, and sample data supply chain metrics that can be used to measure the performance of the data supply chain.

Measuring the performance of the data supply chain is critical for optimizing data governance and management practices. The SCOR model provides a useful framework for defining performance attributes and metrics. By monitoring and improving these metrics, organizations can ensure that their data supply chains are reliable, responsive, agile, cost-effective, and efficient in their asset management.

4.5 A New Model for the Data Supply Chain

As data management and governance executives, it's important to have a model for describing the flow of data assets that centers on how each step increases their economic potential. While the classic product/service supply chain model is useful at a high level, it becomes increasingly unrelated to the specific processes relevant to the management and flow of data.

Table 4.2 Metrics for the data supply chain

Performance attribute	Classic supply chain performance attribute definition	Sample data supply chain performance metrics
Reliability	The ability to perform tasks as expected. Reliability focuses on the predictability of the outcome of a process	– Query/update performance, data quality (accuracy, completeness, timeliness, integrity, etc.)
Responsiveness	The speed at which tasks are performed. The speed at which a supply chain provides products to the customer. Examples include cycle-time metrics	– Data accessibility, user request turnaround time, user satisfaction survey
Agility	The ability to respond to external influences and market changes to gain or maintain competitive advantage	– Utility of data for a range of purposes; linked data, metadata, and master data measures; ease of integrating new types of data or changing dimensions
Costs	The cost of operating the supply chain processes. This includes labor costs, material costs, management, and transportation costs. A typical cost metric is cost of goods sold	– Data acquisition cost, data management costs, data delivery costs (each including labor and technology-related costs)
Asset management efficiency (assets)	The ability to efficiently utilize assets. Asset management strategies in a supply chain include inventory reduction and insourcing versus outsourcing. Metrics include inventory days of supply and capacity utilization	– Data timeliness, amount of available history, actual usage (e.g., percent of data touched by users/apps)

To create a more relevant model, we can examine a range of different kinds of recognized assets: material assets, financial assets, intellectual property, human capital, and data "assets." In the next section, we will explore the specific processes and standards for each of these assets.

While financial and material assets are somewhat obvious, it's important to recognize employees as "human capital." This concept emerged in the 1960s with the publication of Gary Becker's book *Human Capital*. Today, human resources executives and the concept of human capital are widely used in business. However, employees are not recognized as assets on the balance sheet because ownership and control are key asset determinants, and employees are considered "at will" and not owned by the organization.

Similarly, data is not recognized as an asset according to accounting standards. However, we execute a similar set of activities on assets to ensure they don't lose value and can generate future value. We collect or obtain assets, produce, and inventory them, enhance their potential economic benefit, move them, integrate them, and protect them.

By recognizing data as an economic asset and creating a supply chain model that centers on its economic potential, organizations can more effectively manage and

Table 4.3 Activities for data supply chain (ISC)

Collect	Prepare	Combine	Enrich
Produce	Inventory	Locate	Secure
Organize	Distribute	Govern	Monitor

Table 4.4 Fundamental activities to execute over data in the data supply chain (ISC)

Sell	Lend or license	Share
Spend	Trade	Apply

govern their data assets. This can lead to better decision-making, increased efficiency, and greater value creation.

These life cycle primitives classify the activities we do to assets and represent the supply side of the supply chain (see Table 4.3):

These activities are familiar when compared to the SCOR framework and can be applied to any class of asset (or *proto-asset*). There is no specific order or sequence of steps implied, as they can be combined and sequenced as necessary. The activities focus solely on augmenting the value of the asset and do not facilitate its realization. In order to realize the economic value of an asset, we must take action with it. These fundamental activities categorize the actions we take with assets and represent the demand side of the supply chain (see Table 4.4).

In the case of financial assets, we typically spend or invest them to meet our demands, whether it's for personal or business needs. Material assets are sold or used to produce goods and services to meet the demands of customers. Human assets are utilized to meet the demands of business processes, and intellectual property is often utilized to meet the demands of innovation and competitive advantage. Similarly, data, as a valuable asset, can be utilized to meet the demands of various business processes and enable better decision-making. By understanding the supply and demand of data within an organization, data management and governance professionals can better identify opportunities to maximize the economic potential of their data assets.

Figure 4.1 illustrates a continuum that shows how data value potential is enhanced leading to its realization, with three main stages of a data supply chain (ISC): acquisition, administration, and application. The ISC may intersect with another organization's ISC in a data supply network, where the arrow may loop back on itself. This is due to the nondepleting, non-rivalrous, pro-generative nature of data, where sold, lent, or analyzed data may become raw data for another organization, and so on.

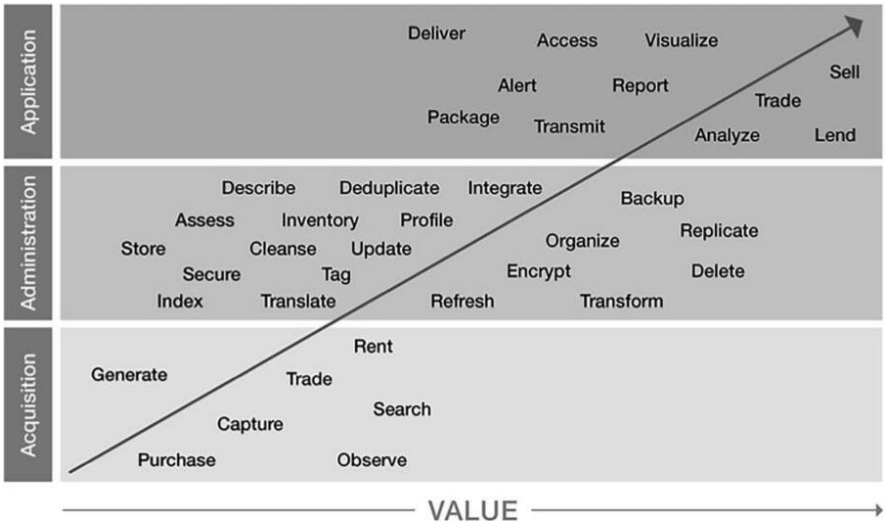

Fig. 4.1 Data value increments through stages of data supply chain

4.6 Data Ecosystems

As the business environment becomes more dynamic, it is important to rethink the way we view the flow of data. While supply chains and supply networks have been useful models, they can be too linear and procedural for today's needs. Instead, we can view data as part of an ecosystem, which allows for more adaptability and responsiveness to environmental changes.

In Japan, *keiretsus* are well-known corporate ecosystems built around trust, collaboration, and coordination. Similarly, companies like Walmart and Coca-Cola have formed data *keiretsus*, which enable partners to easily share and utilize each other's data. This behavior of data within and among entities is similar to something flowing or thriving within an ecosystem.

The importance of data flow is even more pronounced as businesses turn to ecosystems to fuel their digital growth. Top-performing companies create or participate in ecosystems and expect to double their ecosystems in 2 years. This shift to a more dynamic networked digital ecosystem requires a rethinking of the traditional linear value chain business model.

In the context of data management, an ecosystem allows for a more adaptive and responsive approach. It is important to understand what an ecosystem is and how it works, so that we can apply its principles to data management. An ecosystem is a community of living and nonliving things that interact with each other in a specific environment. In a data ecosystem, this community includes data, people, technology, and processes, which interact and influence each other. By understanding and utilizing these interactions, we can create a more effective and efficient data management system.

Ecosystem
[ek-oh-sis-tuh m, ee-koh-]
noun, Ecology.

1. a system, or a group of interconnected elements, formed by the interaction of a community of organisms with their environment.
2. any system or network of interconnecting and interacting parts, as in a business.

An ecosystem can be defined as a community of organisms along with the inanimate parts of their environment, linked via nutrient cycles and energy flows. While the web may be considered a global data ecosystem, it is more useful to consider ecosystems on a more localized scale.

4.6.1 Data Within an Ecosystem

Thinking of data as a resource or energy source, such as "the oil of the 21st century," is a common analogy. However, it disregards the unique economic and behavioral characteristics of data. Alternatively, we can think of data within an ecosystem as an organism itself. This perspective suggests that data is born, thrives, replicates, evolves, and is affected by climate and topography. Data doesn't have DNA within it to program its behavior, but emerging technologies are beginning to shift the processing to the data, suggesting a more inside-out approach to data processing.

In fact, some organizations such as the New York Stock Exchange and retail market intelligence company IRI offer analytic environments for customers to process data in situ rather than extracting and downloading it, which reflects an ecosystem-like perspective on data processing.

It is important to note that viruses can infect data just as they can infect systems, suggesting that we must be mindful of the security of our data ecosystem. As an industry, we tend to use related terms such as "value," "asset," "life cycle," and "management" without a common understanding. By examining classic, biological ecosystem concepts, we can better adapt them to explain the world of data. As we move toward a *more dynamic business environment*, understanding the concept of a data ecosystem and its implications for data management will be crucial to our success.

In the digital age, it is natural to think of our data and data as part of a complex and dynamic ecosystem. Just like in a biological ecosystem, the various components of the data ecosystem interact with each other in a network of processes and systems.

4.6.2 Ecosystem Entities

In a biological ecosystem, organisms, organic matter, nutrients, and energy are the main actors. In the data ecosystem, however, data is the central focus. Additionally, resources such as processing power, storage, and bandwidth are also critical components. The "nutrients" that support the growth of data are events, such as transactions, that add to the datasets.

4.6.3 Ecosystem Features

Both biological and data ecosystems involve interactions among the organisms or components and with the environment. In data ecosystems, these interactions occur during processes such as lookups, queries, and reporting. The system architecture and business climate also play a role in the ecosystem's topography and climate. Like biodiversity in biological ecosystems, *infodiversity* is an important feature of data ecosystems, providing the variety of data upon which businesses and consumers depend.

4.6.4 Ecosystem Processes

In biological ecosystems, energy flows, nutrient cycling, and the movement of matter are the primary sub-processes. In the data ecosystem, similar processes occur, such as the filtration, cleansing, and application of algorithms to alter data. Reproduction of data involves making copies or extracts of it. Movement of data is crucial, and growth is due to nutrients and available resources.

4.6.5 Ecosystem Influences

Disturbances or occurrences influence ecosystems, and it is essential to prepare for such events. Security breaches, natural disasters, new competitors, or business collapses can cause disturbances to the data ecosystem. Such events may require structural changes in the way we manage and leverage data.

4.6.6 *Ecosystem Management*

To ensure the optimal production and consumption of organisms in biological ecosystems, ecosystem managers may introduce or reduce resources, supplement resources, or artificially repair organism imbalances. Similarly, in the data ecosystem, ecosystem managers must maintain the optimal production and consumption of data. They perform similar tasks, such as reconfiguring hardware and networks, cleansing data, and backing up data to prevent its loss. Effective ecosystem management requires a comprehensive vision, strategy, governance, and tools.

4.7 Applying Sustainability Concepts to Managing Data

The six "R"s of sustainability provide a useful framework for managing data as an asset. By adopting these principles, organizations can improve their data management and governance strategies, reduce costs, and minimize their environmental impact. By **r**efusing unnecessary data, **r**educing data storage, **r**eusing data, **r**epurposing data, **r**ecycling data, and **r**emoving data, organizations can create a more sustainable and effective data management strategy.

Refuse: The first step in managing data as an asset is to refuse any unnecessary data. This means that organizations should only collect and store data that is essential for their business operations. Refusing data can help reduce storage costs and simplify data management. It also helps organizations comply with data privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), which require organizations to minimize the amount of personal data they collect and process.

Reduce: The second "R" of sustainability is to reduce the amount of data that is collected and stored. Organizations should regularly review their data storage practices and eliminate any *redundant, outdated, or trivial* (ROT) data. This can help reduce storage costs and improve the overall quality of data. By reducing the amount of data they store, organizations can also improve their data security, as they will have fewer data sources to secure and protect.

Reuse: The third "R" of sustainability is to reuse data whenever possible. Organizations should establish a data reuse policy that encourages data sharing and collaboration across different departments and business units. This can help improve decision-making, reduce duplication of effort, and improve overall efficiency. By reusing data, organizations can also reduce the amount of data they need to collect and store, which can help improve data quality and reduce costs.

Repurpose: The fourth "R" of sustainability is to repurpose data for different use cases. Organizations should explore new ways to use their existing data assets to create new business value. This could involve combining different data sources to create new insights or using data to train machine learning models. By repurposing

data, organizations can unlock new business opportunities and improve their competitiveness.

Recycle: The fifth "R" of sustainability is to recycle data. Organizations should consider the environmental impact of their data management practices and adopt strategies to minimize their carbon footprint. This could involve using energy-efficient data storage solutions or using renewable energy sources to power data centers. By adopting sustainable data management practices, organizations can reduce their environmental impact and contribute to a more sustainable future.

Remove: The final "R" of sustainability is to remove data that is no longer needed. Organizations should establish a data retention policy that specifies how long different types of data should be kept and when it should be deleted. This can help reduce storage costs and improve data security. It also helps organizations comply with data privacy regulations that require the deletion of personal data after a certain period.

4.8 Data Management Standards

There are many data management standards in existence today, each designed to address different aspects of data management and governance. Some of the most widely used data management standards include:

- ISO 8000-1x0: This standard specifies a set of data quality requirements and metrics for data exchange between organizations. It provides guidelines for data formatting, encoding, and validation and helps ensure that data is accurate, complete, and consistent.
- ISO 27001: This standard provides a framework for data security management. It specifies a set of policies and procedures for protecting sensitive data and data assets from unauthorized access, disclosure, and destruction.
- GDPR: The General Data Protection Regulation is a set of regulations developed by the European Union to protect the privacy and security of personal data. It requires organizations to implement strong data governance and security policies and to obtain explicit consent from individuals before collecting and processing their personal data.
- HIPAA: The Health Insurance Portability and Accountability Act is a US regulation that governs the use and disclosure of protected health data (PHI). It sets standards for data privacy, security, and breach notification and requires organizations to implement comprehensive data management and security practices to protect PHI.
- COBIT: The Control Objectives for Information and Related Technology is a framework developed by the Information Systems Audit and Control Association (ISACA). It provides guidelines for IT governance, risk management, and compliance and helps organizations align their IT operations with their business goals.

- DAMA-DMBOK: The Data Management Body of Knowledge is a framework developed by the Data Management Association (DAMA). It provides a comprehensive guide to data management best practices and techniques and helps organizations establish effective data management strategies.
- Open Data Initiative: The Open Data Initiative is a collaborative project between Microsoft, Adobe, and SAP that aims to promote data interoperability and data exchange between different applications and platforms. It provides a set of standards for data modeling, data format, and data exchange and helps organizations share data across different systems.
- DCAM: The Data Capability Assessment Model (DCAM) is a framework developed by the Data Management Association (DAMA) and the Enterprise Data Management Council (EDMC) to assess an organization's data management capabilities. It provides a maturity model for data management and helps organizations identify areas for improvement.
- ISO 15489: The International Organization for Standardization (ISO) 15489 is a standard for records management that provides guidance on the creation, management, and disposition of records. It helps organizations ensure that their records are accurate, complete, and accessible over time.
- CMMI: The Capability Maturity Model Integration (CMMI) is a framework developed by the Software Engineering Institute that provides guidance on software development but also includes a data management maturity model. The data management maturity model helps organizations assess their data management capabilities and identify areas for improvement.
- DAMA CDMP: The Certified Data Management Professional (CDMP) program is a certification program developed by DAMA that provides a standardized approach to assessing and validating data management knowledge and skills. The program covers 14 data management disciplines and provides a useful benchmark for individuals seeking to develop their data management expertise.
- FAIR: The Findable, Accessible, Interoperable, and Reusable (FAIR) data principles provide a framework for making research data more discoverable, accessible, and reusable. They provide guidelines for data management that are particularly relevant to scientific research but can be applied to other domains as well.

Overall, these standards can provide a valuable guide for organizations seeking to establish effective data management and governance practices. By adopting these standards, organizations can ensure that their data is accurate, secure, and compliant with regulatory requirements and can build a strong foundation for data-driven decision-making and business success.

However, there are several limitations to their implementation. Some of the key limitations include:

- Compliance vs. implementation: While standards provide useful guidelines for data management, they do not guarantee successful implementation. Many organizations may struggle to implement data management standards due to lack of resources, expertise, or cultural barriers.

- Rapidly evolving technology: Data management standards may become outdated or irrelevant as technology evolves. For example, emerging technologies such as AI and machine learning may require new data management approaches that are not covered by existing standards.
- Cost: Implementing data management standards can be expensive, especially for smaller organizations with limited resources. Organizations may need to invest in new technology, staff training, and consulting services to meet the requirements of data management standards.
- Complexity: Data management standards can be complex and difficult to understand for nonexperts. This can lead to confusion and misinterpretation of the standards, which may result in ineffective or inefficient data management practices.
- Lack of harmonization: There are many data management standards in existence, and they are often developed independently by different organizations or regulatory bodies. This can lead to inconsistencies and conflicts between standards, which can make it difficult for organizations to achieve compliance with multiple standards.
- Cultural barriers: Data management standards may be met with resistance from stakeholders who are unwilling or unable to change their existing practices. This can result in poor adoption rates and suboptimal data management practices.

While data management standards provide useful guidance for organizations seeking to establish effective data management and governance practices, they are not without limitations. To successfully implement data management standards, organizations must carefully consider the cost, complexity, and cultural factors that may impact their implementation. They should also be mindful of the rapidly evolving technology landscape, which may require new approaches to data management that are not yet covered by existing standards.

4.8.1 Adapting IT Asset Management (ITAM) to Data Management

The ISO 19770 family of standards for ITAM provides a process defining best practices for software asset management (SAM), an XML standard for inventorying and identifying software deployed on devices, a schema for describing entitlements and rights associated with software licenses, and a standard for reporting on resource utilization. These standards help educate end users on compliance, aid budget managers in making technology redeployment decisions, guide IT service departments on warranty and other service data, and offer procedures on invoice and inventory level data for finance departments.

Substituting the phrase “data asset” for “technology” or “IT asset,” one may ask if data management departments or leaders have a global standard for data best practices, an inventory standard for data assets, a standard way to document

contractual rights and privileges for data usage, or a recognized standard for reporting on data utilization. The answer to any of these is hardly, and yet, data assets are critical to organizations.

4.8.2 Adapting ITIL to Data Management

The Information Technology Infrastructure Library (ITIL) is a widely adopted framework for IT service management that provides a comprehensive set of best practices for managing IT services. Although ITIL was not specifically designed for managing data assets, many of its principles and processes can be adapted to managing data assets effectively.

One of the key principles of ITIL is the focus on delivering value to the business. In the context of data asset management, this means that data assets should be managed with a clear understanding of their business value and with a focus on ensuring that they meet the needs of the organization.

Another key principle of ITIL is the focus on service management processes, including service design, service transition, service operation, and continual service improvement. These processes can be adapted to managing data assets by establishing processes for data asset design, implementation, operation, and improvement.

For example, in the service design phase, organizations can establish a data asset design process that includes requirements gathering, data modeling, and data quality assessment. In the service transition phase, organizations can establish a process for data asset implementation, including data migration, testing, and training. In the service operation phase, organizations can establish a process for monitoring and maintaining data assets, including data backup and recovery, access control, and data quality monitoring. In the continual service improvement phase, organizations can establish a process for evaluating data asset performance, identifying areas for improvement, and implementing changes to improve data asset management practices.

ITIL also emphasizes the importance of service level management, which involves defining and managing service level agreements (SLAs) with internal and external stakeholders. In the context of data asset management, this means that organizations should establish SLAs for data assets, including data quality, availability, and security. This can help ensure that data assets are meeting the needs of the organization and that stakeholders are aware of the level of service they can expect from data assets.

Finally, ITIL emphasizes the importance of continual service improvement, which involves regularly reviewing and improving IT services to ensure they are meeting the needs of the organization. This principle can be applied to data asset management by establishing regular reviews of data asset performance, identifying areas for improvement, and implementing changes to improve data asset management practices.

By adopting a service management approach to data asset management and focusing on delivering value to the business, establishing service level agreements for data assets, and implementing a continual service improvement process, organizations can establish effective data asset management practices that meet the needs of the organization.

4.8.3 Adaptations from RIM and ECM

Records Information Management (RIM) and Enterprise Content Management (ECM) are two related concepts that can be applied to managing data assets effectively. RIM focuses on the systematic management of records throughout their life cycle, while ECM focuses on the management of digital content, including documents, images, and multimedia.

One of the key principles of RIM is the need to establish clear policies and procedures for managing records. In the context of data asset management, this means that organizations should establish clear policies and procedures for managing data assets, including data quality, data retention, and data security. By establishing clear policies and procedures, organizations can ensure that data assets are managed in a consistent and effective manner and that stakeholders are aware of their responsibilities for managing data assets.

Another key principle of RIM is the importance of identifying and classifying records according to their business value. In the context of data asset management, this means that organizations should identify and classify data assets according to their business value and establish appropriate retention policies for each type of data asset. This can help to ensure that data assets are managed effectively throughout their life cycle and that they are retained for the appropriate length of time.

ECM emphasizes the importance of managing digital content throughout its life cycle, from creation to disposal. In the context of data asset management, this means that organizations should establish processes for managing data assets throughout their life cycle, including data creation, data capture, data storage, data retrieval, and data disposal. By establishing clear processes for managing data assets throughout their life cycle, organizations can ensure that data assets are managed effectively and efficiently and that they are disposed of in a secure and responsible manner.

4.8.4 Adaptations from Library Science

Perhaps somewhat surprisingly, the field of library and information science (LIS) offers valuable insights and best practices for managing data assets effectively. While the origins of LIS can be traced back to the seventeenth century, its principles continue to be relevant today. Gabriel Naudé, a French librarian who published a text on library operations in 1627, offered valuable insights into the creation and

management of libraries. His principles include the importance of collecting and sharing human knowledge, inspecting the catalogs of other libraries, focusing on the most important data first, and organizing data assets in a way that is easy to locate and nearby other data assets of similar topic interest.

These principles can be adapted to managing data assets by recognizing that there is no greater asset than data, learning what data assets are collected and compiled by competitors and others, focusing on the most important data first, and ensuring the availability of high-demand data assets. Additionally, organizations can collect data from respected sources, capture raw, original data whenever possible, include available metadata on data, recognize that all data has potential and probable value to someone or some process, organize data assets in a way that is easy to locate, and ensure the protection and preservation of data assets.

The International Federation of Library Associations and Institutions (IFLA) is the leading governing body for LIS today. Its principles include the promotion of high standards for the provision and delivery of library data services, encouraging widespread understanding of the value of good library and data services, and endorsing the principles of freedom of access to data. Over the past few decades, LIS has been transformed by the digital age, and the IFLA has developed and published conceptual models and digital formats for bibliographic encoding and sharing and for resource descriptions. Additionally, it offers formal guidelines on the handling and storage of various media, content curation and sharing, artifact digitization and preservation, and overall operations. These guidelines can provide CDOs and other data professionals with fascinating insights and useful ideas to bring into the data asset management fold.

4.8.5 Adaptations from Physical Asset Management

PAS 55, which serves as the basis for the ISO 55001 standard, provides a framework for physical asset management that can be adapted to manage data assets effectively. While the original standard focuses on managing physical assets such as equipment and infrastructure, its principles can be applied to managing data assets in a similar manner.

The *first step* in adapting PAS 55 for managing data assets is *to establish clear policies and procedures for managing data*. This includes developing a data governance framework that defines the roles and responsibilities of stakeholders, as well as policies for data quality, data retention, data security, and data privacy. By establishing clear policies and procedures, organizations can ensure that data assets are managed in a consistent and effective manner and that stakeholders are aware of their responsibilities for managing data assets.

The *second step* is *to identify and classify data assets according to their business value*. This includes establishing a data inventory that lists all data assets and their associated metadata, as well as developing a data classification scheme that categorizes data assets according to their criticality, sensitivity, and other relevant factors.

By identifying and classifying data assets, organizations can ensure that data assets are managed effectively throughout their life cycle and that they are retained for the appropriate length of time.

The *third step* is *to develop an asset management plan for data assets*. This plan should include strategies for acquiring, maintaining, and disposing of data assets, as well as procedures for monitoring and reporting on the performance of data assets. By developing an asset management plan for data assets, organizations can ensure that data assets are managed in a way that maximizes their value and meets the needs of the organization.

The *fourth step* is *to establish performance metrics for data assets*. This includes identifying key performance indicators (KPIs) that measure the effectiveness and efficiency of data asset management, such as data quality, data availability, and data security. By establishing performance metrics for data assets, organizations can monitor and continuously improve their data asset management practices.

In short, PAS answers five key questions:

1. What assets do you have?
2. What is your risk of an asset-related disaster?
3. Do you know the current condition of your assets?
4. What are the costs of corrective versus preventative maintenance?
5. Should you repair or replace any given asset?

Shouldn't any CDO or data governance lead have the answers to questions regarding the organization's data assets?

4.8.6 Adaptations from Financial Management

Financial management standards provide valuable insights into how organizations can better manage their data assets. In particular, the roles and responsibilities of a trustee or fiduciary can be adapted to help organizations manage their data assets more effectively.

A trustee or fiduciary is responsible for managing assets on behalf of a beneficiary or stakeholders. This includes developing investment strategies, monitoring the performance of investments, and reporting on the performance of assets to stakeholders. These responsibilities can be adapted to managing data assets in several ways.

First, *organizations can appoint a data trustee or data fiduciary* to manage their data assets. This individual or team would be responsible for ensuring that data assets are managed in a way that maximizes their value and meets the needs of the organization. This includes developing a data strategy, monitoring the performance of data assets, and reporting on the performance of data assets to stakeholders.

Second, the *data trustee or data fiduciary should establish clear policies and procedures for managing data assets*. This includes developing a data governance framework that defines the roles and responsibilities of stakeholders, as well as

policies for data quality, data retention, data security, and data privacy. By establishing clear policies and procedures, organizations can ensure that data assets are managed in a consistent and effective manner.

Third, the *data trustee or data fiduciary should identify and classify data assets according to their business value*. This includes establishing a data inventory that lists all data assets and their associated metadata, as well as developing a data classification scheme that categorizes data assets according to their criticality, sensitivity, and other relevant factors. By identifying and classifying data assets, organizations can ensure that data assets are managed effectively throughout their life cycle and that they are retained for the appropriate length of time.

Fourth, the *data trustee or data fiduciary should develop an asset management plan for data assets*. This plan should include strategies for acquiring, maintaining, and disposing of data assets, as well as procedures for monitoring and reporting on the performance of data assets. By developing an asset management plan for data assets, organizations can ensure that data assets are managed in a way that maximizes their value and meets the needs of the organization.

Even the responsibilities of a chief data officer (CDO) and a chief financial officer (CFO) share several similarities, as both roles involve managing important organizational assets and providing strategic guidance for the company. Some of the key parallels between the roles of a CDO and CFO include:

- Asset management: Just as a CFO is responsible for managing the financial assets of an organization, a CDO is responsible for managing the data assets. Both roles require identifying the assets, tracking their performance, and maximizing their value to the organization.
- Strategic planning: Both the CDO and CFO play a crucial role in developing and implementing the strategic plans of the organization. They provide guidance on how to use the assets in a way that meets the needs of the organization and its stakeholders.
- Risk management: Both roles are responsible for identifying and mitigating risks associated with their respective assets. For example, a CFO might manage financial risks such as credit risk and market risk, while a CDO might manage risks associated with data quality and data privacy.
- Reporting: Both the CFO and CDO are responsible for providing accurate and timely reporting to stakeholders. The CFO provides financial reports, while the CDO provides data reports to ensure that data is being used effectively to drive business outcomes.
- Compliance: Both the CFO and CDO must ensure that the organization complies with applicable laws and regulations related to their respective assets. For example, a CFO must ensure that financial reporting is in compliance with accounting standards, while a CDO must ensure that data privacy regulations are being followed.

Chapter 5
Data Governance Methodologies: The CC CDQ Reference Model for Data and Analytics Governance

Christine Legner, Martin Fadler, and Tobias Pentek

5.1 Introduction

For most companies – digital natives as well as incumbents – data have turned into strategic assets which they can directly or indirectly monetize through new business models, data-driven insights, and improved business processes. As the importance of data is increasing, so does the awareness that data governance plays a critical role in leveraging the value of data and analytics [1–3]: In fact, "without appropriate organizational structures and governance frameworks in place, it is impossible to collect and analyze data across an enterprise and deliver insights to where they are most needed" [1, p. 417]. Having clear responsibilities ensures that data is "fit for purpose" for analytics and other use cases and that data issues are solved. While data governance undeniably is the foundation for sustainable data quality improvements and for regulatory compliance, it is increasingly considered an important enabler of value creation and data-driven innovations.

Despite the increasing awareness, many organizations still struggle with implementing effective data governance. On the one hand, it is demanding to get management support and justify investments in data governance programs. The value from these programs is difficult to demonstrate and measure, as it is mostly indirect – without data governance organizations may miss out on data-driven innovation, waste employee's resources for non-value-adding tasks, and increase their risks of noncompliance with an increasing number of regulations [4]. On the other hand, implementing and scaling data governance in medium to large

C. Legner (✉) · M. Fadler
Faculty of Business and Economics (HEC), University of Lausanne, Lausanne, Switzerland
e-mail: christine.legner@unil.ch; martin.fadler@unil.ch

T. Pentek
CDQ AG, St. Gallen, Switzerland
e-mail: tobias.pentek@cdq.ch

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_5

organizations is far from being trivial. Inside the organizations, data governance knowledge is scarce and often tacit and has traditionally focused on control and compliance for a small subset of enterprise data, most importantly master data. These traditional governance approaches are often perceived as overly rigid and constraining when it comes to satisfying the increasing demand for data and using them in innovative scenarios. Thus, they fall short of providing a comprehensive guideline to govern data management and analytics delivery with the overarching goal to support data-driven innovations.

To conclude, we lack methodological guidelines that go beyond outlining roles and responsibilities (i.e., structural governance) and extend data governance's focus to enable and maximize value creation from data and analytics. To address these gaps, this chapter presents a reference model as a three-step approach toward data and analytics governance, which has been developed in an industry-research collaboration and tested with companies from different industries. It presents the view of the Competence Center Corporate Data Quality (CC CDQ), which unites 20 multinational companies and researchers in the field of data management. In this chapter, we will first elaborate on the foundations and paradigm shifts in data governance before we elaborate on key principles for effective governance design. We will present each of the three steps of the CC CDQ Reference Model for Data and Analytics Governance in detail.

5.2 Paradigm Shifts in Data Governance: From Control to Value Creation

5.2.1 Data Governance: Definition and Mechanisms

The term *governance*, originating from old French, refers to "the way that organizations or countries are managed at the highest level, and the systems for doing this" [5]. Governance should, thereby, not be confused with management. While *governance* assigns the fundamental accountabilities and builds the organizational structure that sets the guardrail for value generation, *management* uses this governance system to allocate resources and run day-to-day operations [6]. In enterprises, different governance systems exist that aim to moderate value generation from specific investments and comprise, for instance, corporate governance, IT governance, and data governance.

Building on these foundations, data governance defines the framework with the decision rights and accountabilities for the management and use of data [7]. It encourages desirable behavior concerning the conduct of data within an organization, by defining the policies, procedures, and standards for the effective use of an organization's structured and unstructured data assets.

Data governance is often associated with a set of generally applicable governance mechanisms that are borrowed from IT and corporate governance literature

[8, 9]. They can be classified into (1) structural governance mechanisms that define the organizational structure and assign responsibilities, (2) procedural governance mechanisms that define and structure decision-making processes, and (3) relational governance mechanisms that focus on collaboration, communication, and knowledge sharing.

5.2.2 *Data Governance 1.0: Focus on Control, Data Quality, and Regulatory Compliance*

Data governance has traditionally focused on data quality and regulatory compliance (*Data Governance 1.0*) as main goals and thereby emphasized control over data. In this defensive orientation, dedicated data management teams are in charge of improving the quality of corporate data residing in operational systems and most importantly master data, for example, master data on materials, suppliers, and customers. Analytics teams oversee data quality in data warehouses and business intelligence (BI) tools that deliver financial or other corporate reports. In these controlled environments, major effort is invested up front to clean data at the source and then load it into a pre-defined schema (schema-on-write) to achieve a single version of the truth (SVOT).

5.2.3 *Data Governance 2.0: Extending Beyond Control to Enable Value Creation*

With the explosion of data and the widespread adoption of data science, enterprises seek new value creation opportunities from data and aim at monetizing it in indirect or direct form. The view of data as an asset and the reuse of data for a variety of analytical purposes, however, have direct implications on the way how they are governed and eventually managed. On the one side, a more flexible approach is required to explore and experiment with data from different sources. This implies a shift from data warehouses as controlled analytics environments to more flexible data lake infrastructures. Here, data from multiple sources are loaded without a pre-defined structure (schema-on-read) in their "raw" format to enable multiple versions of the truth (MVOT). The up-front effort for cleaning and integration is thereby kept to a minimum. Data lakes are not only used to explore data and develop data science pipelines, but they also serve data science pipelines in production which are used in downstream systems to enhance day-to-day operations. Thus, the dependencies between operational, transactional, and analytical systems are increasing. For instance, without the assignment of clear roles and responsibilities for onboarding data, data scientists must wait for their data a long time, or data lakes may become "data swamps." In this example, responsibilities are needed in both

worlds. In the transactional world, the data owner must grant fast access to his/her data, while in the analytical world, a data engineer most likely onboards data to the platform according to the analytical need. Therefore, data governance today must support not only data quality control and regulatory compliance but also enable (direct or indirect) data monetization and a variety of use cases in both operational and analytical contexts (*Data Governance 2.0*).

5.2.4 Need for Guidelines Supporting Data and Analytics Governance

In line with the changing role of data, the focus of data and analytics governance needs to shift from control toward value creation, and governance practices have to adapt accordingly (see Table 5.1). In the past, frameworks or reference models have proven to be very popular among practitioners and often guide their data management and governance initiatives [2]. Their popularity in this field can be explained by the fact that "data management involves a set of interdependent functions, each with its own goals, activities, and responsibilities. [. . .] There is a lot to keep track of, which is why it helps to have a framework to understand the data management comprehensively and see relations between its component pieces" [10, p. 33].

Most of the existing data management frameworks encompass data governance as one component (see Table 5.2), while only few dedicated data governance

Table 5.1 Paradigm shifts in data governance

	Data Governance 1.0	Data Governance 2.0
Orientation	*Defensive*	*Offensive and defensive*
Data governance goals	Control: Improve data quality and ensure regulatory compliance	Value creation: Extend data use and enable value creation from data in analytical and operational use cases Control: Improve data quality and ensure regulatory compliance
Data types	Structured data, with strong focus on master data, reference data, and transaction data	Structured data (all types) Unstructured data (text, photos, videos, etc.)
Data use	Operational business processes and reporting (known purposes)	Operational business processes and reporting (known purposes) Various analytical use cases, including machine learning and artificial intelligence (previously unknown purposes)
Applications	Systems of record, such as systems for enterprise resource planning (ERP) or customer relationship management (CRM) Data warehouses, data marts Reporting, ad hoc analysis	Systems of record (ERP, CRM, etc.) and systems of engagement that enable collaboration and interactions Data warehouses and data lakes Self-service BI, advanced analytics, and AI-empowered applications

Table 5.2 Data management and data governance frameworks

Framework	Focus	Governance mechanisms		
		Structural	Procedural	Relational
DAMA-DMBOK (Data Management Body of Knowledge) [10]	General data management framework Data governance as one of the knowledge areas and center of the DAMA wheel	✓	✓	
Data Capability Assessment Model (EDM Council) [11]	General data management framework Data governance as one of the seven data management capabilities	✓	✓	
Data Governance Framework (Data Governance Institute) [12]	Data governance framework with three components: rules and rules of engagement, people and organizational bodies, and data governance processes	✓	✓	
Information Governance Implementation Model (ARMA)	Information governance, implementation	✓	✓	
Reference Model for Data and Analytics Governance (CC CDQ)	Reference model for data and analytics governance (complements Data Excellence Model as data management framework): three-step approach to define effective governance setups	✓	✓	✓

frameworks exist. One reason might be that the border between what belongs to data management and what belongs to data governance has been rather blurry in the past.

The DAMA-DMBOK [10], as the most popular data management body of knowledge, cites data governance as one of the main knowledge areas and in the center of the DAMA wheel "since governance is required for consistency within and balance between the functions" [10, p. 35]. The Data Management Capability Assessment Model (DCAM) published by the EDM Council [11] outlines data governance as one of the seven data management capabilities, with sub-capabilities such as *Data governance structure is created* or *Cross-organizational enterprise data governance is aligned*. Compared to these frameworks which consider data governance as part of data management, the Data Governance Institute's framework [12] outlines three dedicated components for data governance: *Rules and rules of engagement* define the long-term direction but also data rules and definitions as well as accountabilities and controls. *People and organizational bodies* encompass data stakeholders, data governance office, and data stewards. *Processes* comprise 12 proactive, reactive, and ongoing data governance processes, such as *establishing decision rights* or *specifying data quality requirements*. The Information Governance Implementation Model outlines eight key areas necessary

for implementing a successful Information Governance (IG) program: steering committees, authorities, supports, processes, capabilities, structures, and infrastructure. It also provides a maturity assessment.

From the comparison of data management and governance frameworks, we find that comprehensive guidelines for data governance are still scarce and much of the experience knowledge in this field is yet to be documented. The existing frameworks have a strong focus on structural and procedural governance mechanisms, although relational mechanisms are found to be essential for scaling governance. They also put more emphasis on controlling critical data assets, such as master data, than on enabling data-driven innovation and value creation in an extended network of data creators and users. Consequently, there is a need to shift the perspective on data and analytics governance from control and compliance to develop governance practices that align with the overall organization's goal to generate value from data assets.

5.3 The CC CDQ Reference Model for Data and Analytics Governance

The CC CDQ Reference Model for Data and Analytics Governance is the outcome of an extensive industry-research collaboration. It aims at supporting organizations in designing and implementing structural, procedural, and relational governance mechanisms with the goal of generating value from data assets. To provide some background, we will briefly introduce data governance research in the Competence Center for Corporate Data Quality (CC CDQ). We will then elaborate on key principles for effective governance setups and provide an overview of the CC CDQ Reference Model for Data and Analytics Governance.

5.3.1 Data Governance as Key Theme in the Competence Center Corporate Data Quality

The Competence Center for Corporate Data Quality (CC CDQ) has been founded in 2006 as industry-research collaboration to develop concepts, methods, and tools that advance data management. Today, it comprises practitioners from 20 multinational companies, many of them Fortune 500 companies (for instance, Bosch, Merck, Nestlé, Siemens, Tetra Pak, or ZF), and a team of academic researchers from the Faculty of Business and Economics at the University of Lausanne (HEC Lausanne). Since the beginning, data governance has been one of the main areas of interest in the CC CDQ. As one of the first research activities, the CC CDQ has defined the roles and boards for master data management, resulting in a first reference model for data governance [7]. These roles and their responsibilities were later further detailed and complemented by master data management processes [13]. With its focus on master

data quality, the reference model reflected the defensive orientation of data governance (*Governance 1.0*). In 2018, the CC CDQ members realized that the changing role of data in their organizations impacted on data governance. They decided to revise and extend the CC CDQ framework and data governance model with the goal to also support companies in data-driven innovation [8]. In the following, the CC CDQ reference model for data governance was extended to embrace analytics (*Governance 2.0*).

5.3.2 Design Principles for Data and Analytics Governance

The CC CDQ Reference Model for Data and Analytics Governance does not prescribe a concrete governance design, but guides companies in defining the governance design which is most suitable for their context. Independently of the specific governance design, two principles summarize the key considerations for effective data and analytics governance setups.

5.3.2.1 Principle 1: Governance Linking Strategy to Operations

Generally speaking, governance implements a strategy by means of oversight and control mechanisms and complements strategic as well as operational tasks [14]: Strategy is doing the right things, operations are doing things right, and governance is ensuring that the right things are done right. Thus, data governance takes place between strategy and operations: "Data governance should be a bridge that translates a strategic vision acknowledging the importance of data for the organization and codifying it into practices and guidelines that support operations, ensuring that products and services are delivered to customers" [15].

- At the *strategic level*, the objective and long-term direction for data and analytics are defined. This includes sponsorship, strategic direction, funding, and the coordination of data management and analytics activities at an enterprise-wide level.
- The *governance level* implements the strategy through oversight and control mechanisms. While enterprise-wide data and analytics governance is cross-functional, defines the overarching governance framework, and controls its implementation, it needs to be detailed for the different business units or departments by defining the standards and the policies of the areas of responsibility.
- The *operations level* executes the strategy through day-to-day activities, operates the data and analytics product life cycle based on the defined standards, and takes responsibility for the correctness of the data content and the use of analytics products.

5.3.2.2 Principle 2: Federated Data Governance Involving Data and Analytics, Business, and IT Experts

Data management and analytics activities in organizations require alignment and close collaboration between data and analytics experts, business and IT stakeholders. Centralizing all data and analytics activities in an enterprise would potentially increase the economies of scale but also reduces flexibility and speed to deliver value through data and analytics in different business functions. Conversely, decentralization makes business functions more flexible but requires a rather high level of maturity and skills. It may also lead to data silos and hinders data sharing and integration across functions.

As consequence, a federated approach is preferred for enterprise-wide data and analytics governance. This implies assigning data and analytics roles and responsibilities to employees and groups who work in different parts of the enterprise:

- The ownership of data and analytics products lies with the business users [16]. Consequently, *business roles* play an important role in defining business requirements for data and analytics products and ensuring that value is created from them.
- Effective data and analytics governance requires a certain level of coordination at different levels. At the enterprise level, *central teams with core data and analytics roles* are responsible for analyzing business requirements across different divisions and functions and coordinate data management and analytics delivery activities.
- *IT roles* support data management and analytics delivery by means of infrastructure and IT services. This includes the operation of analytics products and the development of analytics platforms (Fig. 5.1).

5.3.2.3 Overview of the CC CDQ Reference Model for Data and Analytics Governance

The CC CDQ Reference Model for Data and Analytics Governance builds on the principles defined in the previous section. It comprises three sequential steps that help in answering the fundamental questions related to governance design (see Fig. 5.2):

- Step 1: *What? Set the scope for data and analytics governance.*

  This step suggests taking an end-to-end perspective to identify the most relevant data and analytics products for the organization and set the governance scope in alignment with business priorities.
- Step 2: *Who? Identify decision areas/processes, roles, and responsibilities for data and analytics governance.*

  This step starts by defining the key decision areas related to data and analytics (based on the processes), before defining the required roles and boards, and

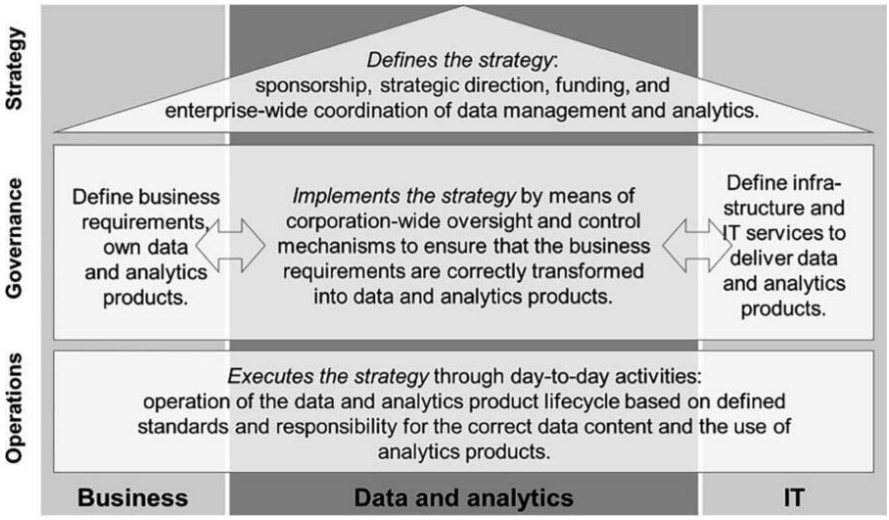

Fig. 5.1 Data and analytics governance linking strategy and operations (based on [17])

assigning the responsibilities to them. It lays the foundation for establishing the structural and procedural governance mechanisms.

- Step 3: *How? Establish the operating model and interactions for data and analytics governance.*

  In this last step, decisions are made regarding the required headcount and organizational structure and nomination of employees to roles. This step concretizes structural and procedural governance mechanisms and adds the interactions between the roles and units to explicate the required collaboration and communication (relational governance mechanisms).

5.4 Step 1: Set the Scope for Data and Analytics Governance

5.4.1 End-to-End Perspective for Defining Scope and Requirements

The first step consists in defining the scope and requirements toward data and analytics governance. Here, the CC CDQ Reference Model suggests taking an end-to-end perspective covering the most important activities related to data and analytics – starting from the source systems where data is generated to the delivery of data and analytics products, which create business value. Setting the scope of data and analytics governance therefore requires answering three questions:

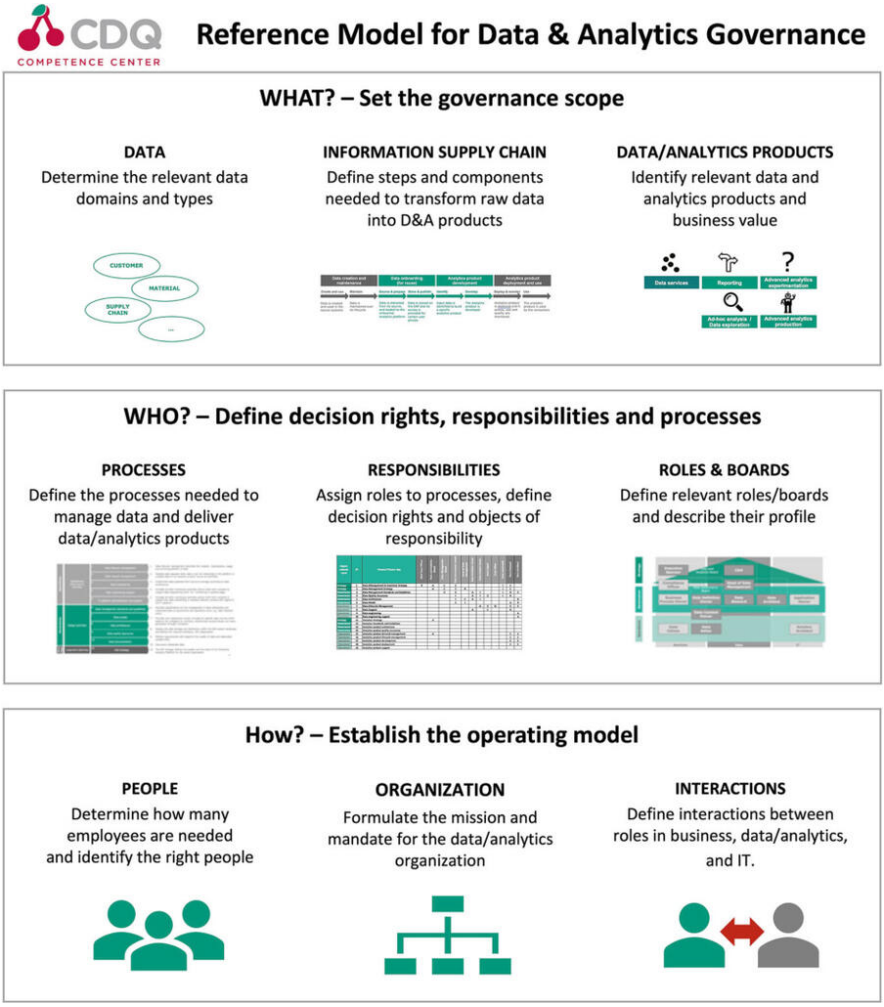

Fig. 5.2 CC CDQ Reference Model for Data and Analytics Governance

- Identify the most relevant data and analytics products for the organization (*output*).
- Identify the required datasets, domains, and data types (*input*).
- Define the phases and steps needed to transform raw data into data and analytics products, including the relevant platform and components (*transformation*).

This approach helps aligning the governance scope with priorities for data and analytics products while considering data management and analytics delivery.

5.4.2 Data and Analytics Products and Their Information Supply Chains

Each data and analytics product can be conceptualized and associated with a specific information supply chain, i.e., the successive processing steps and technical components required to produce and deliver it in a scalable way (see Fig. 5.3).

In the following, we illustrate the information supply chain for five typical data and analytics products that most companies have:

1. *Reporting:* Reports are the most common analytics product and enable an organization to make operational and strategic decisions based on structured data. It comprises periodical reports, as well as dashboards summarizing the business transactions in the form of key performance indicators and visualizations. A common way to implement corresponding pipelines are data warehouse and data mart architectures. Structured data from operational systems, i.e., master data and transactional data, are integrated in a pre-defined schema. The data mart extracts, aggregates, and processes data for the common domain of interest of the report to support the decision. Also, behavioral data such as sensor data stemming from machine equipment are used to create reports. For these scenarios, the data must oftentimes be processed in real time.
2. *Ad hoc analysis/data exploration:* To democratize data and increase its use in daily decision-making, companies provide self-service analytics tools, such as Tableau or Power BI, to their employees. With these tools, users can easily analyze and aggregate data without programming skills and visualize data in an interactive way. When it comes to data onboarding, master and transaction data is extracted from operational systems, transformed, and loaded into a data warehouse with a unified format. The data warehouse holds data from various domains. To analyze data of interest, data needs to be loaded first into a data mart before it can be accessed with self-service analytics tools.
3. *Advanced analytics experimentation:* For developing advanced analytics use cases, data scientists explore and work with data in dedicated environments, typically called data labs or sandboxes. In these environments, data scientists can use the tools they are most comfortable with and experiment with the provided data as they wish. For a specific use case, data needs either to be newly onboarded or is already accessible. Following this "pull principle" for data onboarding, it is avoided to load data into a data lake which is not used at the end. Within their dedicated environments, data scientist can explore and develop pipelines using the distributed infrastructure of the data lake in a scalable way.
4. *Advanced analytics production:* Those models that prove feasible are deployed and made accessible with the *analytics production* capability, which in turn ensures that the analytics models remain up-to-date throughout their life cycle. A business user accesses an analytics model in business applications. In technical terms, the pre-trained analytics model is accessed from an endpoint and makes a prediction based on the user input. However, the data pipelines become more

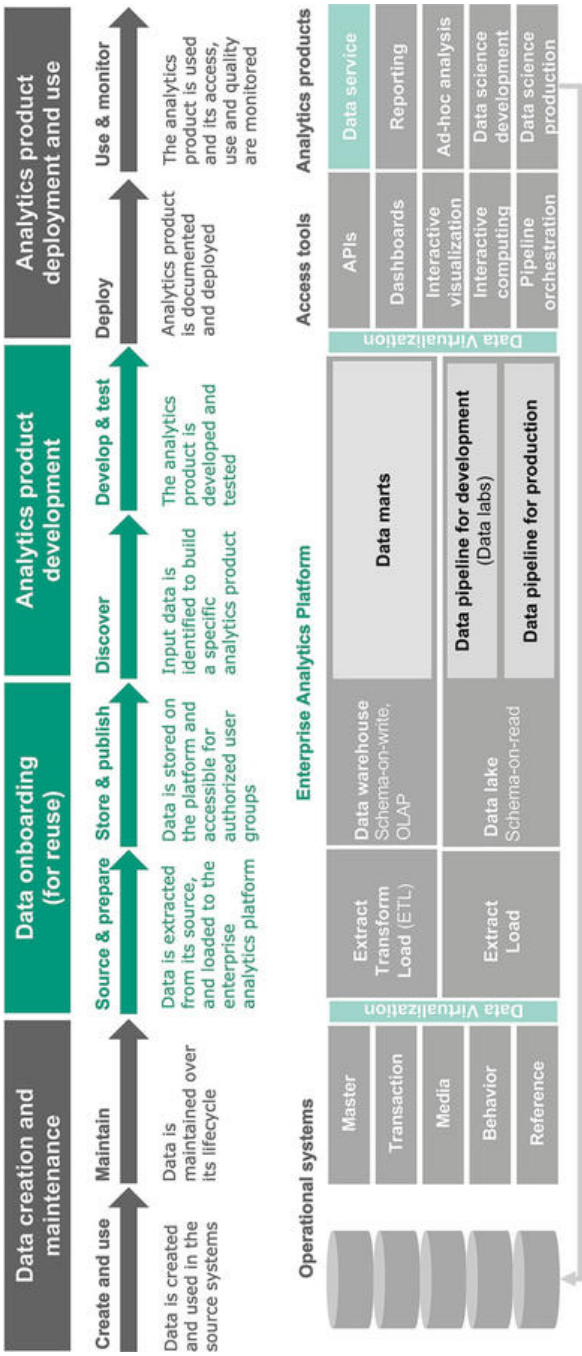

Fig. 5.3 Information supply chain

complicated when an analytics model will be automatically retrained on each interaction with the user, for instance, which is a common case when applying active learning strategies. Also, the newly available data from the user input must be validated against the dataset which was used to train the model in order to detect possible concept drifts and initiate a new training phase.

5. *Data service:* In addition to the analytics products, a data service provides data on agreed service agreements and makes them available through APIs.

5.5 Step 2: Who to Govern? – Processes, Roles, and Responsibilities

The second step in the CC CDQ Reference Model for Data and Analytics Governance defines the relevant roles and responsibilities, according to the defined scope. To answer the leading question "Who to govern?", we proceed as follows:

- Identify the *decision areas (here: processes)* on a strategic, governance, and operational level.
- Assign the *roles* and boards needed to manage data and deliver analytics products.
- Assign the *responsibilities* by mapping roles to decision areas/processes, roles (including boards), and their responsibilities.

While the process view defines procedural governance mechanisms, the roles/board view details structural mechanisms for data management and analytics delivery. The responsibilities connect the role and process view through a RACI chart which assigns responsibilities to each role and process on a granular level and also defines the relation between different roles.

5.5.1 Decision Areas (Processes)

A pragmatic approach for defining the decision areas related to data and analytics starts from outlining the high-level processes at strategic, governance, and operational level. We distinguish between two types of processes, which are interdependent and facilitate the delivery of the defined data and analytics products:

- The *data management processes* – or "left operations" – aim at making data fit for use in data and analytics products. They comprise managing data at the source level and supporting the onboarding process to the enterprise analytics platform.
- The *analytics delivery processes* – or "right operations" – aim to deliver different types of analytics products, for example, reports, ad hoc analysis, data science experiments, and production. Thus, these processes focus on managing data on the enterprise analytics platform and delivering analytics products.

In terms of governance, the most relevant decision areas are related to the (1) overarching frameworks and principles for data and analytics management, (2) the life cycle management for data and for analytics products, (3) the data and analytics architecture, and (4) applications supporting data and analytics (Table 5.3).

5.5.2 Data and Analytics Roles

An effective data and analytics governance design relies on roles and responsibilities for both the data management and analytics side.

5.5.2.1 Data Management Roles and Responsibilities

On the data management side, an effective data governance design requires data ownership to remain with the business functions [16]. It also relies on data stewards and data architects, who, for instance, set and enforce enterprise-wide standards for data documentation or facilitate data unification activities to enable experimentation with and exploration of data lakes.

The *data owner* is accountable for the data definition, creation, and maintenance (data life cycle) in specific areas of responsibility (e.g., a specific data domain such as business partner or product). He or she collects business requirements for the defined area of responsibility from business and other stakeholders, for instance, the compliance officer. The role is usually assigned to a senior executive responsible for a defined business domain (for instance, a business function or process) and who has strategic responsibility (for instance, head of sales or head of purchasing). In large organizations, the role can be split into a *data definition owner*, who is accountable for data definitions, business and quality rules, data access policies, data life cycle, and the conceptual data model, and a *data content owner* who manages the data creation and life cycle. The role of data content owner is usually assigned to executives (e.g., the head of sales of a specific country) who have operational responsibilities for the employees creating data according to the relevant data definitions.

In respect to the data in his/her domain, the data definition owner is accountable for data definitions, business and quality rules, data access policies, data life cycle, and the conceptual data model. He or she collects business requirements for the defined area of responsibility (e.g., a particular data domain like a business partner or product) from other business process owners and other stakeholders, for instance, the compliance officer.

While the data (definition) owner is accountable, the *data steward* performs the daily work and is responsible for the data definition in the specific areas of responsibility. Here, the data steward takes care of a data object (with all or a subset of attributes) in a specific data domain. This includes defining data while enforcing data quality measures and ensuring that data is *fit for use*. The *data architect* supports the

Table 5.3 Data management and analytics processes as key decision areas – strategic, governance, and operational processes

Data management	Analytics delivery
Strategic processes	
The data strategy defines targets and value proposition of data and analytics for the organization	
Governance processes	
Data management standards and guidelines prepare and communicate the specifications for data management. These include the data management framework, data definitions and life cycle, and authorization concepts	*Analytics standards and guidelines* prepare and communicate the specifications for analytics delivery. These include the analytics management framework, the definitions of analytics products and life cycle, and authorization concepts
Data performance management defines the performance monitoring system for data quality and use, compliance and other relevant aspects (i.e., metrics framework and reporting structure), and action plan for improvements	*Analytics performance management* defines the performance monitoring system for analytics product quality and use, compliance and other relevant aspects (i.e., metrics framework and reporting structure), and action plan for improvements
Data architecture ensures that data definitions are consistent and defines the structure of data, relevant rules, and metadata (independent from an application perspective). It also designs the data storage and distribution within the system landscape and defines the required interfaces	*Analytics architecture* defines the components supporting the development and deployment of analytics products and defines the required interfaces per analytics product type
Data applications define and manage the dedicated applications to manage data and support data users (e.g., data catalog, data quality tool)	*Analytics platform* defines and manages the components and enterprise analytics platform and components to develop and deploy analytics products
Operational processes	
Data life cycle management comprises the creation, maintenance, and usage of data according to the defined data architecture, standards, and guidelines	*Analytics product life cycle management* develops, deploys, and maintains analytics products according to the defined analytics architecture, standards, and guidelines
Data engineering answers data request, implements data pipelines to onboard data to analytics platforms, and contributes to developing analytics products according to data models and data architecture	*Analytics demand management* collects and discovers analytics product requests and use cases across the business, translates them, and manages the prioritization of analytics products
Data enablement includes all activities to promote data value and data awareness and to support knowledge sharing	*Analytics enablement* includes all activities to promote the use of analytics, develop skills, and support knowledge sharing
Data support processes include all other continuous activities and/or projects to support data, incl. monitoring of quality/usage	*Analytics product support* processes include all other continuous activities and/or short-term projects to support the management of APs, including monitoring of quality/usage

data steward by designing, creating, deploying, and managing conceptual and logical data models, as well as with mapping to physical data models. In the role model defined by [7], the data architect role corresponds to the technical data steward

Table 5.4 Data roles

Role	Decision right and area	Allocation
Data owner	Accountable for the data definition, creation, and maintenance (data life cycle) in specific areas of responsibility (e.g., a specific data domain). This role can be split into *data definition owner* and *data content owner*	Business (executive level)
Data steward	Responsible for the data definition in a specific area of responsibility, typically a data object (with all attributes or a subset of them) in a specific data domain	Data and analytics organization or business
Data architect	Responsible for designing, creating, deploying, and managing conceptual and logical data models as well as for the mapping to physical data models. Accountable for the implementation and maintenance of data pipelines	Data and analytics organization/IT
Data editor	Responsible for data creation and maintenance (data life cycle) according to a specific area of responsibility's data definition	Business/shared service center
Data expert	Responsible for communicating data definition and for training data editors	Business/shared service center

role and complements the business steward. To address new analytics, use cases, and new data types (for instance, data acquired from sensors or smart devices), the data definition needs to be continuously adapted and serves as a central element to ensure easy data access and use across the enterprise. The data steward is therefore in charge of handling data requests from different business functions (Table 5.4).

The *data expert* is another typical role on the operations level. This expert has no other major responsibility besides communicating the data definitions to the data editors and training them.

5.5.2.2 Analytics Roles and Responsibilities

An effective analytics governance design (see Table 5.5) requires the requestors and users of analytics products to collaborate with the data and analytics organization and IT.

On the business side, executives in business domains who sponsor and request analytics products take the *analytics product (requirement) owner*'s role. In this role, they are accountable for the specification of business requirements toward an analytics product and for realizing the business value from using it. Accordingly, they must stimulate the identification and use of analytics products in their area of responsibility in order to increase data-driven decision-making and communicate with important business stakeholders. A *business analyst*, in the analytics product requirement owner's area of responsibility, is responsible for the specification of the analytics product on the operations level. While the *analytics (product requirement) owner* specifies the business requirements, the *analytics (product life cycle) owner* is accountable for implementing these requirements in a specific analytics product, doing so by coordinating its development, deployment, and maintenance. In

Table 5.5 Analytics roles

Role name	Decision right and area	Allocation
Analytics product requirement owner	Accountable for the business value and the specification of the business requirements of an analytics product	Business (executive)
Analytics product architect	Responsible for the design of analytics products and analytics product architecture	Data and analytics organization/IT
Analytics product life cycle owner	Accountable for the implementation (development and deployment) and maintenance of an analytics product Responsible for analytics product standards and guidelines, quality assurance, and the life cycle management	Data and analytics organization
Business analyst	Responsible for the business value and specification of an analytics product's business requirements	Business
Data analyst	Responsible for the implementation (development and deployment) and maintenance of reports and ad hoc analyses	Data and analytics organization
Data scientist	Responsible for the implementation (development and deployment) and maintenance of advanced analytics models	Data and analytics organization
Data engineer	Responsible for data pipelines' implementation and maintenance	Data and analytics organization/IT
Analytics expert	Responsible for the training of analytics product users	Business/data and analytics organization

addition, this analytics product life cycle owner is responsible for defining analytics product standards and guidelines, assuring quality, and for managing the life cycle as part of her or his governance responsibility. On an operations level, he or she coordinates the *data analysts*, *data scientists*, and *data engineers* responsible for analytics products' development and deployment. In order to do so, she or he involves the business stakeholders to ensure that the business requirements are met. The analytics product life cycle owner is typically a person with project management experience with technical know-how of analytics product development. The analytics product architect's role is meant to ensure applications' reusability and scalability across the enterprise. This architect is responsible for analytics products and analytics product architecture's design, which requires close collaboration with the IT organization. Consequently, this role is allocated to the bordering area of analytics/IT.

Two data governance roles are of particular importance for the analytics organization. The *data architect* is accountable for data pipelines' implementation and maintenance by providing the data models that data engineers use. The *data steward*, a key role for data governance, is responsible for managing analytics projects' data requests and for supporting the data onboarding process. This support is of particular importance to increase the analytics practitioners' efficiency and reduce the time spent on finding and preparing data.

5.5.2.3 Organization-Wide Coordination of Data and Analytics

The *role of the chief data officer (CDO)* – also called head of data and analytics or chief data and analytics officer (CDAO) – is becoming of major importance in enterprises. A CDO is the head of the central data and analytics organization, responsible for the overall data management and analytics strategy, and accountable for its implementation. This range of activities requires continuous exchanges with the data and analytics organization's executive sponsor on the business side, as well as with the chief information officer (CIO) on the IT side. In the role model suggested by [7], a CDO fulfills the chief data steward role and extends his or her accountability to the analytics organization.

A *central data and analytics organization* ensures that requests for new analytics products (e.g., data science use case) are prioritized and specified within an enterprise-wide demand management process. Although all companies still distinguish between the delivery of BI (e.g., reporting) and advanced analytics products (e.g., predictive modelling), they seek an integrated, unified view on analytics products' demand and delivery in the long term, in order to bundle resources and facilitate their analytics capabilities. Business roles' involvement guarantees that the business requirements are met and the domain knowledge is transferred to analytics products.

In addition, companies increasingly establish a dedicated data and analytics board comprised of C-level executives to align the stakeholders on the enterprise level. This board is accountable for defining the data and analytics strategy, controlling its implementation (including compliance requirements), and setting priorities.

5.5.3 Assigning Roles to Responsibilities

Once the decision areas (or high-level processes) and roles have been defined, it is possible to assign responsibilities on more granular level. A RACI matrix can be used to define for each of the processes the person or board:

- Responsible for the process or task
- Accountable for the process or task
- Consulted, who needs to be involved in the process or tasks
- Informed, who needs to be informed about the results

5.6 Step 3: How to Govern? – Deriving the Operating Model

5.6.1 Mapping Roles, Responsibilities, and Processes to the Organizational Context

The third step aims at answering the question "How to govern?" and defines the operating model. Thus, the tasks are to map roles, responsibilities, and processes to the specific organizational context:

- Define the headcount and structure of the data and analytics organization, and assign roles and responsibilities.
- Identify the relevant (cross-)functional and divisional data and analytics domains, and assign roles and responsibilities.
- Define interactions between the different groups and roles in data and analytics, business, and IT.

The derivation of the operating model starts with structuring and organizing the way of working in the central data organization. Assigning the roles and responsibilities in an organization depends on many factors – most importantly, the maturity of the company and the mandate for data management and analytics. In practice, many variants can be found. Once the scope and way of working in the central data management organization have been clarified, team sizes must be determined and the responsibilities assigned to employees in the organization.

5.6.1.1 Typical Configurations

While this organizational design is contingent on various factors and hence depends on the unique situation of a company, we identified typical data governance design patterns through an in-depth analysis of several case studies. These patterns can be associated with different stages of maturity:

1. **Pattern 1 (improve master data quality):** Companies belonging to the first pattern have a narrow data governance scope, focusing on improving data quality for master data in a few data domains, typically, product and finance, but do not prioritize analytics products beyond reporting. Companies use this initial structuring along the key business objects to define distinct areas of responsibilities and extend them to additional domains in later stages. However, in pattern 1, a central data team is granted main operational responsibilities for collecting business requirements, setting up data quality measures, monitoring data quality, and supporting projects that involve data quality issues. Hence, responsibilities are mainly centralized, although the data content is created in business units.
2. **Pattern 2 (enable enterprise-wide data management):** Companies belonging to this data governance pattern follow a broader governance scope: they have

defined their data strategy and set their focus on the most relevant data domains and data types for operational and analytical use cases. While data quality remains a key central responsibility, the central data team assumes broader responsibilities related to executing the data strategy. To improve data quality and promote data access and use, the responsibilities are gradually decentralized to business roles, who collect business requirements in structured ways and maintain data according to domain-specific standards and guidelines. In this pattern, relational mechanisms are more intensively used than in the first design pattern. For instance, roles and responsibilities are communicated and collaboration and alignment happen in regular meetings and steering committee with business professionals.

3. **Pattern 3 (coordinate data network to enable data monetization):** Companies belonging to this pattern recognize data as strategic asset and major driver of their digital transformation. They usually bring an extensive experience in data management and aim at finding new ways for monetizing data. As data and analytics are major value drivers for the company, they promote an integrated view of data and analytics through which they foster synergies and manage data quality and usage in a seamless way. The central data team mostly undertakes strategic responsibility and is closely aligned with C-level executives while coordinating a network of decentral data and analytics teams in different teams. This pattern is closely connected to establishing the role of the chief data officer, which fosters the alignment and steers data monetization activities at enterprise-wide level.

5.7 Summary

The CC CDQ Reference Model supports practitioners in the governance design process by answering three fundamental questions: (1) "What to govern?" (scope), (2) "Who to govern?" (roles and responsibilities), and (3) "How to govern?" (operating model). As an important contribution, this model bridges the distinct perspectives and independent responsibilities for data management and analytics delivery. We emphasize this end-to-end perspective as value can only be created when data management and analytics delivery are governed in close conjunction to enable value creation and innovation from data.

Acknowledgments This work was supported by the Competence Center Corporate Data Quality (CC CDQ, www.cc-cdq.ch). The authors would like to thank all CC CDQ partner companies for their financial support and their active contributions to the development of the Reference Model for Data and Analytics Governance.

References

1. Grover, V., Chiang, R.H.L., Liang, T.-P., Zhang, D.: Creating strategic business value from big data analytics: a research framework. J. Manag. Inf. Syst. **35**(2), 388–423 (2018)
2. Legner, C., Pentek, T., Otto, B.: Accumulating design knowledge with reference models: insights from 12 years' research into data management. J. Assoc. Inf. Syst. **21**(3), 735 (2021)
3. Vial, G.: Data governance and digital innovation: a translational account of practitioner issues for IS research. Inf. Organ. **33**(1), 100450 (2023)
4. Petzold, B., Roggendorf, M., Rowshankish, K., Sporleder, C.: Designing Data Governance that Delivers Value, pp. 1–8. McKinsey Technology (26 June 2020)
5. Cambridge Dictionary. Governance [Online]. https://dictionary.cambridge.org/dictionary/english/governance. Accessed 31 January 2023
6. Khatri, V., Brown, C.V.: Designing data governance. Commun. ACM. **53**(1), 148–152 (2010)
7. Weber, K., Otto, B., Österle, H.: One size does not fit all - a contingency approach to data governance. J. Data Inf. Qual. **1**(1), 1–27 (2009)
8. Tallon, P., Ramirez, R.V., Short, J.E.: The information artifact in IT governance: toward a theory of information governance. J. Manag. Inf. Syst. **30**(3), 141–178 (2013)
9. Abraham, R., Schneider, J., vom Brocke, J.: Data governance: a conceptual framework, structured review, and research agenda. Int. J. Inf. Manag. **49**, 424–438 (2019)
10. DAMA: DAMA-DMBOK: Data Management Body of Knowledge. Technics Publications (2017)
11. EDM Council. DCAM (Data Management Capability Assessment Model), Version 2.2 (2020)
12. Data Governance Institute. Data Governance Framework [Online]. https://datagovernance.com/the-dgi-data-governance-framework/. Accessed 31 January 2023
13. Reichert, A., Otto, B., Österle, H.: A reference process model for master data management. In: Proceedings of the 11th International Conference on Wirtschaftsinformatik (WI2013), Leipzig (2013)
14. Kim, A., Tiwana, S.K.: Discriminating IT governance. Inf. Syst. Res. **26**(4), 656–674 (2015)
15. Vial, G.: Data governance in the 21st-century organization. MIT Sloan Manag. Rev. (2020)
16. Fadler, M., Legner, C.: Data ownership revisited: clarifying data accountabilities in times of big data and analytics. J. Bus. Anal. **5**(1), 123–139 (2022)
17. Fadler, M., Legner, C.: Toward big data and analytics governance: redefining structural governance mechanisms. In: Proceedings of the 54th Hawaii International Conference on System Sciences, 2021. HICSS (2021)

Chapter 6
Data Governance Tools

Kash Mehdi

6.1 Introduction

In the entire history of Data Management, more Data Governance tools[1] are available today than ever; understanding them can be overwhelming. The current state of the Data Governance space continues to witness a massive rise in technological innovation as more organizations look for ways to retrieve value from their data assets. [2]

Technology plays a crucial role in augmenting labor-intensive human tasks such as connecting raw data with business context; running scan and discovery engines to break data silos; establishing and tracing enterprise-wide data ownership, stakeholder accountability, and their decision rights; and tracing data from source systems to target consumption points, mobilize stakeholders to collaborate on data issues (e.g., poor Data Quality, inaccurate KPIs, Reports), and maintain appropriate security and privacy compliance levels. The bigger picture around Data Governance technologies is to enable organizations to transform the entire company culture to lead with data.

In this chapter, we will explore the following four topics:

1. The business need for Data Governance and its importance
2. Southwest Airlines case study and the role of technology on business outcomes
3. Key functionalities needed in the Data Governance tools
4. Four must-have technology focus areas to kick-start Data Governance

[1] To name a few, DataGalaxy (www.datagalaxy.com), Collibra (www.collibra.com), Alation (www.alation.com), Informatica www.informatica.com), data.world (https://data.world)

[2] See https://www.imarcgroup.com/data-governance-market.

K. Mehdi (✉)
DataGalaxy, Lyon, France
e-mail: kash@datagalaxy.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_6

6.2 The Business Need for Data Governance and Its Importance

Without discriminating the industry type or geography, the Data Governance space continues to grow exponentially as more organizations across the globe build hierarchical structures around data ecosystems and have increased spending on data-related technologies. The Chief Data Officer role has strategically evolved to handle such data ecosystems. It continues to gain mindshare with board members and C-suite executives who realize the critical need to manage data effectively as an asset for innovation and to gain competitive business advantage. Many organizations assign the Chief Data Officer role with the expectation to help stakeholders understand how the entire business runs on data and, more importantly, monetize data to deliver business outcomes. Depending on the type of industry or its focus, the Chief Data Officers spearhead any of the following business outcomes either internal to the company or externally facing:

6.2.1 Common Business Outcomes Led by Chief Data Officers

- **Insights and Analytics Operational Efficiency:** An internally facing business outcome where Chief Data Officers focus on achieving operational efficiency by creating a self-service capability for the Analytics community. For example, Data Analysts and Data Scientists need access to trusted data when performing business activities such as creating business intelligence reports, performing predictive analytics, building analytical models, understanding data flowing from source systems to target consumption points, accessing usage guidelines for appropriate use in projects or marketing campaigns, and much more.
- **Regulatory Compliance:** An externally facing business outcome applicable in almost all industry sectors is needed to maintain trust and compliance with industry operating standards. The ever-changing regulations warrant organizations to establish strong data management best practices to ensure data transparency, traceability, security, and privacy compliance. A few trending examples include, by industry, *Environment, Social, and Governance (ESG; applicable in all industries), General Data Protection Regulation (GDPR; applicable in all industries), California Consumer Protection Act (CCPA; applicable in all industries in the state of California, United States), Basel Committee on Banking Supervision's standard number 239 (BCBS239; applicable in the financial services industry), International Financial Reporting Standard (IFRS17; applicable in the insurance industry), Medical Device Regulation (MDR; applicable in the medical device industry), and much more.*
- **Organizational Data Literacy:** An internally facing business outcome. Today, organizations are experiencing the volume of data doubling more frequently than

ever and consuming more data daily, generating unprecedented human knowledge, which, when utilized, could provide a competitive business advantage. Unfortunately, despite the availability of traditional data governance technologies, organizations are challenged with user data literacy, which impacts their ability to move in concert to deliver innovation or gain a competitive business advantage. Also, much of the data is stored in black box data silos lacking appropriate business context, which impacts data consumer trust for usage in business activities. There are far greater expectations of the data governance tools to break such data silos, build user adoption, and uncover data patterns to enable organizations to enhance customer experience and product and service offerings.

- **Digital Transformation:** Both an internally and externally facing business outcome. Cloud migrations are happening at an accelerated phase than ever before. Since the beginning of the COVID-19 pandemic and shifting of work environments, many organizations have shifted their operating model to the cloud to meet customer expectations and scale their technology ecosystem, with more joining each day. Harvard Business Review states: "Digital Transformation is about improved visibility of resources and better resource management, enhanced flexibility and organization agility, lower costs, smoother supply chain management, better customer experience, improved productivity, faster product development, and superior human resource planning." The journey to the cloud requires data-related technology to help lift and shift data assets from legacy on-premise data ecosystems to a more modern and scalable technology infrastructure. It also warrants data trust, which can be curated with appropriate business definitions, ownership, source-to-target traceability, quality, and privacy standards during its life cycle. More specifically, the role of Data Governance tools can be viewed as a data filtering mechanism between the on-premise and the cloud ecosystems.

Chief Data Officers are not limited to the above list of business outcomes. They continue to cross paths with changing market landscapes, macroeconomic conditions, adverse events, growing stakeholder demands, and regulations, to name a few. The above list represents industry-agnostic macro themes applicable to any organization irrespective of its shape or form. More business outcomes are expected to evolve based on each organization's internal and external focus areas.

6.3 Case Study: Southwest Airlines and the Role of Technology on Business Outcomes

Southwest Airlines is the world's largest low-cost carrier and one of the major airlines in the United States. Many lessons can be learned from the Southwest Airlines case study.

During the 2022 holiday season, we saw a never-like-before record bitter cold storm in the United States, impacting many in the transportation industry. Southwest

Airlines came to the spotlight for its record cancellations stranding several passengers scrambling to connect with the airline staff seeking help. In speaking with the CNN news channel, Pete Buttigieg, the US Secretary of Transportation, reported, "The airline was unable to locate its staff members, let alone their passenger's baggage."

According to FlightAware[3] data on airline cancellations, Southwest Airlines[4] recorded 2500+ flight cancellations, the highest among its peers. This tragedy highlighted the impact of operational efficiency on Southwest Airlines' business outcomes, severely impacting its brand reputation, exposing technology vulnerabilities, and impacting the airline's ability to scale and resume normal business operations.

Running a successful airline business requires a concerted effort from all parts of the organization. The role of Chief Data Officers is far more critical when responding to such adverse events, especially winter storms that can halt an airline's operations in their tracks. Many in the transportation industry face numerous data challenges, as outlined in Fig. 6.1:

6.3.1 Data Challenges in the Transportation Industry

The challenges include the following:

- Ad hoc business processes to collect and protect customer data and coordinate information with local authorities to comply with safety standards.
- Internal alignment seeks to appropriately store customer data, fleet management, and reporting on operational metrics to measure system performance, trends, and remediation plans in case of technology failures (for Southwest Airlines, effectively managing flight cancellations and rebooking so passengers can reach their destination without much delay).
- Lack of standards around internal and external data sharing practices to avoid operational failures
- Lack of shared understanding of data to enable predictive analytics

Use of poor data quality and privacy standards in customer-focused initiatives

The availability of clean and trusted data can unlock many benefits for the transportation industry. It is not just Southwest Airlines that desperately needs operational efficiency; many in the industry are plagued with such data challenges.

In favor of Southwest Airlines, they are one of the leading carriers in the industry. They have made strides in building a business around communities and not hubs, providing affordable airline tickets for passengers from all walks of life. While affordability is crucial to find a mindshare with customers and capture market share, there are other variables to retain a top stop in the industry, especially amid

[3] https://www.linkedin.com/company/flightaware/

[4] https://www.linkedin.com/company/southwest-airlines/

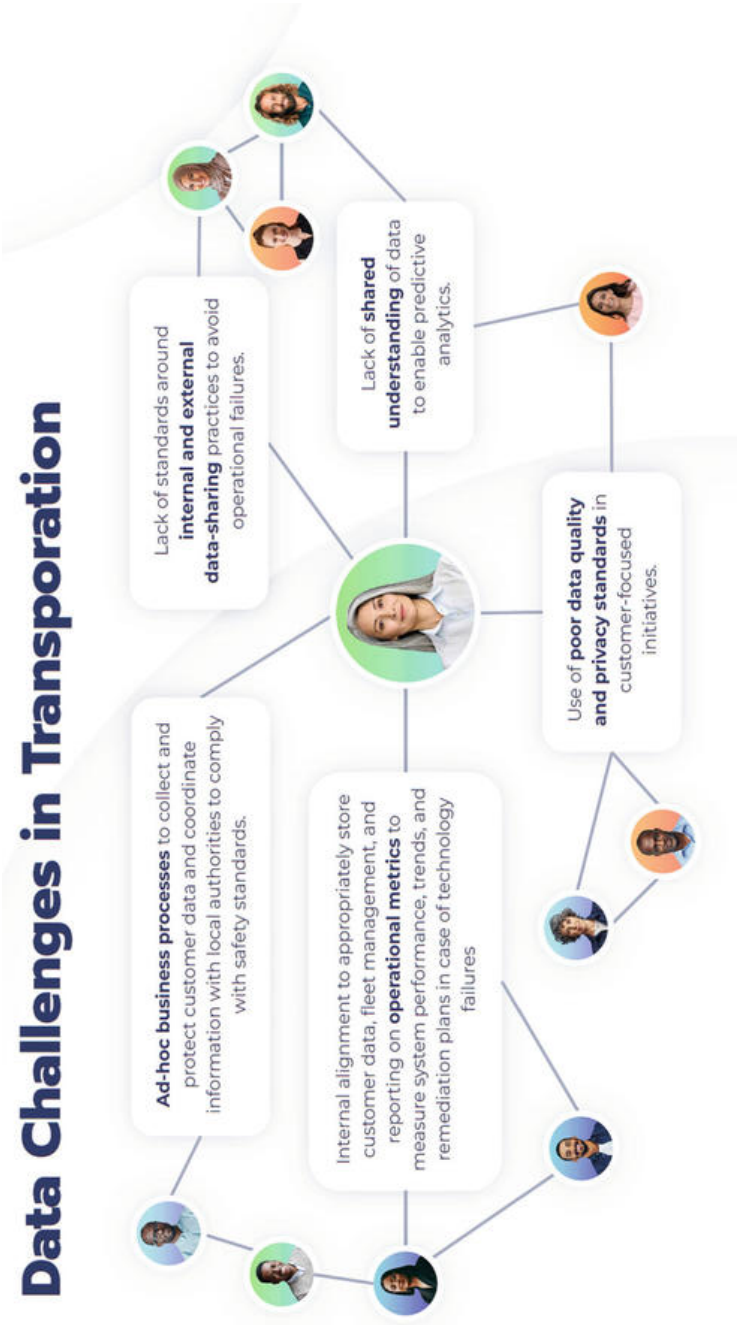

Fig. 6.1 Data challenges in transportation industry

growing competition taking swift action and potentially impacting monetary gains. Many of Southwest Airlines' competitors, including American Airlines, [5] Delta Air Lines, [6] and United Airlines, [7] introduced fare caps in some cities where the airline operated. [8]

Customer experience and acquisition remain the top driver for most customer-facing businesses. Organizations need access to reliable data to make data-driven business decisions, which is precisely the value the Chief Data Officer role brings to the table. There are many ways in which companies can unlock competitive advantage and deliver fit-for-purpose customer experience by effectively governing data. Most data governance initiatives' ultimate goal is to support business outcomes. However, many such data initiatives do not survive due to a lack of adoption of business, user, and technology. The role of technology becomes more critical in driving data user productivity when managing business activities to predict growing customer demands and action to create meaningful solutions for the business.

6.4 Key Functionalities Needed in the Data Governance Tools

Data Governance tools are critical in bringing the business and technology teams together like never before. They must offer rich user experiences to enable Chief Data Officers to help stakeholders understand how the entire business runs on data and build back a better future of scale and organizational readiness to respond to any adverse event, be it a pandemic like COVID-19, macroeconomic conditions, or even climate change-related circumstances.

As a value-add to any industry type, Data Governance tools must offer valuable capabilities empowering the Chief Data Officer role. They must combine industry best practices and practical customer experiences to enable organizations in three major categories:

- **Share**: Ability to share trusted data to enable data consumers at all levels of the organization when performing business activities (e.g., creating the report, data sharing agreements, data contracts, data products, predictive analytics, and model creation).
- **Manage:** Provide a data workspace to enrich data with trust attributes and increase user productivity by reducing the time to find data to move in concert

[5] https://www.linkedin.com/company/american-airlines/

[6] https://www.linkedin.com/company/delta-air-lines/

[7] https://www.linkedin.com/company/united-airlines/

[8] https://www.linkedin.com/pulse/what-chief-data-officers-can-learn-from-southwest-airlines-kash-mehdi/?trackingId=tNEddvuwT%2BeIUsekuo1flg%3D%3D

to convert data into actionable insights. Hence, with the might of the entire workforce, deliver innovation and gain competitive business advantage.
- **Scan:** Break black box data silos by operationalizing intelligent scanning and discovery capabilities to unlock data patterns and insights to enhance customer experience, business intelligence, and more.

6.4.1 *Twelve Technology Features Chief Data Officers Can Use to Become Data-Driven*

Under each major category (Scan, Manage, and Share), 12 technology features can be outlined (see Fig. 6.2). While most traditional players cover some of the functionalities, they are often challenged with user experience and fail to drive user adoption.

6.4.2 *Data Governance Technology Challenges*

While not limited to the above 12 valuable capabilities under the Share, Manage, and Scan categories, traditional Data Governance tools encounter challenges driving data culture and change management. While they cover some or most of the functionalities listed above, the most significant gap is felt when they need to connect with the end users of the technology. Such a problem warrants Data Governance technology vendors lead with a user-experience-first mindset when designing new features and functionalities.

Traditional Data Governance technologies lacking user adoption have severely impacted the Chief Data Officer role. According to an MIT Sloan Management study, [9] it is reported that Chief Data Officers stay in their role for only 2 to 3 years, which, compared to a CEO, is 7 years and 4 years for a CIO.

One key question comes to mind as we navigate the Data Governance landscape: "As a Chief Data Officer, have you realized the full potential of your data governance initiative, and what business outcomes would you say you have achieved?" Only a few Chief Data Officers can say that, and many still face user adoption and change management challenges.

[9] Source: https://mitsloan.mit.edu/ideas-made-to-matter/chief-data-officers-dont-stay-their-roles-long-heres-why

Fig. 6.2 12 Ways Chief Data Officers can become data-driven

6.5 Four Must-Have Technology Focus Areas to Kick-start Data Governance

Data Governance is an exciting journey, at least what it feels like all day, every day when engaging with customers across various industries and geographies. While no one size fits all, the essential elements to kick-start a data governance program are necessarily the same.

Before getting into the four must-have technology focus areas for kick-starting a governance program, let us take a step back and zoom in on the challenges around the data itself, to name a few:

1. Building and operationalizing a holistic data and analytics strategy
2. Delivering clean and trusted data with appropriate security and privacy compliance controls
3. Digital Transformation to support the lift and shift of data from legacy ecosystems to the cloud
4. Maximizing the impact of Insights and Analytics and Master Data Management programs
5. A centralized data inventory of logical data assets spread across multiple systems, applications, and data silos
6. Managing risk exposure on existing data and dealing with growing regulatory compliance needs (e.g., ESG, GDPR, CCPA, BCBS239, IFRS17, MDR)
7. Leveraging Artificial Intelligence and Machine Learning to drive insights from existing data and drive automation
8. Capturing the data flow from the cradle to the grave (what it means, where it comes from, ownership, life cycle, and more)

The list continues to grow in time and space as the data universe expands. Inevitably, data, without a doubt, has become a strategic asset for companies going through Digital Transformation, which is also a massive business drive or motivation for companies to undertake Data Governance initiatives and spend on relevant technologies.

A common question that gets asked by organizations is: "What must-have technology focus areas do I need to kick-start a data governance program?"

Having spent the last decade in the Data Governance space and seeing it mature across various industries, there are four must-have technology focus areas to operationalize it.

6.5.1 Flexible Operating Model

Data Governance Tools must offer a flexible operating model to help organizations align their operating hierarchical structure.

Fig. 6.3 Types of operating data governance models

The operating model is the base for any Data Governance program. It relates to various activities for defining enterprise roles and responsibilities across the line of business. The idea is to establish an enterprise governance structure. Depending on the type of organization, Data Governance structures could take different shapes or forms, covering the ones shown in Fig. 6.3.

As such, Data Governance tools must provide flexible functionalities to cater to different operating model needs. Many of today's traditional data governance players either offer too much flexibility or have a rigid operating model, severely impacting the Chief Data Officer's ability to stay the course with project timelines and gauge the level of effort around initial product setup covering installation, stakeholder alignment on the operating model, cultural considerations in technology, user productivity focus, and much more.

6.5.1.1 Insurance Customer Story

A primary insurance provider in New York City was working on its first Data Governance project. They started the journey by interviewing leaders from each business line, such as Finance, Insurance, Sales, and Marketing. As part of the process, they identified two key representatives from each business line, one in the Business and one in the Technology, named Business and Technical stewards. The business side was marked as the owners of data and information technology as the owners of the infrastructure supporting data. Similarly, various Data Stewards were identified for other business lines to form nested Data Governance layers, which then rolled up to the leaders of Business and IT.

A draft operating model was created to represent an enterprise data governance structure. The Corporate Data Governance Council committee was formed with the Chief Data Officer at its helm.

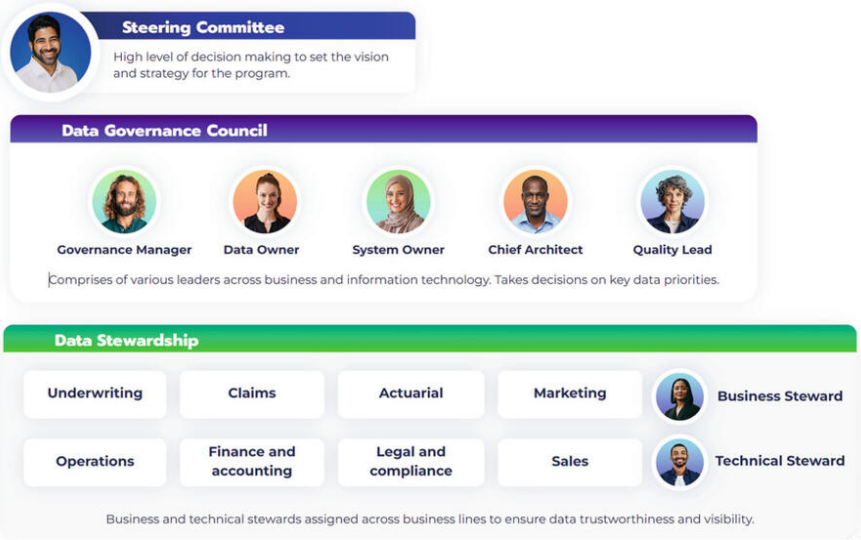

Fig. 6.4 Sample Enterprise Data Governance Council

Note: Defining the realm of ownership across your organization is essential. Determining authority will help socialize the data governance program and establish an intelligence structure to tackle data programs as a single unit of force.

Business and Information Technology (IT) members from different groups align to a reporting structure, often called the Data Governance Council or the Data Stewardship Committee. They are engaged in data discussions and are responsible for most everyday data-related decisions and dissemination of information across the organization. Also, they are responsible for ensuring formalized data ownership and determining the right Data Governance tools to support Business and Technical Steward productivity goals.

The diagram introduced in Fig. 6.4 depicts a simplified example of an Enterprise Data Governance Council:

A flexible operating model is a must-have technology focus area for organizations getting started with data governance or even the ones that have gone through a journey that requires change based on post-implementation learnings. Also, various studies around the data governance approach suggest that no one size fits all organizations. The Data Governance tools must provide a degree of personalization and customization backed by appropriate best practices, training, and education.

6.5.2 *Identification of Data Domains*

Data Governance tools must help organizations break black box data silos lacking appropriate business context and promote user trust during data usage.

Once the operating model is finalized, the next step is identifying relevant data domains for applying Data Governance.

For most organizations, data is categorized either in terms of data domains, business lines, or projects. Data domains could be organized differently depending on the business line's needs. Customer, Vendor, and Product are commonly used data domain examples. One of the biggest challenges ahead of any organization when starting data governance is identifying the most critical data domains without boiling the ocean. Also, it is equally important to link business outcomes and data consumer needs and identify a data domain.

The role of Data Governance tools is far more critical to provide the necessary connectivity to retrieve data from existing technology infrastructure. Data could be lost in the universe of systems, applications, unstructured file formats, ETL transformation logic, Data Archives, SharePoint, a random file on someone's desktop, and much more. In addition, the Data Governance tools must offer data stewards a business-friendly user experience to organize data effectively to match business needs.

For instance, let us consider Customer, Vendor, and Product as three data domains to view various artifacts listed in Fig. 6.5:

6.5.2.1 Financial Services Customer Story

Typically, identifying data domains starts with a business need or a problem. Using one of the Financial Services client's experiences, here is an example outlining a list of operational goals:

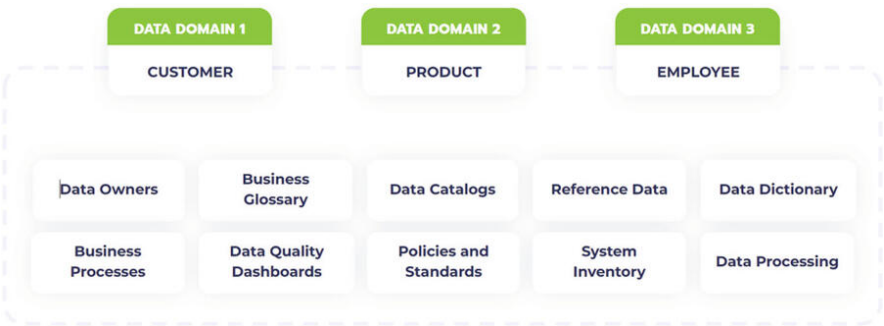

Fig. 6.5 Data artifacts within a given data domain

1. Increase customer experience.
2. Establish control over validating customer needs.
3. Manage customer usage of the product and services.
4. Increase upsells on storage billing cycles.

Note: Data governance is about people, processes, and technology. It can be enabled by identifying a data governance structure, assigning roles and responsibilities, and managing critical information assets through a technology platform for governance.

The Financial Services company used the aforementioned operational goals. It tied them to the business problems, i.e., to control visibility and understanding of customer data, which initially was spread across multiple systems and applications with no defined ownership and business context.

One of the hardest things for Chief Data Officers is to link business outcomes with data challenges. The Data Governance tools must provide capabilities to enable organizations to bridge the gap, wherein they can identify and assign stakeholder ownership, capture business processes generating data and datasets for each data domain (Customer, Vendor, Product), and establish quality and privacy controls throughout the data life cycle. The Data Governance tools are expected to facilitate functionalities to understand where the data comes from, its ownership, and who should be involved in processes when changes are made. It is also critical for the tools to help capture end-to-end data lineage.

Within the Financial Services company, they established a simple rule around its Business Intelligence Reporting metadata: "If you cannot tell me where you got the data from, your report is not certified." The critical exercise in the above example was to link business metadata with technical metadata, including interconnected systems and applications. The model becomes scalable if Data Governance tools can help figure out the system to trace one report.

Figure 6.6 below shows a sample framework around the Report Certification use case:

6.5.3 Identification of Critical Data Elements (CDEs) Within Data Domains

Data Governance tools must help bridge the business and technical knowledge gap. Following the steps from defining an operating model and identifying data domains for governance, the next step is to zoom in on each data domain to mark critical data elements, often called CDEs, wherein business and technical metadata are linked (often a labor-intensive exercise). With the availability of modern data governance technologies, it becomes manageable to identify and enrich each CDE with trust attributes (e.g., security classification, ownership, and data definitions, to name a few).

Fig. 6.6 Report watermarking

In today's reality, after identifying the data domains, most organizations find themselves at the pinnacle from which they see data domains touching tens, hundreds, and thousands of systems and applications containing critical reports, CDEs, business processes, and much more. Most traditional data governance players offer connectivity. However, they fail to consider the user experience needed in the aftermath of scanning and discovery, i.e., to guide organizations to not boil the ocean by simultaneously focusing on all the data assets. Instead, it is to the Data Governance tool's advantage to enable organizations to identify CDEs most critical to the business.

6.5.3.1 Federal Government Agency in Washington, D.C., Story

A Federal Government Agency in Washington, D.C., started a Data Governance initiative to attain commonality across the enterprise. A centralized technology platform was desperately needed to manage and control changes and provide visibility into critical data assets. A Data Governance tool was procured to serve as a platform to create a vibrant ecosystem fostering collaboration around the data life cycle and its management and retaining audit logs for past and future analysis.

6.5.3.2 Technology Company Story

A technology company out of California, United States, needed to validate financial reports and related source systems. They started by identifying ten key reports and documented information about the corresponding system of origin. Later, the initiative was scaled and called "The Report Certification" process, which applied to all reports showing certification and related source system information. A Report cannot be certified if the owners cannot prove its data lineage to the system of origin where the data gets generated. This particular exercise around capturing report lineage enabled the organization to automate data cataloging, wherein they scanned underlying systems and applications.

The Data Governance tools will be advantageous if they consider functionalities enabling Chief Data Officers with tangible quick wins, such as the example of "Report Certification." Having such technology considerations will advance the field of Data Governance, which is currently plagued by user adoption challenges. It will also allow Chief Data Officers to evangelize their work backed by concrete data examples.

6.5.4 Enable Control Measurements

Data Governance tools must help organizations apply quality and privacy control measurements and enable them to track the adoption of Data Governance over time.

So far, we have learned three must-have technology focus areas for data governance: Operating Models, Data Domains, and Critical Data Elements. The last focus is establishing and maintaining control to sustain the Data Governance program. Having helped numerous organizations establish data governance across various industries worldwide, including Financial Services, Healthcare, Insurance, Government, Retail, Manufacturing, Higher Education, and much more, my understanding is that data governance is not a one-time project. Amid changing market conditions, data governance is considered an ongoing program to help organizations understand how their entire business runs on data and enable them to create opportunities for the business. Data Governance also helps prepare an organization to meet new business outcomes.

For Data Governance tools, when it comes to defining control measurements, they must offer the following key capabilities:

1. Automated workflow capabilities to enable Business and IT collaboration around data change approvals, escalation, review feedback, voting, issue management, and much more
2. Application of workflow processes to engage at various nested layers of data governance involving stakeholders, relevant data domains, and critical data elements
3. Robust dashboard and reporting to track the progress of Data Governance (e.g., pending ownership assignment, CDEs without business context, list of data inventory captured, tagging of policies and standards along with usage guidelines)
4. Must include social media-like features to encourage stakeholders to provide feedback through automated workflow processes and audit trail views showing historical changes (before and after)
5. Must provide capabilities to create a library of policies and standards and the ability to tag the same to business and technical metadata for risk reporting
6. Must provide capabilities to create a library of data quality rules and standards and a framework to report on quality trends to review poor-quality issues and remediation

6.5.4.1 Technology Company Out of California Story

A technology company out of California started with Data Governance in early 2010. They began by defining ownership, stakeholder roles, and responsibilities, defining business data definitions, and applying workflow processes to facilitate collaboration during change management involving business and technology data stewards.

Ultimately, they established a robust data governance organization supporting an ongoing program for managing all business data definitions and execution of control measurements such as onboarding business data definition, stakeholder workflow

approvals, reviews, data steward collaboration, capturing stakeholder feedback, and applying quality and privacy standards.

Considering the above Technology Company example, the Data Governance tools must enable organizations to maintain control and track the adoption of the program over time.

6.6 Conclusion

There is more to the above four must-have technology focus areas to kick-start Data Governance. Depending on the type of industry, there could be different approaches. The above focus areas stand valid for measuring the effectiveness of various Data Governance tools, which, when done right, can enable organizations to achieve better data quality, security, and privacy compliance and maximize business intelligence and other data initiatives.

Shifting from traditional tools to more modern and flexible Data Governance technologies offers possibilities for achieving business outcomes, which will help organizations prepare for growing internal and external business needs (Insights and Analytics, Regulatory Compliance, Data Literacy, and Digital Transformation, to cite some examples).

Using the highway tolls analogy, one can consider Data Governance tools as a tollgate for data needs. For example, before undertaking any data initiative or project, a Data Governance tool can offer rich insights by enabling rich searchability, business context, ownership, lineage traceability, and quality and privacy controls.

Chapter 7
Maturity Models for Data Governance

Ismael Caballero, Fernando Gualo, Moisés Rodríguez, and Mario Piattini

7.1 Introduction

In recent years, the importance of data has been emphasized, and expressions such as "data is the new currency," "data is the new oil," and "data is the hidden mine" have become popular. In fact, digital transformation is affecting all sectors, from agriculture to industry, tourism, and healthcare, to name a few. The case is that data has become the most potent enabler of any organization. This increment of importance is because, as Aiken points out in [1], data enables organizations to achieve different strategies: data-centricity, industry convergence, hybrid services, and customer-centricity.

All countries are driving the data economy; for example, the European Data Strategy [2] foresees a 530% increase in the overall volume of data generated and moved within the European Union. For this reason, there is a demand for the creation of adequate data governance mechanisms in organizations so that they can be competitive players in the data market and improve the well-being of citizens. Meeting this demand is fundamental to ensure that data is fit for purpose and can be trusted to do any of the necessary tasks of the organization [3].

The expected benefits of data governance are (1) optimization of the organizational value of data through alignment with organizational strategy; (2) optimization of risks related to the acquisition, use, and exploitation of data, ensuring compliance with regulatory standards; and (3) optimization of the human and technological

I. Caballero · F. Gualo
DQTeam/Alarcos Research Group, University of Castilla-La Mancha (UCLM), Ciudad Real, Spain
e-mail: Ismael.Caballero@uclm.es; Fernando.Gualo@uclm.es

M. Rodríguez · M. Piattini (✉)
Alarcos Research Group, University of Castilla-La Mancha (UCLM), Ciudad Real, Spain
e-mail: Moises.Rodriguez@uclm.es; Mario.Piattini@uclm.es

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_7

resources needed and used to provide more efficient support to the various operations involving data.

These data governance mechanisms must address vertical aspects related to the acquisition, holding, sharing, use, and exploitation of data in business processes while addressing cross-cutting aspects related to their management: quality, ethical and privacy aspects, interoperability, knowledge management and control over data assets through the related policies, and deployment of organizational structures with appropriate separation of data governance roles from data management roles. One of the main elements of data governance is the maturity model.

At Grupo Alarcos, we have been working for 20 years on data maturity models [4–7], which we have applied in several organizations, and we have refined and completed them with various standards and frameworks to meet the new concepts that have been progressively appearing over time, like "data governance." This evolution has given rise to the MAMD (Alarcos' Model for Data Maturity) [8], which has recently been updated following the development by the Spanish Government's Data Office and UNE (Spanish Standardization Organization) of four technical specifications for data governance (UNE 0077 [9]), data management (UNE 0078 [10]), data quality management (UNE 0079 [11]), and an assessment framework for the evaluation of organizational data maturity (UNE 0080 [12]) based on MAMD and other standards such as DAMA's DMBOK2 [13] and ISO/IEC 38505 [14, 15].

Section 7.2 summarizes the main existing data maturity models, Sect. 7.3 presents the latest version of the MAMD, and finally, Sect. 7.4 summarizes some practical applications.

7.2 Maturity Models

Similar to what happened in the software field, in which dozens of maturity models appeared – for example, CMM/CMMI [16] by SEI and ISO/IEC 15504/33000 family of standards [17–22] – several maturity models have also been created for data. In this section, we summarize the most relevant ones.

7.2.1 DAMA

Regarding the assessment of data management maturity, DAMA's DMBOK2 [13] proposes a six-level model:

- *Level 0 – no capabilities*. There are no organized data management practices or formal organizational processes to manage data. Very few organizations are typically at this level 0.

- *Level 1 – initial*. General-purpose data is managed using a limited set of tools with little or no governance. Data management is mainly dependent on a few experts. Roles and responsibilities are defined in "silos." Each data owner receives, generates, and sends data autonomously. Controls, if they exist, are applied unconsciously. Data management solutions are limited. Data quality issues are pervasive but not addressed. Infrastructure support is at the business unit level. Evaluation criteria may include the presence of process controls, such as logging data quality issues.
- *Level 2 – repeatable*. At this level, the implementation of consistent tools and role definition for process execution support arises. The organization begins to use centralized tools and provide more oversight for data management. Roles are defined, and processes do not rely solely on specific experts. There is an organizational awareness of data quality issues and concepts. The concepts of master and reference data are also recognized. Assessment criteria may include a formal definition of roles in artifacts such as job descriptions, the existence of process documentation, and the ability to leverage tools.
- *Level 3 – defined*. This level considers introducing and institutionalizing scalable data management processes as an organizational enabler. Characteristics include data replication across an organization with some controls in place and a general increase in overall data quality, along with coordinated policy definition and management. A more formal process definition leads to a significant reduction in manual intervention. This formal process and a centralized design process make process outcomes more predictable. Evaluation criteria may include the existence of data management policies, the use of scalable processes, and the consistency of data models and system controls.
- *Level 4 – managed*. Institutional knowledge gained from growth in Levels 1 through 3 allows the organization to predict outcomes when tackling new projects and tasks and begin to manage data-related risks. Data management includes performance metrics. Level 4 features include standardized data management tools and a centralized governance and planning function. The most notable improvements at this level are a measurable increase in data quality and capabilities across the organization. Evaluation criteria may include metrics related to project success, operational metrics for systems, and data quality metrics.
- *Level 5 – optimized*. When data management practices are optimized, they are highly predictable due to process automation and technology change management. Organizations at this maturity level focus on continuous improvement. At this level, tools allow data to be seen across all processes. Data proliferation is controlled to avoid unnecessary duplication. Metrics are used to manage and measure the quality of data and processes. Evaluation criteria may include change management artifacts and process improvement metrics.

7.2.2 Aiken's Model

Aiken et al. proposed in [23] a model whose main objective is to increase data management maturity levels to positively impact the coordination of data flow between organizations, human resources, and systems. To improve the organization's data management practices, this model proposes to start with a self-assessment against the maturity level and develop a road map to achieve improvement. The model states that data management consists of six interrelated and coordinated processes:

1. Data coordination program, the purpose of which is to provide an appropriate data management process and technology infrastructure
2. Organizational data integration, which is intended to achieve appropriate organizational data exchange
3. Data management, which consists of achieving the integration of data from the thematic area of the business
4. Data development, to achieve the exchange of data within a business area
5. Data operations support to provide reliable access to the data
6. Active use of data, the purpose of which is to leverage data in business activities

All organizations implement their data management practices in a way that can be classified into one of the five maturity model levels, detailed in Table 7.1.

7.2.3 Data Management Maturity (DMM) Model

The SEI (Software Engineering Institute) published the DMM (Data Management Maturity) Model [24], which is analogous to the maturity model for software processes, CMMI (Capability Maturity Model Integration), but focused on data governance, management, and quality processes.

This model was withdrawn at the end of 2021. Its content is supposed to be subsumed by the CMMI V2 model.

7.2.4 IBM Model

The IBM Data Governance Maturity Model has been developed by the IBM Data Governance Council and is focused on helping to make the strategy more effective. The maturity model defines the scope and who should be involved in governing and measuring how organizations govern their data. This model measures data governance competencies based on 11 maturity categories [25].

This maturity model consists of four interrelated groups:

Table 7.1 Data management maturity levels proposed by Aiken et al. [23]

Level	Name	Practice	Quality and predictable results
1	Initial	The organization lacks the necessary processes to sustain data management practices. Data management is characterized as ad hoc or chaotic	The organization is totally dependent on individuals, with corporate visibility into cost or performance or even awareness of data management practices. There are variable quality, low predictable results, and little or no repeatability
2	Repeatable	The organization has some knowledge of data management and can replicate some best practices and success stories	The organization delivers results with a certain quality. The most qualified personnel are assigned to critical projects to reduce risk and improve results
3	Defined	The organization uses a defined set of processes, which are published for use	Good results are obtained most of the time
4	Managed	The organization statistically forecasts and directs data management based on defined processes, cost selection, planning, and customer satisfaction. The use of data management processes within the organization is required and monitored	Reliable and predictable results and the ability to determine the progress are achieved
5	Optimizing	The organization analyzes existing data management processes to determine which ones can be improved, making changes in a controlled manner and reducing operational costs by improving performance or introducing innovative services to maintain its competitiveness	The organization achieves high levels of accurate results

- *Outcomes* are the intended outcomes of the data governance program, which tend to focus on reducing risk and increasing value and which, in turn, are driven by reduced costs and increased revenue.
- *Enablers* include areas of organizational structures and knowledge, policies, and data stewardship.
- *Core disciplines* include data quality management, data life cycle management, and data security and privacy.
- *Supporting disciplines* include data architecture, classification and metadata, and logging and audit reporting.

In each of these groups are the following 11 categories:

- *Data compliance and risk management*. A methodology in which risks are identified, rated, quantified, accepted, avoided, mitigated, or transferred

- *Value creation*. A process by which data assets are qualified and quantified to maximize the value created by the data assets
- *Organizational structures and knowledge*. Refers to the level of mutual accountability between business and IT and the recognition of fiduciary responsibility for governing data at different levels of management
- *Stewardship*. A quality control discipline designed to ensure data stewardship for asset enhancement, risk mitigation, and administrative control
- *Policy*. The written articulation of organizational performance
- *Data quality management*. Refers to methods for measuring, improving, and certifying the quality and integrity of production, testing, and archive data
- *Information life cycle management*. A systematic approach to the policy-based collection, use, retention, and disposal of information
- *Information security and privacy*. Refers to the policies, practices, and controls an organization uses to mitigate risks and protect data assets
- *Data architecture*. The architecture design of structured and unstructured data systems and applications that enable availability and distribution to appropriate users
- *Classification and metadata*. Refers to the methods and tools for creating standard semantic definitions for business and IT data models and repositories
- *Audit logging and reporting*. Refers to the organizational processes for monitoring and measuring the value and risks of data and the effectiveness of data governance

7.2.5 Gartner's Enterprise Information Management Model

Gartner states that enterprise information management cannot be implemented as a single project but that organizations must implement it as a coordinated program that evolves over time. Therefore, it proposes an information management maturity model called EIM (Enterprise Information Management), which can be adapted to support a small business unit or the entire organization.

The EIM identifies what stage of maturity organizations have reached and what actions they need to take to reach the next level. The maturity model has five levels that look at seven dimensions or building blocks that Gartner has identified as essential for information management maturity: vision, strategy, metrics, governance, people, process, and infrastructure [26].

The maturity levels and indicators themselves are aligned with the organizations' current and near-term capabilities:

- *Level 1:* Organizations are aware of key issues and changes but lack the resources, budgets, and/or leadership to address or make significant changes in EIM.
- *Level 2:* Organizations work reactively application-centric until information-related problems manifest themselves significantly in business losses or lack of competitiveness.

- *Level 3:* Organizations have become more proactive in identifying particular areas of information management and have begun identifying the organization in information systems. Some programs are operational and effective, but little leverage or alignment exists between programs and investments.
- *Level 4:* They take a managed approach to information management, committing to coordination across the organization with influential people, processes, and technologies.
- *Level 5:* Typically, model organizations in which many (if not most) aspects of information acquisition, management, and application have been optimized as tangible organizational assets with high-performance organizational structures and advanced technologies and architectures.

7.2.6 DCAM

The Data Management Capability Assessment Model (DCAM) [27] was created by members of the Enterprise Data Management (EDM) Council as a set of assessment standards to measure the level of data management capability. DCAM documents 38 capabilities and 136 sub-capabilities associated with developing a sustainable data management program.

These capabilities are specific to components, which are the artifacts to be considered in creating a data management program, according to DCAM. The components are (1) data strategy and business case, (2) data management program and funding, (3) business and data architecture, (4) data and technology architecture, (5) data quality management, (6) data governance, (7) data control environment, and (8) analytics management. Coordination of the components into a cohesive operational model ensures that controls are consistently placed throughout the life cycle in alignment with organizational privacy and security policies.

DCAM proposes a capability scoring framework with six levels, from "Not Initiated," the first level, to "Enhanced," the last level. The model is summarized in Table 7.2.

7.3 MAMD (Alarcos' Model for Data Maturity)

When developing a maturity model, it seems fundamental to us that it should be based on international standards, especially on the ISO/IEC 33000 family of standards, which bring the following advantages:

- It facilitates self-assessment.
- It provides a basis for use in process improvement and process capability determination.

Table 7.2 DCAM maturity model ([27])

Score	Category	Description
1	Not Initiated	Ad hoc management (performed by heroes)
2	Conceptual	Initial planning activities (whiteboard sessions)
3	Developmental	Engagement underway (stakeholders being recruited and initial discussions about roles, responsibilities, standards, and processes)
4	Defined	Data management capabilities established and verified by stakeholders (roles and responsibilities structured, policy and standards implemented, glossaries and identifiers established, sustainable funding)
5	Achieved	Data management capabilities adopted and compliance enforced (sanctioned by executive management, activity coordinated, adherence audited, strategic funding)
6	Enhanced	Data management capabilities fully integrated into operations (continuous improvement)

- It supports the evaluation of other process characteristics in addition to process capability.
- It produces a process rating.
- It addresses the capability of the process to achieve its purpose.
- It is appropriate for different application domains and organization sizes.
- It can provide an objective benchmark across organizational processes.

These advantages were already proven in developing maturity models for software processes such as COMPETISOFT [28] or MMIS [29] and for data processes such as MAMD [7]).

7.3.1 ISO/IEC 33000 Standards Family

The ISO/IEC 33000 family of standards for process assessment is intended to provide a structured approach to process assessment that enables an organization to (i) understand the status of its processes to process improvement, (ii) determine the suitability of its processes for a particular requirement or set of requirements, and (iii) determine the suitability of another organization's processes for a specific contract.

The process assessment includes the determination of the organization's needs, an evaluation (measurement) of the processes used by the organization, and an analysis of the current state of those processes. The results of the analysis will be used to guide process improvement activities or to determine the capability of the processes employed by an organization.

The following paragraphs summarize the ISO/IEC 33000 parts used as the basis for the development of MAMD:

- *ISO/IEC 33001: Concepts and terminology* [18]. This standard provides a glossary of terms related to the conduction of process assessment and a general introduction to the concepts and standards for process assessment in the ISO/IEC 33000 family of standards. It provides general information on the concepts of process assessment, the application of process assessment to evaluate compliance with process quality characteristics, and the application of process assessment results to process management. It describes how the parts of the family of standards for process assessment fit together, provides guidance for their selection and use, and explains the requirements in the suite and their applicability to the conduct of assessments.
- *ISO/IEC 33002: Requirements for performing process assessment* [19]. This standard establishes the requirements for performing an assessment to ensure consistency and repeatability of the values and results obtained during process assessment. These requirements help to ensure that assessment results are consistent and provide evidence to substantiate ratings and verify compliance with requirements.
- *ISO/IEC 33003: Requirements for process measurement frameworks* [20]. This standard provides requirements that apply to process measurement frameworks that support and enable the assessment of process quality characteristics.
- *ISO/IEC 33004: Requirements for process reference models, process assessment models, and maturity models* [21]. This standard establishes requirements for constructing and verifying process references, process assessment, and maturity models. The requirements defined in this international standard form a structure that specifies:
  - The relationship between the classes of process models associated with the performance of process evaluation
  - The relationship between the process reference models and the prescriptive/normative models of process realization
  - The integration of process reference models and process measurement frameworks that establishes process assessment models
  - A standard set of process realization and quality assessment indicators that are used in process assessment models
  - The relationship between maturity models and process assessment models and the degree to which a maturity model can be constructed using elements from different process assessment models
- *ISO/IEC 33020: Process measurement framework for process capability assessment* [22]. This standard defines a process measurement framework that supports assessing process capability following ISO/IEC 33003 requirements. The process measurement framework provides an outline for building a process assessment model (according to ISO/IEC 33004), which can be used during the process capability assessment following the requirements set by ISO/IEC 33002. The

standard considers the capability of the process to meet current or future business objectives. The process measurement frameworks defined in this part of the standard form a structure that (a) facilitates self-assessment, (b) provides a basis for use in process improvement and process quality determination, (c) applies to all domains and sizes of the organization, (d) produces a set of process attribute ratings, and (e) enables a process capability level to be derived.

7.3.2 MAMD Overview

MAMD is two-dimensional (Fig. 7.1), whose first dimension defines the different *processes* to be evaluated and their expected outcomes if correctly implemented. In the case of MAMD, the processes to be used will be those defined in the technical specifications for data governance [9], data management [10], and data quality management [11]. In the second dimension, the model deals with the *capability of the process*, which consists of a series of process attributes grouped into capability levels and which identify whether the process, in addition to being implemented (level 1), is managed (level 2), established (level 3), predictable (level 4), or innovating (level 5).

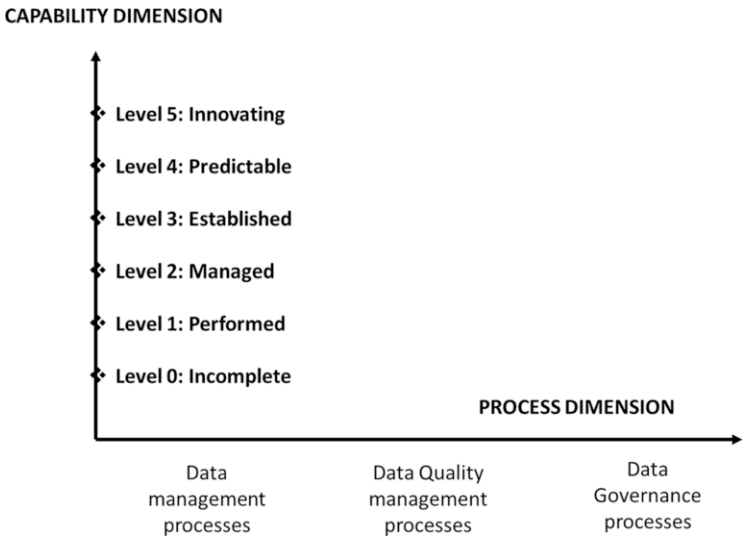

Fig. 7.1 MAMD overview

7.3.3 The Capability Dimension

For the measurement of the capability of a process, ISO/IEC 33020 defines a set of process capability levels and their corresponding *process attributes (PA)*. It is important to note that a process must meet the process attributes of that level and the process attributes of the levels above it to achieve a capability level. The list of process attributes and capability levels is shown in Table 7.3.

Within the process measurement framework proposed by the ISO/IEC 33000 family of standards, a process attribute is a measurable property of the process capability, which is measured using the following ordinal scale:

- *(N) Not implemented*: There is little or no evidence of achievement of the defined process attribute in the assessed process. As an indication, a process attribute is considered "not implemented" if the degree of its achievement is in the range between 0 and ≤15%.
- *(P) Partially implemented*: There is some evidence of a focus and some achievement of the process attribute defined in the assessed process. The process attribute is considered "partially implemented" if the degree of achievement of the attribute is >15% and ≤50%.
- *(L) Largely implemented*: There is evidence of a systematic approach and significant achievement of the defined process attribute in the assessed process. If the degree of achievement of the attribute is >50% and ≤85%, then the process attribute can be evaluated as "largely implemented."
- *(F) Fully implemented*: There is evidence of a complete and systematic approach and full achievement of the defined process attribute in the assessed process. The process attribute is considered to be "fully implemented" if the degree of achievement of the attribute is >85% and ≤100%.

To evaluate the first capability level, which includes the process attribute "PA 1.1 Process realization," it is necessary to check that the specific process achieves the specific process outcomes indicated in the process definition as gathered in the

Table 7.3 Capability levels and process attributes

Capability level	ID	Process attribute
Level 0. Incomplete process		
Level 1. Performed process	PA 1.1	Process realization
Level 2. Managed process	PA 2.1	Realization management
	PA 2.2	Work product management
Level 3. Established process	PA 3.1	Process definition
	PA 3.2	Process deployment
Level 4. Predictable process	PA 4.1	Quantitative analysis
	PA 4.2	Quantitative control
Level 5. Innovating process	PA 5.1	Process innovation
	PA 5.2	Innovation implementation

Table 7.4 Process for the establishment of organizational structures

Process Id.	OrgStr
Name	Establishment of organizational structures for data governance, management, and use
Purpose	This process aims to create and maintain the organizational structures necessary to assume the responsibilities related to the governance, management, and use of data; these structures must be provided with sufficiently skilled human resources to address these responsibilities successfully
Process outcomes	PO1. The most appropriate working model for data governance, management, and use is chosen PO2. The organizational structures necessary to perform data governance, data management, and data quality management are created and maintained PO3. Chains of authority, responsibility, and accountability are established to enable decision-making and conflict resolution in data governance, management, and use PO4. Escalation mechanisms are established for decision-making and problem-solving PO5. The skills, knowledge, and competencies required for the roles that will perform the established responsibilities are identified PO6. It is ensured that the people who perform the specific roles related to the data have the identified knowledge and skills PO7. The performance of organizational structures is monitored
Base practices	Define an organizational structure for data governance, management, and use [PO1, 2, 3, 4] Establish the necessary skills and knowledge [PO5, 6] Monitor the performance of organizational structures [PO7]
Work products • Organizational structures for data governance, management, and use [PO1, 2] • Authority levels of the components of organizational structures [PO3] • Chains of responsibility and accountability of organizational structures [PO3, 4] • Stakeholder communication and control mechanisms [PO4] • Knowledge, skills, and competencies needed to perform the responsibilities assigned to each role [PO5, 6] • Reports on the degree of performance of organizational structures [PO6, 7]	

process reference model (see Table 7.4). This evaluation is particular for every process since the process outcomes are specific for every process. On the other hand, for evaluating capability levels 2 to 5, the process attributes in Table 7.3 are used; evaluating these attributes is cross-cutting to all processes.

The process and process attribute results can be characterized as an intermediate step to provide a process attribute rating. Based on the results obtained in assessing each of the process attributes of a specific process under evaluation, a rating of the capability level of that process can be issued. This is achieved by an aggregation method based on the assumption that a process has a given capability level if all process attributes of the previous levels have a rating of "Fully Achieved" (F) and the process attributes of that capability level have a rating of at least "Largely Achieved" (L).

7.3.4 Process Dimension

The process dimension is constituted by the processes of the three technical specifications mentioned above [9–11]. Each process is described in terms of its name, purpose, and process outcomes; base practices, work products, and their relationship to the process results are also included. For example, the establishment of organizational structures process is presented in Table 7.4.

7.3.5 Organizational Maturity Model

MAMD is aligned to ISO 8000-61 [30] and ISO 8000-62 [31] and consists of five maturity levels, as shown in Fig. 7.2.

The maturity levels proposed in MAMD, along with their meaning and the processes included, are detailed below:

- *Maturity Level 1 – Accomplished*
  At this level, the organization can demonstrate the use of a set of best practices to provide the minimum necessary support for managing the data required in its business processes. An organization at this level pays no attention to data governance or quality. The processes that are included in maturity level 1 are:
  - Data processing
  - Data technology infrastructure management

- *Maturity Level 2 – Managed*
  The organization can demonstrate the execution of best practices to control the data quality used in its business processes. Therefore, there is some evidence of

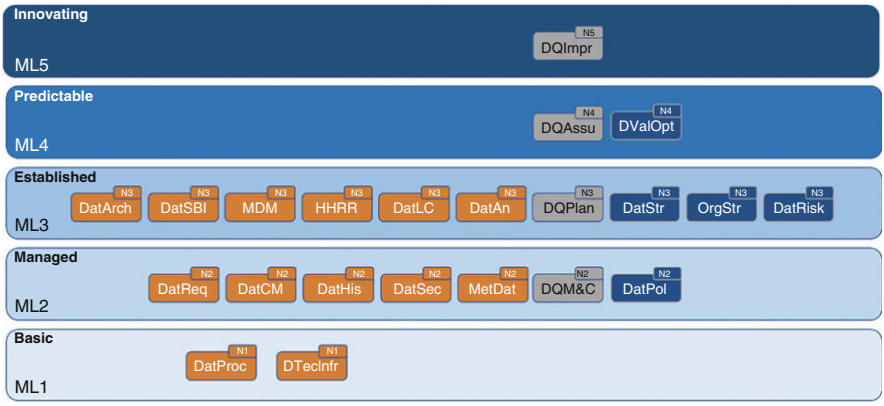

Fig. 7.2 MAMD maturity model for data governance, data management, and data quality management

the assurance that the organization has the minimum necessary data management processes in place to provide an acceptable outcome for its business processes. The processes included in maturity level 2 are:

- Data requirements management
- Data configuration management
- Historical data management
- Data security management
- Metadata management
- Data quality monitoring and control
- Establishment of data policies, best practices, and procedures related to data governance

- *Maturity Level 3 – Established*

  The organization can demonstrate that it uses the complete set of data management best practices to ensure that the data used in its business processes are of appropriate levels of quality and that the data used in its business processes are aligned with organizational strategy. The processes included in maturity level 3 are:

  - Data architecture and design management
  - Data sharing, brokerage, and integration
  - Master data management
  - Human resources management
  - Data life cycle management
  - Data analytics
  - Data quality planning
  - Establishment of data strategy
  - Establishment of organizational structures for data governance, management, and use of data
  - Data risk optimization

- *Maturity Level 4 – Predictable*

  The organization can demonstrate that they use a set of best practices to monitor that the organizational data strategies are genuinely effective, enabling it to ensure data quality and optimize data value. The processes included in maturity level 4 are:

  - Data quality assurance
  - Data value optimization

- *Maturity Level 5 – Innovation*

  The organization can demonstrate that it uses a set of best practices to ensure that data governance, management, and data quality management processes are continuously improved to optimize data value and reduce risks, contributing to the organizational strategy. The process included in maturity level 5 is:

  - Data quality improvement

7.4 Practical Applications of MAMD

MAMD has been successfully applied to different organizations, public and private, with mainly three purposes:

1. Define projects to select and implement or improve the data governance, data management, and data quality processes that most contribute to better support of the organizational data strategy. Examples of experience covering this purpose are listed in Subsections 7.4.1–7.4.4.
2. Assess the level of organizational data maturity to improve the less capable processes. Examples of these experiences are introduced in Subsections 7.4.5 and 7.4.6.
3. Combine MAMD as a body of knowledge with some other domain-specific frameworks to tailor new maturity models for domains considering the specific concerns of data governance, data management, and data quality management in the domain. Examples of this type or purpose are covered in Sects. 7.4.7–7.4.9.

In the following subsections, we describe some interesting experiences of using MAMD.

7.4.1 Regional Government: Improving the Performance of Authentication Servers

This experience was conducted in a Spanish regional government. The people in charge of the IT area discovered that they have severe problems with the performance of authentication servers for the applications supporting public services. The reason was that too many user accounts were created: some of them for the regular functioning of public services (e.g., new public servants were hired, and their corresponding user accounts needed to be created to let them work), some others for temporal services (e.g., teachers or physicians who were hired for limited and seasonal time, and their user accounts were blocked but not removed), and some others for some uncontrolled purposes (e.g., IT technicians created some user accounts as part of a testing process, and not later removed after the testing).

The problem was faced from the point of view of data quality management, considering the user account log files to be explored as a "user account" data repository. In this context, MAMD's "data quality monitoring and control" process was used to create a better systematic and rigorous approach. The idea was to define some business rules about "user authentication management" to identify and reduce the number of unnecessary user accounts. The expected consequence was an increase in the performance of authentication servers as they need to manage fewer users. After this first stage, the goal was to consider other MAMD processes as a reference to provide well-defined and customized procedures for IT users to prevent the authentication servers from suffering the same problems again.

7.4.2 Insurance Company: Building a "Source of Truth" Repository

The second experience is related to a large insurance company. Due to regulatory compliance, the insurance company must build specific reports to be submitted to the national agency to meet Solvency II's requirements. These reports were built upon data from different transactional databases related to the insurance operations (e.g., new insurance policies contracts and customers' claims). These reports are vital to determine the company's capability to keep on the market, as Solvency II requires. Consequently, the data used to produce this report must be of the highest quality possible. The insurance company invested lots of resources in assuring the quality of the data coming from every transactional data source, and consequently, they were at high risk of making any mistakes.

To prevent these mistakes, the insurance company decided to build a master data repository that could be used as the only source of truth from which data required to build the report was extracted. After introducing MAMD to the people in charge of the project, the initiative was no longer understood as a technological project but also a managerial project in which data had to be conveniently governed. The project was structured in three stages:

1. Development of the "source of truth" repository. This repository was a master data repository. Several data management processes from MAMD ("data requirements management"; "data architecture and modeling"; "data technology infrastructure management"; "master data management"; "data sharing, brokerage, and integration"; "data security management"; and "data configuration management") were considered the essential reference for this subproject's stage.
2. Improvement of the data quality in the "source of truth" repository. Once the repository was populated with data from various sources, people in charge of the initiative considered some MAMD processes ("data quality management" processes of "data quality planning," "data quality monitoring and controlling," and "data quality assurance") to improve the current state of quality of the data. They also use these processes to revisit the ETL processes feeding the master data repository to ensure that the collected data quality requirements were correctly implemented. Interestingly, they consider this stage as iterative and incremental, assuming some risk in every iteration of the stage and leading efforts to reduce the existing risks continuously.
3. Governance of the data in the "source of truth" repository. People in charge of the project understood over time that the repository became an essential asset for the company. Consequently, they became convinced of the need to maintain the data contained in the master data repository aligned with the organizational strategy. To achieve this goal, they complemented the data life cycle information with some policies (not only to meet Solvency II requirements and better performance in other data operations). Some other MAMD processes, like "data life cycle

management" and "establishment of data policies, best practices, and procedures related to data governance," were followed to support these last stages.

7.4.3 Bicycle Manufacturer: Enabling Better Analytics

This experience was conducted in one of the largest Spanish bicycle manufacturers and vendors, which sells their products all over the globe. They were interested in improving their capability to produce better sales data analytics to characterize their customers better and become closer to their needs. They maintain an extensive database of the products (not limited to bicycles) they have been selling during the last years, their customers, and the punctual iterations that any potential client could have done on their landing web page. The main problem in achieving their goal is the inadequate levels of quality of this data.

Consequently, they launched a data quality assessment project for a sold product data repository. This project was grounded on the MAMD's "data quality monitoring and control" process. Several weaknesses in the organizational way of working with data were revealed during the project. Consequently, the company realized that it had a structural problem (which generated the decay of almost data repositories in the organization) that had to be addressed not to threaten its sustainability.

To provide a solution, MAMD was introduced to the people in charge of data management as a reference framework to adapt their working methods. On this occasion, the project embraced two stages:

1. Improvement of the quality of the data repositories. As the structure of their applications involved several isolated data repositories, the people in charge of the project were more worried about how to define systematic procedures to act consistently over the various data repositories. Their goal was first to clean their databases to have data with adequate levels of quality to launch the analytics initiatives. The need to improve data quality came after realizing the unsatisfactory results of the first stage of the analytics process, which motivated them to focus on data quality to avoid the waste of resources in the analytics projects. Thus, they felt highly motivated to develop typical data quality evaluation and improvement procedures following the MAMD's "data requirements management" and "data quality monitoring and control" processes. One attractive advantage of this approach was that they could connect the data quality requirements of the several types of analytics with monitoring and controlling the levels of quality of the datasets, producing better-fitted data for the analytics.
2. Improvement of the way of working. One stunning discovery was that having databases simultaneously in preparation (e.g., cleansing) and production: as soon as the database came into production, the level of quality began to decay. The reason was that the data production processes were not working correctly, and they put data with inadequate quality into the just-cleansed database. Consequently, the need to review the data production process (mainly those related to stock management) and define some data policies quickly became one crucial part

of the main project. Once again, MAMD was proposed as a reference framework to implement and put into production the corresponding artifacts. In this sense, the processes "data requirements management," "data processing," "data analytics," and "establishment of data policies, best practices, and procedures related to data governance" were considered.

7.4.4 Telco Company: Building a Data Marketplace

This experience was conducted in a large telco company. This company has invested many resources in developing a data lake as part of the infrastructure to provide new data services to different business processes. Nevertheless, the data lake was not the only internal data provider: some other data resources (multiple types of master data repositories, several data warehouses, and several analytical units) were available to provide, most of the time, overlapping data services. This situation caused a great deal of distrust on the part of the workers, who did not know which data source they should use for their purpose. People in charge of the data lake project had launched previously specific local data governance initiatives and were acquiring solid knowledge. As they realized the risks of having several data providers, they wanted to share the acquired knowledge with the other data providers for the company's benefit. One of the most critical conclusions the company raised was that they need to unify and demure all the overlapped data services creating a data marketplace and providing as much information as possible to the potential stakeholders about the provided services and the possible utilization as part of the various business processes of the company.

MAMD was introduced to face the development of the data marketplace. It was agreed that several processes could primarily help to design a solution, which was not only a technological concern. In this sense, the processes "data requirements management"; "data architecture and design management"; "data sharing, brokerage, and integration"; "data quality monitoring and control"; "establishment of data policies, best practices, and procedures related to data governance"; and "metadata management" were considered essential as reference.

7.4.5 Hospital/Faculty of Medicine: Assessing the Organizational Maturity

This experience corresponds to evaluating the organizational maturity concerning data governance, data management, and data quality management of a hospital/faculty of medicine [32]. To this aim, the maturity assessment introduced in ISO/IEC 33000 was followed. The assessment first involves selecting several business processes to look for evidence of implementing the best practices related to data

governance, data management, and data quality management. In this case, these three processes were selected:

– As main process (MP): pharmacology data repositories maintenance
– As auxiliary process 1 (AP1): biostatistics report generation
– As auxiliary process 2 (AP2): clinical software maintenance

The assessment scope was established at the maturity level 2. Consequently, the inspection of the MP, AP1, and AP2 involved the searching of evidence for all data governance, data quality management, and data management included in the maturity levels 1 and 2 for the process attributes PA 1.1, PA 2.1, and PA 2.2.

Based on the strength of the found evidence, a score was given for every process/process attribute, and the conclusion was that the maturity level of the hospital/faculty of medicine had consolidated only the maturity level 1.

With this information, the people in charge of the hospital/faculty of medicine decided that the obtained maturity level was insufficient to ensure adequate results for the selected business processes, and they launched several projects to fix the problems.

7.4.6 University Library: Assessing the Organizational Maturity

This experience was conducted in a Spanish university library [7]. This project's main aim was to assess the organizational maturity level of the library to determine how well they were governing and managing the data. This requirement was essential for them because they needed to internally share data with other university organizations and externally with other university libraries and other institutions of public administration.

Similar to the previously described experience, several business processes were chosen as the source of evidence of the adequate implementation of the data governance, data quality management, and data management processes included in MAMD. On this occasion, the selected processes were:

– As main process (MP): cataloging procedure
– As auxiliary process 1 (AP1): funds movement procedure
– As auxiliary process 2 (AP2): user load procedure/external users

The maturity assessment was scoped to maturity level 2. It was relatively easy to determine that the university library has achieved maturity level 1. As the head of the library considered that achieving maturity level 2 would bring significant benefits to the institution, they decided to launch a process improvement project to amend the various problems found during the internal audit. Several corrective actions affecting the working methods and the data repositories were successfully executed in this sense. As a consequence, almost all problems were fixed. The university library

decided to go ahead with an external certification audit to be granted a certificate. AENOR International oversaw conducting the external audit for certification, and they found that the university library has achieved the required rating for the process attributes for the maturity level 2. Consequently, AENOR International granted the university library a certificate of maturity level 2 (see Fig. 7.3).

7.4.7 DQIoT: Developing a MAMD-Based Maturity Model for IoT

As part of the Eureka Project DQIoT (UCTR170338), [1] an adaptation of MAMD has been made for the IoT [33]. The adaptation includes a Process Reference Model and a Maturity Model.

7.4.8 Regional Institute of Statistics: Developing a MAMD-Based Model for the Official Statistics Domain

In this experience, MAMD has been combined with other international standards to develop the Statistic Business Process Reference Model (SBPRM) following the recommendations provided in ISO 9001 [34] and those provided by the Generic Statistical Business Process Model (GSBPM) [35], the reference framework for statistics production defined by UNECE.

The contribution of every framework is the following:

- ISO 9001 provides the structure of the processes included in the framework and the necessary mechanisms related to the quality management of the process. Three groups of processes have been identified: strategic processes, main processes, and support processes.
- GSBPMv5.1 provides the concepts and the content for every statistic process.
- MAMD enables the enrichment of the processes, including the best practices of data governance, data quality management, and data management.

This Statistic Business Process Reference Model is to be used as the basis for running the official statistics of the Regional Institute of Statistics. The regional government will use the results to develop policies that will improve the well-being of the citizens.

[1] Executed in collaboration with the Spanish University of Castilla-La Mancha, the Korean University of Myongji, the Spanish companies Lucentia Lab and IE, and the Korean company GTOne. More information at https://alarcos.esi.uclm.es/proyectos/DQIoT/index.php

AENOR

Certificado de Nivel de Madurez
Procesos de Gobierno y Gestión de la Calidad de Datos

AENOR ER Madurez Datos 2 ISO 33000

NMD-2018/0001

AENOR certifica que la organización

BIBLIOTECA UNIVERSITARIA

se ha evaluado la madurez de los procesos de gobierno y gestión de la calidad de datos conforme con la Norma ISO/IEC 33000 e ISO/IEC 8000 (MAMD)

y se ha alcanzado el nivel: 2

para las actividades: Gestión de Datos

que se realizan en: Biblioteca Universitaria (servicios centrales). Avda. de Camilo José Cela, 6,

Fecha de primera emisión: 2018-02-26
Fecha de expiración: 2021-02-26

Rafael GARCÍA MEIRO
Director General

AENOR INTERNACIONAL, S.A.U.
Génova, 6. 28004 Madrid. España
Tel. 91 432 60 00.- www.aenor.com

Original Electrónico

Fig. 7.3 Certification of data maturity level 2 for a university library granted by AENOR Intl

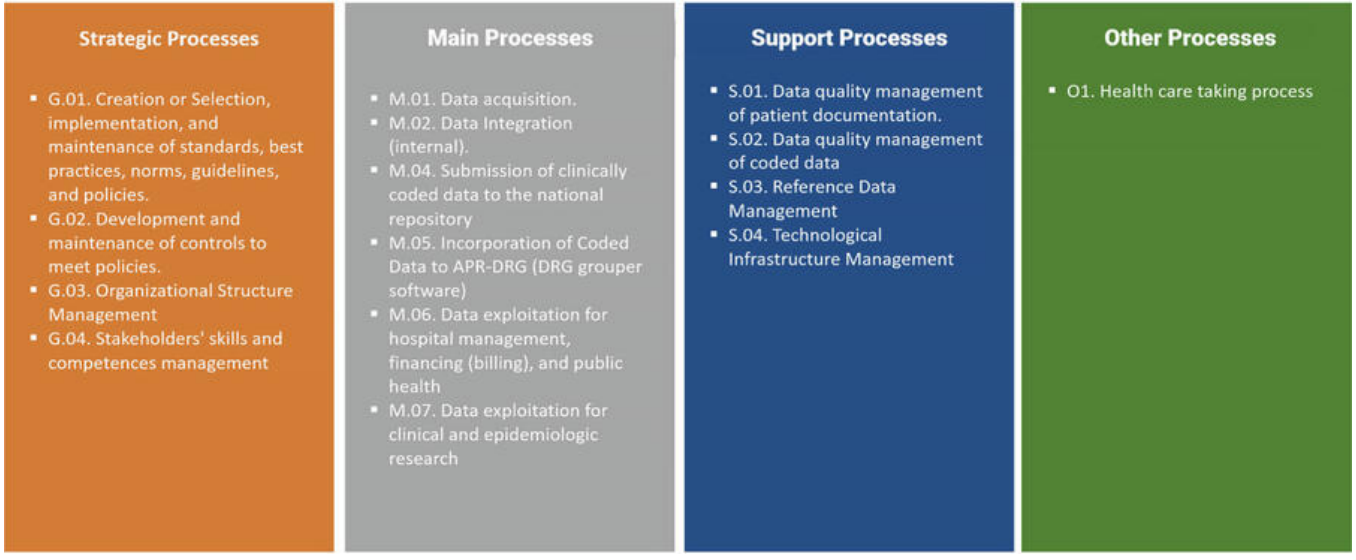

Fig. 7.4 Process included in CODE.CLINIC [36]

7.4.9 CODE.CLINIC: Tailoring MAMD for Coding Clinical Data

Coding medical data is a crucial previous step for many activities in healthcare management since it is the basis for several activities ranging from hospital reimbursement to clinical research [36]. This activity is prone to many types of error, and it was considered necessary to identify the best practices related to clinical coding to prevent healthcare organizations from these errors. However, considering how data-intensive these best practices are, they will be benefited from being enriched with some others related to data quality management and governance. As a result, CODE.CLINIC, a framework that can be used to support institutions in coding their medical data better, was developed. This framework consists of two main components: a Process Reference Model (PRM) and a Process Assessment Model (PAM) based on MAMD. Figure 7.4 gathers the CODE.CLINIC PRM, which gathers 16 processes grouped into 4 blocks. More information about CODE.CLINIC is introduced in Chap. 11 of this book.

Acknowledgments This work has been partially funded by the ADAGIO project (Alarcos' DAta Governance framework and systems generatIOn), JCCM Consejería de Educación, Cultura y Deportes, and FEDER funds (SBPLY/21/180501/000061).

References

1. Aiken, P.: EXPERIENCE: succeeding at data management—BigCo attempts to leverage data. J. Data Inf. Qual. **7**, 1–2 (2016). https://doi.org/10.1145/2893482
2. European Data Strategy: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy. Accessed 02 May 2022
3. Guy Pearce: Beware the traps of data governance and data management practice. ISACA J. **6**, 23–31 (2022)
4. Caballero, I. et al.: Getting better information quality by assessing and improving information quality management. In: Proceedings of the Ninth International Conference on Information Quality (ICIQ-04), 9th edn (2004)
5. Caballero, I., et al.: IQM3: information quality management maturity model. J. Univers. Comput. Sci. **14**(22), 3658–3685 (2008). https://doi.org/10.3217/jucs-014-22-3658
6. Caballero, I., Piattini, M.: CALDEA: a data quality model based on maturity levels. In: Presented at the Third International Conference on Quality Software, 2003. Proceedings. IEEE (2003)
7. Carretero, A.G., et al.: MAMD 2.0: environment for data quality processes implantation based on ISO 8000-6X and ISO/IEC 33000. Comput. Stand. Interfaces. **54**, 139–151 (2017)
8. DQTeam: Modelo Alarcos de Madurez de Datos v4.0. https://dqteam.es/mamd/ (2023)
9. UNE: Especificación UNE 0077: 2023, Gobierno del Dato (2023)
10. UNE: Especificación UNE 0078:2023, Gestión del Dato (2023)
11. UNE: Especificación UNE 0079:2023, Gestión de Calidad del Dato (2023)
12. UNE: Especificación UNE 0080:2023, Gestión de Evaluación del Gobierno, Gestión y Gestión de Calidad del Dato (2023)
13. DAMA: DAMA-DMBOK: data management body of knowledge. Technics Publications, LLC (2017)
14. ISO: ISO/IEC 38505-1:2017 Information technology — governance of IT — governance of data — Part 1: application of ISO/IEC 38500 to the governance of data https://www.iso.org/standard/56639.html. Accessed 09 May 2021

15. ISO: ISO/IEC TR 38505-2:2018 Information technology — Governance of IT — Governance of data — Part 2: Implications of ISO/IEC 38505-1 for data management, https://www.iso.org/standard/70911.html. Accessed 23 May 2021
16. CMMI Product Team: CMMI for Development v1.3. https://doi.org/10.1184/R1/6572342.v1 (2018)
17. ISO: ISO/IEC 15504-1:2004 Information technology — process assessment — Part 1: Concepts and vocabulary. https://www.iso.org/standard/38932.html (2004)
18. ISO: ISO/IEC 33001 -- Information technology -- process assessment -- concepts and terminology (2015)
19. ISO: ISO/IEC 33002 -- Information technology -- process assessment -- requirements for performing process assessment (2015)
20. ISO: ISO/IEC 33003:2015: Information technology — process assessment — requirements for process measurement frameworks. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/41/54177.html. Accessed 11 April 2022
21. ISO: ISO/IEC 33004:2015: Information technology — process assessment — requirements for process reference, process assessment and maturity models. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/41/54178.html. Accessed 11 April 2022
22. ISO: ISO/IEC 33020 -- Information technology -- process assessment -- process measurement framework for assessment of process capability (2015)
23. Aiken, P., et al.: Measuring data management practice maturity: a community's self-assessment. Computer. **40**(4), 42–50 (2007)
24. Mecca, M., et al.: Data management maturity (DMM) model. CMMI Institute (2014)
25. Soares, S.: The IBM Data Governance Unified Process: Driving Business Value with IBM Software and Best Practices. MC Press, LLC (2010)
26. Gartner: Gartner's Enterprise Information Management Maturity Model. https://www.gartner.com/en/documents/3236418 (2016)
27. EDM Council: The Data Capability Assessment Model (DCAM) Framework v2.2 Overview. https://cdn.ymaws.com/edmcouncil.org/resource/collection/AC65DC50-5687-4942-9B53-3398C887A578/DCAM_Framework_v2_Overview_v2.2.1.pdf (2020)
28. Oktaba, H., et al.: Software process improvement: the COMPETISOFT project. Computer. **40**(10), 21–28 (2007)
29. Pino, F., et al.: Modelo de Madurez de Ingeniería del Software V2.0 (MMIS V.2). AENOR, Madrid (2018)
30. ISO: ISO 8000-61:2016: Data quality — Part 61: Data quality management: Process reference model. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/30/63086.html. Accessed 04 August 2021
31. ISO: ISO 8000-62:2018: Information technology — Process assessment — Requirements for process reference, process assessment and maturity models. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/53/65340.html. Accessed 11 April 2022
32. Carretero, A.G., et al.: A case study on assessing the organizational maturity of data management, data quality management and data governance by means of MAMD. In: Proceedings of the 21st International Conference on Information Quality, ICIQ 2016, Ciudad Real, Spain, June 22–23, 2016, pp. 75–84. Curran Associates (2016)
33. Kim, S., et al.: Organizational process maturity model for IoT data quality management. J. Ind. Inf. Integr. **26**, 100256 (2022). https://doi.org/10.1016/j.jii.2021.100256
34. ISO: ISO 9001:2015 Quality management systems — requirements. ISO (2015)
35. UNECE: Generic Statistical Business Process Model, GSBPM v5.1. UNECE (2019)
36. Caballero, I., et al.: Towards a process reference model for clinical coding. In: Quality of Information and Communications Technology - 15th International Conference, QUATIC 2022, Talavera de la Reina, Spain, September 12–14, 2022, Proceedings, pp. 190–204. Springer (2022). https://doi.org/10.1007/978-3-031-14179-9_13

Part II
Data Governance Applied

Chapter 8
Data Governance in the Banking Sector

Raúl Cruces Rufo

8.1 Inception, Challenges, and Evolution

The inception of the data management and governance (DM&G) function in the financial industry, led by the chief data officer (CDO), was mainly regulatory driven.

The European Central Bank [1] published in January 2013 the new risk data aggregation and risk reporting principles [2] (the BCBS 239 principles), applied in full on January 1, 2016, for Global Systemically Important Banks (G-SIBs). They implied improvements in data governance, reporting, metrics, data quality (DQ), and technological infrastructure. On top of that, a data and information self-assessment (DISA) process should measure periodically the degree of compliance.

[1] The European Central Bank (ECB) (https://www.ecb.europa.eu/home/html/index.en.html) is the central bank for the euro and administers monetary policy within the eurozone, which comprises 19 member states of the European Union and is one of the largest monetary areas in the world. Established by the Treaty of Amsterdam, the ECB is one of the world's most important central banks and serves as one of the seven institutions of the European Union, being enshrined in the Treaty on European Union (TEU). The bank's capital stock is owned by all 27 central banks of each EU member state (https://en.wikipedia.org/wiki/European_Central_Bank).

[2] Risk data aggregation and risk reporting principles (the BCBS 239 principles) (https://www.bis.org/publ/bcbs239.pdf). BCBS 239 is the Basel Committee on Banking Supervision's standard number 239. The subject title of the standard is "Principles for effective risk data aggregation and risk reporting." The overall objective of the standard is to strengthen banks' risk data aggregation capabilities and internal risk reporting practices, in turn, enhancing the risk management and decision-making processes at banks. The standard was published in January 2013 and applied in full on January 1, 2016, for Global Systemically Important Banks (G-SIBs) who were defined as such no later than November 2012, otherwise 3 years after their designation as G-SIBs. The standard also recommends that it is, by the national supervisors, applied to Domestic Systemically Important Banks (D-SIBs) 3 years after their designation as such (https://en.wikipedia.org/wiki/BCBS_239).

R. C. Rufo (✉)
Banco Santander, Madrid, Spain

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_8

However, the world is changing. Communication between people and between them and companies is not the same as those of the past. IT leads communications and millions of data are transferred per minute around the world. Banking industry, as a result of this change, is facing five major challenges:

(i) Banks must continue the transformation of their business to better serve their customers in the future. In the last few years, thanks to their global platforms, banks have made great progress serving most segments as Wealth Management & Insurance (WM&I), Corporate & Investment Banking (CIB), and trade, merchants, and payment services for small and medium enterprises (SME). However, banks continue to see a huge opportunity to improve how they serve individual customers. Banks do many things extremely well, but they still have far too many products, and they have still room to improve their customer experience.

The relationship with customers evolves. Branches have ceased to be the physical meeting space for managers and customers. Market leadership is no longer determined by the density of the branch's physical network. Today banks' customers manage their own agenda, communicating with them through their smart devices. We have entered in the digital age and banks who best understand, process, and use the data derived from their clients' digital interaction will be the leaders of the coming decades. So, the main channel is changing very quickly. Bank branches are moving to multichannel, meaning increase of interactions on digital channels.

Some banks' vision to win in individual segment is to become digital banks with branches. To cope with it, they are implementing plans to deliver on this vision based on simplification, leveraging service automation through innovative and common global technology, and developing value-add branch solutions.

(ii) In fact, new players who neither are bankers nor have ever had a branch to open accounts, grant loans, or process insurances have emerged in the banking scenario. And they are trying to mix with banks at the same level. Until recently competitors were other financial entities with the same regulatory requirements. However new major competitors are emerging from other sectors, without the same regulatory requirements.

(iii) Personalized knowledge by person-to-person contact has changed into a one by data analytics (DA), machine learning (ML), and artificial intelligence (AI). We have entered the age of knowledge, of the client. The greater and higher quality of the information banks have about their activities and needs, the closer they will be to maintain their privileged position as a reference bank.

(iv) Sectorial and geographical diversification matters and is differential. Geographical diversification means holding business from different regions. Banks do not want all their business in a single country or region for the same reason they do not want it all in a single sector. The failure of that would be a huge blow to their performance. So, banks are very keen on investigating the relationship between the diversification and performance through several data sources. For

example, banks use Return on Assets (ROA) and Return on Equity (ROE) as measure of performance and Herfindahl Index (HI) [3] as a measure of diversification. The number and the amount of credits, deposits, credit cards, and insurance are employed as control variables. According to the result of the analysis, it is determined that dependent variables ROA and ROE are explained by diversification.

(v) The leader was the one who had the best bank managers and now the leader is the one who has more and better data. Banks do not play with data. They progress and strengthen their ability to respond immediately to clients and markets.

These major challenges lead to a necessary evolution of the DM&G function market trends, from a transformational leader in 2014 and 2015 to a business and analytics enabler from 2016 to today. However, this is not the end of the trip, as the critical goal is to become a data-driven bank enabler in the near future.

The CDO role emerged to provide appropriate DM&G throughout the whole bank. Core functions performed included data controls and governance, quality and metadata. Regulation and compliance acted as big levers of pressure to create the CDO role which focused mainly on implementing foundational technologies.

From 2016 on the CDO role starts to take ownership of additional responsibilities, starting to deliver tangible business value through advanced analytics, both by creating centers of expertise and addressing analytics problems. Well-established data strategy is focused on delivering prioritized use cases, supported by a multi-year road map. There is a material progress in implementing a strategic data architecture based on reputable golden sources, simplification, and new technologies, as well as the enablement of processes optimization. Also, the focus is on a fully implemented operating model and data control across critical elements and reports, enabling transparency and increased DQ.

But this is not the end of the trip. What about the future? DM&G function led by the CDOs must become a data-driven bank enabler. The need of a continued emphasis on the role as a strategic business enabler is required as data becomes a valuable asset for the company and a source of competitive advantage, being treated like that at a company board level, enabling data monetization, full end-to-end process optimization, and cost reduction. CDOs must drive a data-driven organization, which means:

(i) Data culture embedded across the board for all decision-making processes
(ii) Use of advanced data analytics, machine learning, and artificial intelligence to solve complex problems, as well as for the long tail of day-to-day issues

[3] The Herfindahl Index (also known as Herfindahl–Hirschman Index, HHI, or sometimes HHI-score) is a measure of the size of firms in relation to the industry they are in and is an indicator of the amount of competition among them. Named after economists Orris C. Herfindahl and Albert O. Hirschman, it is an economic concept widely applied in competition law, antitrust, and also technology management (https://en.wikipedia.org/wiki/Herfindahl%E2%80%93Hirschman_index).

(iii) Robust technological platform and fit-for-purpose DM&G tools ecosystem to effectively manage and exploit data

Accordingly, banks created DM&G functions with two main objectives: (1) to position themselves as the best banking institution providing positive customer service and support and (2) to comply with the new regulations enforced in response to the 2008 financial crisis. This means to keep balance between two dimensions, customers and regulation.

On the one side of the balance scale are increasingly demanding customers looking for uniqueness and multichannel experience, leading banks' urgency to know their behavior patterns and needs, and on the other side of the scale is more demanding regulatory environment in terms of additional regulatory information, DQ, data protection, confidentiality and portability, and open data.

Regulations are slower than the interest of banks' customers. Regulations are reactive moving behind customers. So, if banks want to be data-driven, they cannot only abide by regulators and supervisors. Banks must anticipate them and, more importantly, to their own customers, who are proactively demanding bank exclusivity, new products, and multichannel digital experience.

Therefore, having precise and quality data for the development of analytical models to help to maintain and take care of current customers, as well as to attract new ones, is fundamental for banks.

8.2 Data-Driven Bank

Data is a global strategic pillar at every bank, which is the driver across them for the data-driven journey to grow the business with data. Data-driven bank vision means:

(i) A data-driven corporation, i.e., consistent, live, data-driven processes and fast decisions and operations
(ii) Leveraging scale in data processing and reusing architectures, components, tools, and experiences at scale
(iii) New skills and data-aware talent, as data skills are a key asset for their talent, to find new insights
(iv) Efficiencies and cost savings, migrating systems, and reducing total costs (carve-out, migration, sunset, [4] decommissioning, etc.)
(v) Business growth on data insights, i.e., the use of data for growth, making business simple, personal, fair, and fast

Data-driven bank vision simplifies data flow to value moving from fragmented technology, data, and teams to the enablement of a fluid data flow to insights and

[4] To expire (or run out, shut down, terminate) at its predetermined time. The setting sun symbolizes the completion of a journey. This journey could be an information technology (IT) system itself. The twilight of IT components or systems is often compared metaphorically with the setting sun.

value. This means the definition and implementation of a new data value chain concept, linked to the data life cycle, considering data ingestion (sources, transfer, storage, and landing), DQ (ETL, [5] clean, join, stage, quality, and governance), DA (clustering, predictive and accuracy), data insights (360°, risks, churn), and data value.

Fit-for-purpose DM&G in banks requires CDOs accountable for DM&G function, supporting the digital transformation, participating in transformation projects, and ensuring customer and business orientation for data.

They must define and develop the banks' global DM&G strategy, working together with all stakeholders and subsidiaries in:

(i) Gathering inputs from subsidiaries to ensure compliance with local regulatory requirements and overall banks' risk appetite
(ii) Securing the approval for the global DM&G strategy, including necessary adjustments to the banks' data framework, policies, procedures, and standards at the relevant governing bodies
(iii) Also, managing the data value chain globally, ensuring DM&G and control

DM&G strategic vision must aim to cope with the four main requests currently being faced by the vast majority of banks:

(i) Senior management requesting to increase the data scope under DM&G, moving forward faster, setting a clear data accountability in the business areas, and showing the achieved level of progress
(ii) Data owners and data producers demanding to ease their DM&G duties so they can focus on their business
(iii) Data consumers claiming to move from an only reporting-focused DM&G to a data-driven one, leveraged in DA, ML, and AI, aimed to improve reporting and decision-making to get business value via additional revenues and/or cost savings
(iv) Business requiring improving data sharing by reducing data ingestion timings, enhanced data accessibility, shortened process of making data available, and creation of business added value (speeding value)

The answer to these main requests to be a data-driven bank is fourfold: data stewardship, Single Data Marketplace ecosystem (SDM), DM&G dashboard, and Data as a Service (DaaS).

[5] Extract, transform, and load. In computing, extract, transform, load (ETL) is the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s). The ETL process became a popular concept in the 1970s and is often used in data warehousing. Data extraction involves extracting data from homogeneous or heterogeneous sources; data transformation processes data by data cleansing and transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, data lake, or a data warehouse (https://en.wikipedia.org/wiki/Extract,_transform,_load).

8.3 Data Stewardship

The creation of a data steward role with a dependent team in each of the data domains, to guarantee execution capacity, is a key cornerstone to improve accountability on data in business areas.

The objective is to strengthen the CDO role, ensuring connection with the business, with resources (data stewards and budget) and accountability to remediate data issues.

A data steward is a subject matter expert in a given data domain, with the best knowledge of the data and their uses. The data steward is identified by business, and, within this scope, he or she develops and implements granular data actions and road maps for strategic initiatives together with the data owners. Data steward also monitors their progress, ensuring execution and escalating risks. In any case, the data owners retain accountability for the data they own, even if tasks and functions have been delegated to a data steward.

To ensure engagement and accountability, data stewards would have to co-report to the CDOs additionally to each of their business heads. Data ownership remains in the business, but now accountability can be established to the data steward–CDO pair.

Data stewards and their team ensure execution and have an end-to-end view of data initiatives in their data domains. They drive divisional level accountability and ensure responsibilities are embedded in the first line of defense, defining with the CDO a data management strategic plan with medium-term goals. They also lead execution and remediation plans in collaboration with data owners and CDO. Their main tasks are grouped in two blocks:

- Data management strategic plan:
  - Identify their data domains and the area priorities/data across business, set deadlines, and establish specific objectives.
  - Coordinate and raise DM&G actions within their business for critical data as driving key data element (KDE) identification, DQ and controls, data flows (lineage), and the use of data across the information life cycle.
  - Enable accountability within their data domains to identify data risks; coordinate on the data aspects of project/change initiatives and third-party relationship management (suppliers/data services vendors).
- Business as usual (BAU):
  - Measure DQ and remediation needs; lead their teams to ensure execution and have a holistic view of data issues in their data domains being requested by their domains or by others.
  - Assure fixing of DQ issues and coordinate with other areas their resolution.

An initial prioritization of banks' areas must be performed. Usually, initially priority focus is on finance, accounting and management control, risk and compliance, responsible banking, ESG climate, green finance, human resources,

technology and operations, wealth management and insurance, cards, recovery and resolution, and digital marketing data domains.

8.4 Single Data Marketplace Ecosystem (SDM)

SDM is a best in class solution to implement DM&G in data lakes/repositories/platforms in banking. It is based on Amazon's and Microsoft's DM&G models, leaders in the data industry.

SDM allows old/siloed data sources decommissioning, improving efficiency, simplifying DM&G, saving costs, and generating new leads and business opportunities.

The strategy must be to expand it based on deploying business cases, identified, and agreed as global goals with business, and not in a whole implementation of data lakes.

Functionalities defined by CDOs cover main DM&G aspects: data definition and classification, accessibility, security, availability, traceability, lineage, and quality, among others. Technological aspects and solutions are defined by chief technology officer (CTO), following CDO requirements.

Tone of the top, sponsorship and empowerment on this initiative must allow banks to go beyond the road to become a data-driven bank, exploiting data capacities to the next level.

It is an innovative approach based on a new data sharing experience. SDM means to evolve and simplify data-related roles, accountabilities, and responsibilities to three key concepts around SDM:

1. **Data producers:** They integrate datasets through automated and simplified ingestion processes. They provide automated DQ checks and certification.
2. **Data contracts/data sharing agreements:** These are mechanism that connects data consumers with data producers, regulating the data sharing.
3. **Data consumers:** Through a metadata search engine, they quickly find datasets they look for, subscribe, and access data.

The ecosystem allows to move from siloed data to the SDM per bank. It enhances automation and scalability of DM&G model, automating participant access (profiling and monitoring). Also, it integrates a simplified and overall DM&G supporting tool.

SDM allows the CDOs to foster, guard, and enforce end-to-end DQ in the data value chain (collect, store, discover, subscribe, deliver, and analyze).

What is involved in moving to the SDM?

- Data availability, i.e., the identification of where the required data are (data sources) and making them available to the repositories.
- SDM is implemented over the identified sources:

- Data from different sources is modeled to improve their consumption.
- Metadata is added, both technical and functional, including the sources, definitions, and ownership. Also, the control model is implemented.
- Consumers can subscribe to the information through the data contracts/data sharing agreements.
- Data are made available to be consumed on the same platform or in other applications in order to exploit and analyze them and apply business intelligence (BI), models, etc.

Metadata, data in context, is one of the key elements within SDM. It helps to detect data, understands data relationships, tracks data, and assesses the value and risks associated with their use. Metadata also helps with the identification and remediation of the "data sicknesses":

- Initial stages: Not all data needed is available and/or processing is well defined (availability metadata).
- Fostering accessibility: Existing data is not accessible for the required purpose (global, governance and access and uses metadata).
- Improving DQ: Data are available and accessible but lack quality/consistency (quality, traceability, and IT metadata).
- Better decision-making: Quality data are available, but not embedded in business decisions yet (security and social metadata).

In order to properly implement and assess the main aspects involved in the SDM, most of the banks have acquired DM&G tools as Microsoft Azure Purview, Informatica, Anjana Data, Ab Initio, or Stratio, but some have decided to develop them in-house.

8.5 DM&G Dashboard

Once a new DM&G strategy is approved, and started its development, with the aim to become a data-driven bank, creating a culture of innovation that positions data and analytics at the core of business strategy, a fit-for-purpose dashboard must be developed to provide an overview and a forecast of the progress of DM&G extension, as one of the priorities of the strategy.

This dashboard aims to reinforce the monitoring of DM&G overall activity in a quarterly basis, ensuring a robust control over the strategic ambitions by the data governing bodies and relevant stakeholders, including:

- An overview of data under DM&G in BAU basis with classical DM&G's key performance indicators (KPIs) on business glossary, DQ, DISA, and DQ models
- Progress on DM&G extension efforts through four axes: (i) data projects (including strategic ones), (ii) consolidated vision by data stewards' initiatives, leveraging on data management strategic plan, (iii) DQ models' initiatives, and (iv) data

lakes management and governance status, including information about data consumption
- A forecast view with annual and long-term target follow-up
- Also, the dashboard that must include a link to a data value dashboard, showing all the granular detail of the data consumption in the data lakes

8.5.1 Overview

The dashboard must show an overview of the KDEs being managed in all BAU aspects (business glossary, DQ, DISA, and DQ models) and the different DM&G extension axes (data projects, data steward initiatives, DQ models' initiatives, and data lakes' status, including information about data consumption) through which data under DM&G are being worked and will gradually increase the BAU.

Regarding BAU aspects, different KPIs must be included related to:

- Business glossary, including the information related to the data dictionary and reports library. This repository will include all the required attributes in the regulation for each data or report.
- DQ assessment along the end-to-end data life cycle, allowing to measure the quality of those critical data, approved in data governing bodies, used in some reports at aggregated level (group, unit, data ontology, and KDEs).
- DISA, an exercise that certifies all those critical data, approved in data governing bodies, and the different systems and processes that are part of their generation up to the final report.
- DQ control models identified by the CDOs. Each system must have a control model guaranteeing the DQ along the end-to-end life cycle (input, processing, and output) of the systems. It relates to the below mentioned DQ models' initiative included in DM&G extension axes.

Regarding the four DM&G extension axes, different KPIs can be included related to:

- Data projects in development phase, managed through the data dictionary and DQI inventory, to be included in BAU according to the defined road map and their target date. The dashboard must include information about the data perimeter (KDEs), data attributes, applicable DQ controls, business areas involved, or the affected reports.
- Information, which must be shown, related to the different strategic data projects and sub-projects showing them according to the affected business areas, the DM&G requirements applying in each case (data flows, business glossary, metadata, DQ), the targeted date (projects' end date), and the number of associated KDEs and DQIs. It must also include information related to the distribution of strategic data projects and sub-projects based on the source systems.

- Data steward initiatives under defined data stewards' scope, leveraging in the data management strategic plan.
- DQ model identification in order to standardize and include them under DM&G in BAU. It relates to the abovementioned DQ control models' initiative included in BAU aspects.
- Data lakes management and governance status, including information about data consumption. It refers to data lakes managed and governed in BAU based on the defined metadata model for the metadata. It must show technical data managed in the data lakes according to the governance standards:

  (i) Organization of projects by business area
  (ii) Number of technical data by business area and their increase compared with the last execution
  (iii) A distribution of information according to each source system and business area

Work must be done to link the technical with the functional data. Once the critical data applicable to each project have been identified, they must be analyzed to be included in BAU.

8.5.2 Forecast

The forecast must show a projection of the evolution related to KDEs and DQ indicators (DQIs) managed both in BAU and data project basis.

Two major KPIs must be defined to follow up the progress so far to achieve the target of new data to be included under DM&G:

1. Yearly driven data shows the progress so far to achieve the annual target (annual goal of new data to be included under DM&G). Also, it must show the estimated year-end value. It is calculated with the data that are under DM&G in BAU and the data of the projects to be included under DM&G along the year according to the defined road map. Additionally, the historical driven data values must be represented for last executions.
2. Global driven data shows the progress so far to achieve the 3 years target (next years' goal of new data to be included under DM&G and the estimated value). It is calculated with the data under DM&G in BAU and the data of the projects to be included in BAU in the next years. Additionally, the estimated global driven data must show the future goal of this KPI considering the data of projects that will be included in BAU each year.

8.5.3 Data Value

Finally, a data value dashboard, detailing on the data consumption, must be included. It must show a bunch of KPIs related to:

Access:

- Exploitation:
  - Data usage: percentage of real consumers over the potential consumers. It is aimed to know the use of the data in the data lakes, allowing us to know the percentage of users who consume the data:

    Real consumers: those users who consume the data constantly over time or on recent dates.
    Potential consumers: those users who have accessed to the data, regardless of whether they consume or not.
    Relevant consumers: number of critical consumers of each system. It is calculated as a percentage of consumers identified as relevant over the total consumers.
  - Access time: speeding value, i.e., the period of time to make data available. End-to-end calculation of the time it takes from making the data available on the data lakes/platforms until the data consumer can access the information. It is calculated as the sum of the average of ingestion times plus the average to perform a data contract per data.

Data:

- Data strategy:
  - Degree of coverage: number of strategy lines for each data lake/platform. It is calculated as the percentage of strategic lines governed by data lake over total strategic lines in the strategic DM&G plan.
  - Core data: percentage of critical information for each system. It is calculated as the percentage of core data over the total data lake data.
- Data availability:
  - Historical depth: depth of data consumed. It is calculated as the percentage of the sum of consumed historical depth over the sum of available historical depth.
  - Reputed data: data with a minimum quality criteria. It is calculated as the percentage of the data from reputable golden sources over the total data lake data.
- DQ: degree of literacy, i.e., how the DM&G is. It is necessary to know which governance we have applied on each managed and governed data by each data lake/platform. It is calculated based on the weighted completeness of the critical metadata of each data and the weighted DQ assessment. The objective is to be

able to measure the robustness of a data, obtaining a metric allowing us to quantify how good a data is. A data will be considered governed when the degree of literacy is equal to or greater than 60 percent in strategic projects or 45 percent in nonstrategic projects.

- Data usefulness:
  - Reused data: data consumed by third parties. Once the data that have been made available for consumption are known, we must be able to quantify how many of them are being consumed by third parties without considering their data producers. The objective is to avoid working in silos, making it easier to reuse existing information, avoiding duplication of information. It is calculated as a percentage of data consumed by third parties over the total data made available.
  - Data heat: i.e., data temperature. Based on the information that is available in a data lake for consumption, it is necessary to be able to quantify it in such a way that it allows us to know how hot the data is. The objective is to become aware of how hot data is, based on the consumption made in the data lake, distinguishing by users (nominal/machine) who access it, quantifying consumption and what depth of information is being consumed, getting a ranking of data heat by business area, source system, etc.

    We will first obtain the standard data heat as the sum of the weighted number of consumptions for each type of consumption. Then we will obtain the normalized data heat calculated as the percentage between the standard data heat and the maximum value of the dataset with which we work by granularity.

- Datability: development of a data inner value (DIV) score that has a translation to a data inner monetary value (DIMV) leading to the prioritization of use cases.
  - Data inner value (DIV) score: It must be obtained at the most granular level, i.e., data level, being then aggregated until we get at a final score for the complete dataset under analysis (initiative, system, project, data lake, etc.). It is calculated based on DQ, usability, and relevance:

$$\text{DIV score} = \text{Average}\left((\text{DQ}), (\text{Number of use cases} \times \text{Utility of uses}), \text{Relevance}\right)$$

  - Data inner monetary value (DIMV) score: The value added specifically by DM&G functions to the income statement, measured on a use case basis, considering both gross income generated and costs required to do so:

$$\text{Gross Income} = \Delta\ units^{(1)} \times average\ margin\ per\ unit.$$

Δ units = sold units applying DM&G treatments (final scenario) vs. *sold units without DM&G treatment (initial scenario).*

Average margin per unit = quarterly margin per type of product/asset (1) units: refers to the number of sold units depending on the use case (cards, loans, insurance, etc.*).*

$$\text{Costs} = (\text{Cost final scenario} + \text{transformation cost}) - \text{Cost initial scenario}$$

Cost initial scenario = (IT costs + operations costs + DM&G costs) without data treatment.

Cost final scenario = (IT costs + operations costs + DM&G costs) with data treatment.

Transformation cost = any cost related to the transformation process when evolving from the initial scenario to the final one. Some examples could be the cost of technical development or the consulting cost.

8.6 Data as a Service (DaaS)

For a long time, different factors have enabled the existence of many data silos within the banks, which must be deactivated following the next steps:

(i) Intervening in the main information circuits, in order to set an active DM&G over the main data circuits of the banks
(ii) Creating the data offer within the data lake environments, reducing their redundancy
(iii) Aligning IT plan with the data strategy (MIS, CRM, payments, etc.)
(iv) Migrating data users from silos to data lake environments
(v) Shutting down the infrastructure that supports these silos, getting savings in technological infrastructure

Current advances in IT infrastructure and SDM must allow banks to move to the next level, allowing them to streamline key capabilities as data democratization, data ecosystem, self-service, advanced data analytics, promotion on the use of AI and ML models, data volume and complexity, distributed processing, hybrid deployment, expert users and advanced use cases, customized marketing, and recommender systems.

And does this only affect the data flows and the IT infrastructure? The creation of data silos goes beyond infrastructure. There are multiple departmental servers and teams (Retail Banking, Wholesale Banking, Management Control, Risk, Customer Quality & Experience, Business Banking, Human Resources, etc.) processing the same data, working independently, with different focus, without a bank vision, and creating duplicated processes in a product of data silos in high percentage.

Linked to a fit-for-purpose SDM, Data as a Service (DaaS) model enables simpler data processes, guaranteed DQ, and data process reengineering. DA, BI, and business analytics teams do not process data, as these are processed by DaaS.

What is needed to achieve it?

- IT infrastructure working perfectly, without failures and very stable.
- High percentage of progress in eliminating data silos and migrating user connections to SDM.
- After migration to SDM, meet the three "commandments" of a satisfied data user: early and quality data and luxury exploitation experience.
- Just after achieving the above, we can advance to the next level:
- Get sponsorship from human resources and structures and productivity/organization areas.
- Work with them on the business case that makes DaaS viable, even if the investment is very low or zero.
- Present the project, with a bank vision, to the areas having data processing departments. Also agree on the transfer of resources associated with DaaS.
- Design the DaaS implementation plan ensuring the continuity of current data services, analyzing current processes and searching for "Quick Wins," and developing a Data Process Reengineering Plan converging toward a "Modern Data Platform," whose services are fully automated, with the best practices in data models and tools.

8.7 The Magic Algorithm

Banks know that the challenge they are facing is not easy, because it changes the way they work and displays reality. As Copernicus and Galileo, banks must change the way of seeing the world and the way they make their observations. Banks must explore and seek other "Suns" to enlighten business, customers, and regulators and other planets to live on. Banks must order the stars, "the data," to guide themselves on their way.

Customers are the main focus of attention for banks. They are looking for uniqueness and their interactions are increasing on digital channels (multichannel experience). DM&G is key to improve banks' services and build business, customer, and regulators' trust.

Best way to build confidence is through transparency (openness, saying what we know) and integrity (consistency in action, doing only what we say). Banks' commitment to data transparency and integrity with business, customers, and the regulators is critical.

$$\text{The magic algorithm} : \text{Transparency} \ + \text{Integrity} = \text{Trust}$$

Chapter 9
Data Has the Power to Transform Society

Carlos Alonso Peña, Alberto Palomo Lozano, and Javier Esteve Pradera

9.1 Introduction

The emerging European data economy is set to generate an exceptional growth opportunity for industries in the Member States. At stake is an economic market that, according to estimates, will have grown by 300% by 2025, reaching 830 billion euros for the 27 Member States (around 6% of GDP [1]), primarily related to the growth of the Internet of Things and the massive amount of data that these devices are capable of generating. However, this commitment to prosperity and innovation must be harnessed without generating inequalities, without compromising the industrial and technological future of the Union, and without jeopardizing the fundamental values and rights of citizens.

Beyond the more market-oriented view, data is signified by its transformative potential for society. Data can be implemented and governed for public benefit as a resource to address environmental, social, and health challenges from an innovative perspective, enabling collaboration, driving innovation, and improving accountability.

Data, and its essential role in the development of disruptive technologies, such as artificial intelligence, is the differentiating factor of an industrial and technological revolution that will allow us to consolidate a digital economy that is fairer and more inclusive and align with the United Nations Sustainable Development Goals and the 2030 Agenda. Consequently, data is a vital component of any advanced data economy to ensure the development of two crucial and strategic processes such as

C. A. Peña (✉) · A. P. Lozano · J. E. Pradera
State Secretariat for Digitalization and Artificial Intelligence, Ministry of Economic Affairs and Digital Transformation, Madrid, Spain
e-mail: carlos.alonso@economia.gob.es; alberto.palomo@economia.gob.es; javier.esteve@economia.gob.es

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_9

digital transformation and ecological transition. This reality is especially the case in Spain.

The Spanish government is actively working to create a legal, political, and funding environment for the deployment and implementation of the data economy through the various initiatives detailed in the Digital Spain 2026 strategy and deployed in the National Artificial Intelligence Strategy, the Connectivity and Digital Infrastructure Plan, and the Strategy for the Promotion of 5G Technology. These priorities are part of the Recovery, Transformation and Resilience Plan that will leverage NextGenEU funds to drive them forward. In the public sector, the Public Administration Digitalization Plan is aligned with European initiatives and regulations to promote the data economy, and it aims at increasing the effectiveness and efficiency of the administration, thereby laying the foundations for an innovative public administration.

The Data Office of the Government of Spain has a facilitator role, focused on the strategic and conceptual development of data and information infrastructures based on easily transferable methodologies across different sectors. The Office was formally constituted in mid-2020 (Creation Order ETD/803/2020), framed in the State Secretariat for Digitalization and Artificial Intelligence within the Ministry of Economic Affairs and Digital Transformation. The Data Office combines its external vision of promoting and accompanying industrial sectors with its inner vision of reinforcing the digital transformation of the administration permanently to preserve strategic digital autonomy.

Following this duality in the Office, this chapter addresses two distinct but ultimately intertwined topics. On the one hand, it sets out the concepts and constraints underpinning federated data governance as a critical element in achieving strategic digital autonomy. On the other hand, the chapter details the principles that should govern a data-oriented administration to unlock the potential of data as internal and external transformative power.

9.2 Federated Data Governance as a Pillar of Strategic Digital Autonomy

In this context, the European Commission published a timely and ambitious European Data Strategy [1] in February 2020. This political positioning pivots on two disciplines currently in vogue: data and cloud services (or, more generally, the "as-a-service" model offered by the public cloud). Under the framework of a magnanimous road map, with significant investments deployed along the axes of innovative digital infrastructures, digital skills and rights, and regulation and standardization in the use of data for the digital transformation of businesses and public

[1] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_es

services, the strategy aims not only to develop new capabilities to empower European societies and economies based on these two disciplines but also to connect them.

It is noteworthy to capitalize on the interrelationship between infrastructures and their use. We have noticed these synergies recently in fields such as artificial intelligence, whose current momentum has been triggered by a coinciding conjunction between vast data sets' availability and disruptive parallel computing capabilities, a scenario where Moore's Law[2] intersects with Metcalfe's Law.[3] Alternatively, more generically, the focus must be pointed to the rise of technology-based business giants, which use their digital platforms for marketing third-party items beyond providing products or services. This model of "commoditization," so widely used today, has an obvious translation into the data domain. Moreover, let us consider the non-rival status of data. It can be copied and stored at an increasingly low cost and exploited in different contexts without negatively affecting the original owner of its rights.

This is why the strategy seeks to consolidate and promote the Digital Single Market, a *leitmotiv* underlying the very foundation of the European Union. Similarly, to steel and coal from 1951, the EU seeks to generate a distributed market for industrial data where counterparties execute *point-to-point* data transactions serving as an instrument to digitize the different value chains. In this context, data would not only represent the by-product resulting from the interaction of digital applications, useful in audits and process debugging, but also a raw material that can be reused in multiple ways, generating added value even at a cross-sectoral level.

9.2.1 From the Platform Model to the Ecosystem Model

This scenario bears remarkable similarities to current platform models since it leverages the same key ingredients: a dynamic market of supply and demand, in this case for data sets and services, and an instrument connecting participants and mediating the transaction. However, a novel aspect of this strategy is the search for collectivization in creating value. Unlike platform models, where a large part of the value generated is retained in the intermediation process, the EU is committed to a data ecosystem model that, in line with European principles, avoids the generation of dependencies and dominant positions in the markets (see Fig. 9.1).

This ecosystem model aspires to federalism so that minimum joint governance allows for flexible participant interaction. At the same time, the participants in the ecosystem model still retain an autonomy that allows their unilateral participation in such point-to-point transactions, depending on the conditions of the moment. This

[2] https://www.intel.es/content/www/es/es/newsroom/opinion/moore-law-now-and-in-the-future.html

[3] https://blogs.ua.es/airc/2007/10/25/la-ley-de-metcalfe/

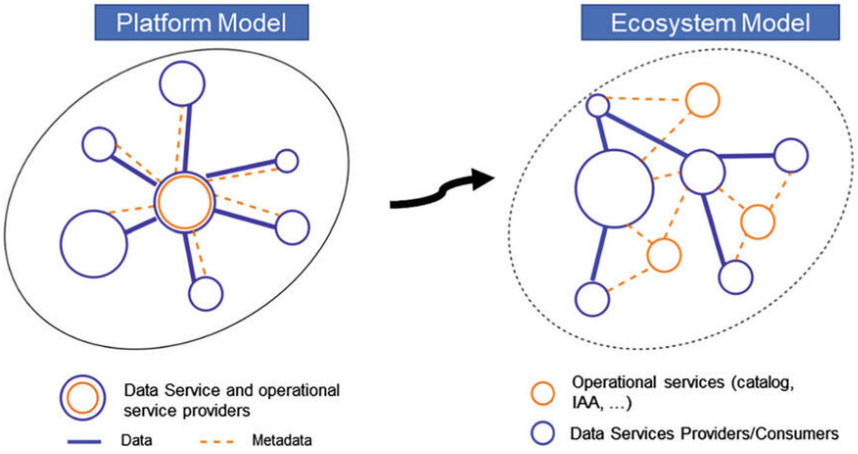

Fig. 9.1 Platform model vs. ecosystem model

ecosystem model creates significant challenges concerning interoperability, as opposed to the platform paradigm. In the platform model, all participants use the exact vehicle, generally in the form of a software application, where semantic and technological interoperability exists by design. Consequently, participants in a federated ecosystem model must agree upon a set of standards and codes of good practice to facilitate such interconnection between distributed ICT systems. However, even beyond the purely technological part, common standards are sought to facilitate interconnection also at the business, legal, and organizational levels, thus ensuring flexibility and easy extensibility for the development of business processes a posteriori, e.g., to stimulate the effervescence of the data market and the resulting uses.

Furthermore, it is precisely for this reason that the European Data Strategy seeks strong coordination between *cloud capabilities* (i.e., the rapid deployment of infrastructure and platform services to match the moment's needs) and the domain of data management, transformation, and exploitation. Without such interactive coherence, the most likely scenario would be one where data fails to break out of its original silos, thus limiting its generous transformative capabilities or giving rise to the creation of large interest groups with a dominant position over the rest.

9.2.2 Features of Federated Data Ecosystems

Specific features of these federated data spaces include the following:

- There must be a *governance model* based on enforceable interoperability rules that guarantee the development of the data-sharing business on an equal footing by the ecosystem's service providers and consumers. The governance rules will guarantee the nonexistence of entry and exit barriers to the ecosystem beyond the guarantee of interoperability and security in the development of the business.
  - A data space operator will be in charge of the technical and operational tasks necessary to run the system (authentication and access identity, support, maintenance, logging, de-registration, and system supervision, to name a few). However, this post may not provide data-sharing services of its own (neither data provision nor data processing). The provision of services will be the responsibility of the supply-side participants in the data space: data service providers or data processing providers.
  - An intermediary service provider will be responsible for offering value-added services that facilitate data sharing. For example, service catalogs, activity logging auditing, or application stores for data processing can be mentioned. The provision of such services will be open to interested parties who wish to provide them as long as they do not meet disqualifying conditions.
  - Decision-making in the data space aspires to be participatory, both in technological and business matters, so that no dominant operator can make unilateral decisions on the characteristics and evolution of the data space. This fact ensures that the generation of innovation and value of the system is sustainable, based on the promotion of the participation of different types of stakeholders.
  - Interoperability rules will allow access to value-added services and the provision and consumption of data-sharing services with security and confidence, but always in an autonomous manner, by data providers and consumers, as part of an agnostic environment that does not offer an advantage to any participant in the data space (this is the principle of "competition on equal terms").
- *Architecture*. The technological architecture of the proposed solution will follow the decentralized federation-based model, in which there is no requirement for centralized components in the provision and consumption of data-sharing services. However, services are provided and consumed directly in peer-to-peer relationships between consumers and providers. The only exception to this rule is for identity and trust services, although decentralized mechanisms can be explored even then.
- *Functionalities*. At least the following functionalities shall be offered:
  - Secure data exchange between participants
  - Data models and data formats for data exchange
  - Traceability and lineage of data sets
  - Data sovereignty as the ability to define and enforce policies for access and use of data by access rights holders

  - Logging of data-sharing activity for auditing and reporting purposes
  - Tools for publishing and searching data (i.e., a catalog)

- *Minimum building blocks*. The European Commission defines building blocks as basic digital infrastructures that can be reused to compose complex digital services. The current state-of-the-art technology makes different reference architectures and components at different maturity levels available to stakeholders for creating data spaces that support the aforementioned European values and strategy.

In order to boost the creation of these ecosystems and mitigate dominant positions that lead to technological dependence, the operational and intermediation components in a data space should be made available as open-source software. By way of example, and not exhaustively, we can mention those assigned to the Gaia-X initiative, the International Data Spaces Association, or the FIWARE community, as well as those developed on the basis of programs funded by the European Commission (e.g., the *Connecting Europe Facility* [4]).

9.2.3 The Pillars of Federated Data Ecosystems

By definition, federated data ecosystems are susceptible to network effects, rapidly increasing in value as their participants grow. This goodness of economies of scale is exponential when articulated around a federated model with weak coupling between parties, as it encourages participation. However, their governance is also a significant challenge. We, therefore, propose four main strategic lines of action to promote the creation and operation of these sovereign ecosystems, aiming to make them more dynamic and gain the capillarity of participants:

- Firstly, given the distributed nature of these ecosystems, we believe that the generation of *community* is a critical element. In any sector, common and differential elements are unique to the scope, resulting from its specific business processes and a progressive and iterative generation of consensus and shared knowledge. This is why any action aimed at boosting federated data ecosystems must not only consider the particularities of the community on which it seeks to deploy but also build on the pre-existing frameworks, reusing artifacts, standards, best practices, and codes of conduct widely accepted in that domain. This community should not only be considered but also enhanced.
- Since these ecosystems seek to break the centralization inherent to platform models, they are always built considering the distributed nature of their members. This makes *transparency and trust* take on a predominant role, as they cannot be assumed by default. Therefore, any governance model to be adopted must

[4] https://wayback.archive-it.org/12090/20221222151902/https:/ec.europa.eu/inea/en/connecting-europe-facility

consider the design of policies and mechanisms that enact these factors, including the correct identification and accreditation of participants and services offered and demanded.

- As already mentioned, these ecosystems appeared to break out of the natural silos in which data was collected and exploited mainly. Considering the data spaces as a productive input, there are numerous meanings under which the same set of data can be considered and exploited by interrelating it with others. This has led several analysts to refer to data as "the oil of the twenty-first century," given its enormous plasticity and transformative capacity in different contexts, and that is why it plays a role in the *innovation of products and services*. It is, therefore, a priority perspective from which to approach the development of federated data ecosystems, which must be able to articulate a differential and novel value proposition based on the high scalability of the proposed model.
- Similarly, we also believe that the use of novel ideation methodologies of concept and the ability to pilot and deploy rapid proofs of concept are instrumental in developing these federated ecosystems, where, by design, there is no predominant system broker to rely on.
- Finally, for an ecosystem to enjoy practicality and continuity, it must evolve from the testing phase to an operational reality where it generates quantifiable business value, i.e., ensure its *scale-up*. This undoubtedly involves deploying processes with guarantees of sustainability, developed on business assumptions and considerations that imply a shared benefit, and whose organizational and legal foundations are solid enough to achieve the desired positive economic and social impact in the medium and long term.

9.2.4 Shared Common Infrastructure

These four pillars, the foundations on which to deploy the formation of federated data ecosystems, are domains that have been widely discussed and analyzed before. In this case, the ability to combine them mutually is genuinely novel. While in centralized or platform environments, the conversation usually revolves around data-driven innovation and its capability of scalability (taking for granted the availability and efficiency of the underlying infrastructures and resources), in the context of federated environments, these two domains must also flourish in collusion with the adequate management of resources from different origins, systems, and owners, and whose reuse raises questions about interconnection, and the identity and trust in them.

Due to this, the transparent orchestration of interoperability between participants and data resources is central to federated ecosystems' digital value chain. It also seeks complete coverage throughout transformative and data exploitation processes, ensuring no *single points of failure* or bottlenecks penalize the optimal deployment of business processes at the technical, legal, and business levels. Therefore, although it is not reasonable to suddenly disinvest from models and tools already adopted and

integrated as part of these processes, the key lies in the generation of an innovative and transversal capacity for interconnecting resources and processes under a federated approach, respecting and encouraging the self-determination of the intervening agents while encouraging their participation.

Just as the Internet emerged to become operationally resilient through a distributed communications model, creating a "common shared infrastructure" [2] layer allows the desired transparent, reliable, and efficient orchestration to be deployed between different combinations of potential participants in federated data ecosystems. Moreover, this orchestration is not only done vertically around a specific domain (as may be the case for already available monolithic sectoral cloud offerings) but based on a virtual and decentralized interconnection between the supply and demand of services from different providers. This way leads to collectivizing value creation among different stakeholders with heterogeneous characteristics, which thus become smoothly and sovereignly coupled.

This model, which can be assimilated into the transversal capacities of a network of fundamental infrastructures of territory (i.e., electricity, water, sanitation), seeks to generate favorable conditions for the development of the desired single market for data on a European scale, providing a global vision to generate network economies and reduce barriers for small- and medium-sized participants while boosting the innovation and resilience capacities of the industries within the Union. However, far from having a physical representation ("hard infrastructure") exclusively, for example, in the form of laboratories, development environments, specific applications, or "run-time environments," the model also adopts softer characterizations in the form of standards and conformity mechanisms, standard reusable software pieces [5] or specific pilots and applications for the various domains. Intangible assets can also be considered, the coordination of ecosystems and the dynamization and incubation of communities and their participants, as well as boosting the reuse of open data held by public administrations, whose value for product and service innovation has been demonstrated.

Therefore, all this common shared infrastructure seeks to accommodate both along the business dimension, based on the analysis of economic models, and along the business dimension, based on the analysis of economic models. This deployment is based on the analysis of economic models and the promotion of cooperation and collaborative innovation considering several dimensions: (1) the legal dimension, offering answers to the contractual and regulatory considerations and needs of the ecosystem participants, and (2) the functional and operational dimension, including (2a) catalogs of resources available under a federated scheme, (2b) the promotion of ecosystem liquidity (to generate a wide range of services to make them more flexible and stimulate their exploitation), or (2c) the characterization of roles and best practices to be exercised, as well as the training and deployment of support communities that treasure and advance shared common knowledge.

[5] Available monolithically in the form of open source code, or even packaged around common functionalities or sectoral requirements

In summary, although the mutualization of these developments and artifacts is not the only piece necessary for the creation of federated data ecosystems, it will undoubtedly serve as a basis for promoting the incipient and ambitious European Data Economy, whose development does not neglect the "capacity of the territory to provide and control [6] those technologies and tools critical for digitization, and therefore for growth, competitiveness, and welfare" [3], i.e., enhancing strategic digital autonomy.

9.3 Data Governance in Public Administrations as a Guarantor of the Generation of Citizen Value

Strategic digital autonomy is also desirable for public digital systems, being applicable to the concepts previously expressed when formulating its data governance, governance of a data-oriented administration guaranteeing the generation of real value for the citizen.

We may think of public administrations as large data banks, combining data generated by citizen service interactions and their relations with companies. As a result of the digitalization process in which public administrations are immersed, their procedures and processes must be reconsidered and reoriented to be more agile, transparent, and responsive. Citizens expect the digital services deployed by the different administrations to be easily accessible, facilitating greater participation and transparency of political processes. Thus, it is impossible to think of an effective digital administration without good data management, and there is hardly any data to manage without deep digitization of the administration.

Data, understood as a public resource, is a critical element of the digital transformation process of public administrations and plays a relevant role in the design of any innovation policy, redefining its relationship with citizens and the different productive sectors, always seeking to enhance the common wealth of society and promote a fair and inclusive economy.

The objective is to achieve a citizen-centered, open, transparent, inclusive, participatory, and egalitarian administration. For this doing, the administration should be data-oriented, ensuring ethical, safe, and responsible use of data, with an improved capacity for objective decision-making through measuring the results produced by its policies. This administration will leave no one behind.

The Spanish administration is diverse in size, competencies, and maturity level regarding the use of data in its different organizations. The most common situation is that the most prominent departments and organizations have begun their journey toward a data-oriented organization, establishing data governance, data management, and data quality management structures. At the same time, the pace of

[6] "Either through the generation of these technologies itself, or by guaranteeing their supply from other territories without this implying unilateral dependency relations"

incorporation is slower in smaller organizations, which, in many cases, are still focused on providing an operational response to day-to-day technological needs without a strategic vision of the possibilities that data could bring them.

The objective should be to maximize the value of data, generating value beyond the system that creates it, breaking data silos within and between organizations, and adding value to the business strategy. Data is not exclusively a matter of ICT interest. Data potential feeds all business areas.

In further developing this last competency, the Spanish Data Office has established the values and design principles to continue advancing the construction of a data-oriented administration capable of taking advantage of the potential of data through the use of innovative technological means, enabling the design, execution, and evaluation of citizen-centered public policies that promote a data-oriented economy that is sustainable and generates social value.

The conception of the principles and strategic lines of a data-driven administration, beyond considerations of efficiency and effectiveness, must take into account the values of objectivity, cooperation, participation, proximity, integrity, transparency, social responsibility, equity, and sustainability, all considered within a culture of pursuit of excellence in a general framework of evaluation and continuous improvement.

This innovative data-oriented administration, capable of taking advantage of the potential of data using innovative technological means, should be established around the principles of effective data governance, ethical treatment of data, reliable data-centered processing, sovereign data sharing, open dissemination of information, evidence-based design and analysis of public policies, and promotion of data culture.

Let us look in detail at the principles laying the foundations to achieve a data-oriented administration.

9.3.1 Principle of Effective Data Governance

A data-oriented administration needs to address data governance, that is, defining who can take what actions, with what data and when, in what situations, and using what methods, and allows maximizing the value of data in support of the organizational strategy. This maximization involves establishing corporate data governance that avoids information silos that lead to inefficiencies, duplication, and stagnation in deploying the potential value of data.

This corporate data governance should be based on a federated approach, following the characteristics described above for the data ecosystems, leaving a sufficient degree of autonomy and responsibility to the different agencies, with the Data Office as the backbone, ultimately enabling the fluid exchange of data and the interoperability of the systems. The role of the person responsible for the data in each agency is fundamental, acting in coordination with the Data Office. The participants' familiarity with their business areas, the vision, the requirements they bring, and the knowledge of their technological systems are crucial elements in the

definition, development, and operation of any data governance and management initiative.

Data governance implies generating policies, standards, and procedures for data management and exploitation; it also implies sharing data between each organization and the Data Office in achieving its federating and enabling mission.

As part of these efforts, the Spanish Data Office has sponsored, promoted, and participated in the generation of technical guidelines by the Spanish standardization body (UNE) regarding the proper governance of data (UNE 0077:2023), data management (UNE 0078:2023), and data quality management (UNE 0079:2023), with which to provide a reference data management framework for both public and private organizations. The availability of well-managed, quality-proven data is essential for progress in data sharing, exploitation, and enhancement.

These national technical guides establish a set of standard processes applicable to the data assets of any organization throughout their life cycle, maximizing their value by applying a structured, managed, consistent, and standardized approach to all data-related activities, operations, and services. It must be ensured that the definition, creation, storage, maintenance, access, and use of data (which implies the need for data management) are done following a data strategy aligned with organizational strategies (this implies the need for data governance) and that the data sets to be used are suitable for the intended use (this implies the need for quality management).

Controls and evaluation procedures, endorsed at the highest level, must be implemented to ensure continuous compliance with data governance policies, establishing the need for a self-assessment model of maturity concerning the data of an organization, to be applied and reported periodically, consolidating the results of the evaluations by the Data Office, allowing to show the degree of progress in the principle of effective data governance. Along these lines, the UNE 0080:2023 technical guide for the evaluation of data governance, management, and quality, based on the ISO/IEC 33000 family of standards and the central IT and data maturity models, makes it possible to evaluate and represent the capacity of the data governance, management, and quality management processes and to obtain the organizational maturity or level of adoption for the three areas.

9.3.2 Principle of Ethical Treatment of Data

A data-driven administration must ethically perform data governance and management. All data practices must be assessed to minimize any adverse impact on people and society. This risk is increased with the use of artificial intelligence technologies. However, the risk is not restricted to this area of knowledge, and it is necessary to delimit, from the ethical point of view of the data treatment, those risks derived from privacy constraints and those that compromise the establishment of fair principles of data management and sharing.

The management and use of data should contribute to the common wealth, minimizing any negative impact, providing equal opportunities to all citizens, ensuring the rights of vulnerable people, complying with the principle of nondiscrimination, and ensuring the proper application of the gender perspective.

The consideration of ESG (Environmental, Social, Governance) criteria must be present in the regular data governance and management decision-making process, enabling the integration of various environmental and social data sources and the appropriate ethical considerations. The availability of well-governed, quality, reliable, mapped, and cataloged ESG data is a first step to consider.

Before implementing automated decisions using algorithms, potential risks to privacy, fairness, and security should be assessed to minimize the likelihood of adverse effects. Methodologies for auditing, monitoring, and verifying executions should accompany any task automation process.

The decisions taken, their justifications, and the results obtained from automated data processing will be communicated in a concise, transparent, intelligible, and easily accessible manner, with clear and straightforward language, avoiding technical terms, so that any citizen can understand them. The traceability of the data sets used in the training and operation of artificial intelligence algorithms will be enabled, as well as their validity, ensuring the absence of biases originating discriminatory results.

High-risk artificial intelligence systems using techniques that involve training models with data shall be developed from training, validation, and test data sets that meet the appropriate quality criteria and are adequately governed and managed. Training, validation, and test data sets shall be relevant, representative, and, to the greatest extent, error-free, complete, and statistically representative of the study's geographic, behavioral, or functional context.

The sustainability of the data treatment shall be ensured, considering the need to meet the principle of not causing significant environmental harm.

9.3.3 Principle of Reliable Data-Centric Processing

The ordinary functioning of the administration must be oriented to the production, exchange, exploitation, dissemination, and enhancement of data, producing a transition in the daily work from the document to the data, always seeking to offer a better service to the citizen. It is impossible to think of an effective digital administration without good data management, and there is hardly any data to manage without deep digitization of the administration.

Electronic administration is a reality in the different Spanish public administrations; it is practically impossible to think of an administrative procedure without considering its deployment in the appropriate information system. The electronic register as the cradle of digital information, the data intermediation platform as a way of not asking citizens for information that already exists in other administrations, the electronic signature, and the prevalence of electronic notification over notification on

paper all constitute a breakthrough in streamlining administrative procedures and simplifying the administrative burden for citizens and economic operators.

However, in many cases, the digitization of administrative procedures in Spain has been conducted by electronically reproducing paper procedures without thoroughly re-engineering these procedures to take full advantage of the new digital medium. As a result of the ease with which documents can be generated and signed (electronic signature holders), the generation of documents has multiplied; in many cases, these documents are created outside the data stored in the corresponding information systems, which creates inconsistencies that can only be later detected. Thus, in many cases, far from taking advantage of the transforming capacity of technology, approaches, practices, and inertia have been consolidated that hinder the deployment of proactive and personalized public services. A data-centered administrative process would overcome most of the problems mentioned above.

Data must be processed efficiently, effectively, and securely, using the appropriate methodologies, technologies, and tools to ensure its non-accidental disclosure, quality, relevance, and accuracy. The search for data quality must be a constant in every organization, playing a pivotal role in the aforementioned UNE 0080 specification for data quality management and the ISO/IEC 25012 and ISO/IEC 25024 standards. Formulating the necessary validations implies the necessary business knowledge and the handling of the necessary tools, all within a general framework of data governance and management.

Data must be processed in a lawful, legitimate, fair, and transparent manner, and the necessary administrative, technical, and physical safeguards must be applied. Explicit and legitimate sanctions may be established for misuse and noncompliance with the established guarantees.

Appropriate privacy and security measures must be considered to ensure their integrity, confidentiality, and availability from the design stage to minimize risks in the event of human errors and technological failures. Consent or the legal basis is the central tool that allows data to be collected, shared, and used fairly, proportionately, accountably, and securely.

The vision of single-use data around the application that generates it, tied to a particular scheme and format, must be overcome so that the data flows where needed for proper decision-making and value generation. Data and applications must be decoupled, facilitating their reuse both internally and by interested third-party ecosystems. The life cycle of data and analytical models must allow for rapid iteration (agile and DevOps approaches) to deploy, optimize, and redeploy new data sets and models.

All high-quality data used as part of any administrative procedure must obey the European "Once Only" Principle, catalyzing an even more intensive use of the data intermediation platform, defining new services, and reaching those assignor and assignee agencies not yet included.

Data-centric administrative processing will enable the use of advanced technologies and tools for descriptive, predictive, and prescriptive analytics (BI, big data, machine learning, deep learning), generative algorithms (LLM, GPT), process

automation (RPA), and advanced information preservation techniques (blockchain), catalyzing new proactive citizen services.

Maximizing the value of using artificial intelligence techniques and tools requires having the necessary quantitative and qualitative data. The primary training data for the algorithms that satisfy and generate the business and service opportunities demanded of the administration at any given moment must be generated naturally. The administration must take advantage of the benefits derived from the economies of scale inherent to its ability to generate much data in its multiple fields of activity.

9.3.4 Principle of Sovereign Sharing of Data

Data is a resource that is not a sole property, and its use does not invalidate but instead favors other additional uses, always respecting the legal framework. Data value grows as its use becomes more widespread (network effect). Sharing data with sovereignty allows the correct design, execution, and evaluation of public policies. However, data sharing must include who can access what data and under what conditions of use, security, and trust concerns.

The public sector data spaces are the place for sharing government data. Data space is an ecosystem where the voluntary sharing of participants' data can occur within an environment of sovereignty, trust, and security, established through integrated governance, organizational, regulatory, and technical mechanisms. Data spaces go beyond the bilateral exchange of information, constituting in their most advanced version authentic business networks where the value of data and its interoperability can occur.

The objective is to project the current methodologies, specifications, and practices on a larger scale, achieving a fluid and continuous data exchange between administrations, economic sectors, and citizens. Considering the very nature of this goal, a much more interdisciplinary and interdepartmental approach and taking advantage of the latest technologies are required. This sharing will generate advantages and opportunities for the different actors involved, always considering the necessary privacy and security considerations.

The data platform was recently created in Spain to promote data-based public management as established within Measure No. 6, "Transparent data management and exchange" of the "Public Administration Digitalization Plan." The data platform is created under the guidelines defined by the Data Office and is implemented by the General Secretariat for Digital Administration (SGAD as per its Spanish acronym).

Public sector data spaces are to be built around the data platform, provided as a standard service to all agencies, taking advantage of their storage capacities, analytical capabilities, and data governance tools and considering the founding principles of European data space building initiatives. Generally, any data sharing or data analytics project should seek to be accommodated within public sector data spaces.

Each public organization will manage its data environments, being able to complete the systems under its responsibility with the functionalities available in

the data platform. In any case, the platform will guarantee each hosted business vertical's independence and specificities and timely publication of the data products. The platform offers controlled access to the specialized personnel of each organization to its business vertical. Analytical results from each business vertical should be easily shared with the proposing agency or other stakeholders. These results may include the necessary data preparations and transformations to meet the needs of a given exchange and become available for future exchanges.

The different agencies will make their data products accessible through the appropriate data services published from the data catalog of the data space. Thus, each agency shall select the relevant data sets for other agencies, proceeding to their creation, establishing their conditions of use, semantic definition, and cataloging. These data sets will be made accessible in a controlled and uniform manner within the corresponding data space. Some of these data will be moved to a central repository or created due to an analytical process, while others will remain accessible from their origins, ensuring uniformity of access and use.

The data space's security must always be present, guaranteeing its compliance with the Spanish National Security Scheme. Data spaces will be combined, aggregated, recomposed, and deployed on common software infrastructures. If such data spaces do not provide the same level of security from the outset, the combined data will always lead to the lowest common denominator for security, weakening its participants' trust. The application of privacy-enabling technologies (PETs) can help to overcome barriers to sharing by solving issues related to privacy or confidential business information, always in strict compliance with the data protection regulatory framework.

The various European interoperability and standardization initiatives (European Interoperability Framework, DCAT-AP, ADMS, Core Vocabulary, CPSV-AP, Once Only Principle, single digital gateway) must be closely followed, ensuring the adoption of those elements required for the practical materialization of public sector data spaces. When approaching the design of an information system, the interoperability of the data managed must be taken into account. If the system is subject to public procurement, this point should be addressed by requiring the appropriate study.

Thus, the data spaces created must be interoperable with those created by other territorial administrations and with the corresponding security measures with the sectoral data spaces of the different industries and the different European initiatives in this respect. Beyond the public sector and considering the European Union's firm commitment to deploying sectoral data spaces, the Data Office coordinates the adaptation, sharing, and exploitation of these new data management paradigms, where the leadership and participation of the different sectoral bodies are fundamental. The Administration, from this innovative attitude, must act from the public sphere as a catalyst for technological innovation in our country. The data treasured by the administrations are a fundamental resource in deploying these sectorial data spaces.

9.3.5 Principle of Open Dissemination of Information

Data can be implemented and governed for public benefit as a resource to address environmental, social, and health challenges, enabling collaboration, driving innovation, and improving accountability. Open data, understood as data that anyone is free to use, modify, and redistribute, with the only limit, if any, being the requirement for attribution of its source or acknowledgment of its authorship, is an integral part of the value of the data economy.

Spain occupies one of the top positions in the European open data maturity index regarding the openness of the policies conducted and their impact, the quality of the data published, and the adequacy of the datos.gob.es portal. The portal *datos.gob.es* includes the catalog of reusable public information, which makes all reusable public sector information accessible at a single point. The catalog has grown over the years to include more than 62,000 data sets. Despite this, there is still room for improvement in data sharing among administrations, industry, and civil society. Administrations should be more involved in the data ecosystem, not only as producers but also as consumers of the information generated by other agencies.

Access to data by citizens, researchers, and other public and private actors is a right. Data production should be oriented toward generating knowledge that can be integrated into individual and collective decision-making processes. It is highly recommendable to enable techniques for comparing the functioning of formal and informal institutions and the impact of the regulatory and public policy measures adopted. This goal must articulate measures for citizen collaboration in creating and improving public services based on the concepts of transparency, collaboration, accountability, and participation.

Public administrations must be a boosting and driving force behind an authentic open data culture, a culture in line with the Digital Spain Plan 2026 and the IV Open Government Plan for Spain 2020–2024. Collaboration between administrations, the private sector, and civil society is essential to complete the data value chain, encouraging the dynamism of private initiative and civil society as a whole when creating new value-added products and services based on data, which ultimately facilitates the achievement of the national and European objectives of promoting a fairer, more inclusive economy in line with the 2030 Agenda.

The publication should consider the FAIR principles (findable, accessible, interoperable, and reusable data), including current and historical information, evidencing the dynamic nature of the data, publishing under simple and homogeneous open licensing conditions, and guaranteeing specific service standards. Practices or agreements that prevent data reuse or limit their dissemination by creating exclusive rights to their reuse should be avoided.

It is not enough to publish data under an open license; its effective reuse must be addressed and published with a purpose while understanding the specific needs of the different sectors and user communities. Potentially reusable information must be identified right from the design of the information systems, making good the

principle of "open documents by design and by default" as expressed in the European Directive 2019/1024.

Open data is a crucial part of the sectoral data spaces that facilitates the development of the role of the data intermediary in offering value-added services on the basic information provided by the administrations. Establishing new data-based partnerships between administrations and industry is imperative, fostering a culture of open data with which the industry can develop new business models.

High-value data sets (HVDS) are very valuable for many beneficiaries and are associated with considerable benefits for society, the environment, and the economy. This recognition, defined by the Data Office in addition to the European Commission's implementing decision, will make accessible information essential for economic growth.

High-value data sets will be accessible from the Data Platform (API, bulk download) with the appropriate levels of service, internally and externally, taking into account whenever possible their geospatial component. The publication of the high-value data set must be accompanied by appropriate actions to measure its actual use and impact on society.

The ubiquity of geospatial data and its interdisciplinary function make it particularly valuable as a database for building other information, and its publication should be encouraged. Geoinformation and Earth observation data are fundamental to finding solutions to societal challenges such as climate change and environmental protection, sustainable supply of raw materials, energy transition, and internal and external security, thus laying the foundations for a digital value creation chain.

9.3.6 Principle of Evidence-Based Public Policy Design and Analysis

Data is an opportunity to facilitate the design and implementation of evidence-based public policies; to make informed decisions based on more accurate and updated data; to provide more and better services with a greater focus on citizens, ensuring their efficiency and allowing their effectiveness to be measured; and to facilitate research activity as a means of creating value and transparency in public management.

Data is essential for understanding society's problems and needs and assessing the impact of public policies. Evaluating the effectiveness of public measures based on the evidence of data on their results is crucial for creating social and economic value.

The actions and communications of administrations can be more effective, evidence-based, transparent, and sustainable when based on valid and solid data provided promptly during decision-making. Intensive use of data can drive innovation in public sector performance, facilitating the contrast of ideas and promoting creativity and the maximum use of resources in the general framework of modern,

participatory, open, and valuable public management to solve or improve social problems and challenges.

In the framework of public policy design, agencies will verify what data are relevant to them, whether these data are already available or can be collected, and what results can be derived from them. Harnessing the power of big data technologies and tools provides a more complete and accurate view of reality by enabling the collection and analysis of tremendous amounts of data from various sources, making it possible to identify problems or needs of society and to evaluate, even in real time, the impact of the policies implemented. The information available in public sector data spaces and the analytical tools deployed within the Data Platform are essential to this process.

Administration information should be accessible for social research or public policy analysis. Public and private researchers duly authorized by the competent authority must be able to access the information in an agile manner, beyond the physical rooms, and the information present in a single organization, guaranteeing, in any case, the privacy of the information processed and the moderate difficulty of its reidentification.

Public administrations' achievements and results directly impact the improvement of public policies, demonstrating the appropriateness of investing in the effective deployment of this line of action. Intersectoral access for research purposes will be articulated through the Data Platform, offering a single point of access where the different data catalogs, controlled vocabularies, computing resources, basic and advanced data analytics tools, and researcher support will be made accessible.

By adopting the role of data donors, citizens, within the framework of "citizen science" initiatives, can contribute valuable information to creating public sector data spaces, which should be considered in the pool of information accessible for research purposes and subsequent publication as open data. A data altruism policy should be designed, defining the objectives of general interest for which citizens would be willing to donate their data and creating a platform to exercise such altruism.

9.3.7 Data Culture Promotion Principle

Any process of organizational change requires the strong support of its staff. Proper data governance and management requires the creation of new positions, responsibilities, and units in each organization related to working with data, profiles such as data analysts, data engineers, data stewards/custodians, statisticians, data scientists, and data visualizers, with a deep knowledge of the area of activity and closely linked to the business.

Building on recent experience and expertise, a network of data experts, coordinated by the Data Office, should be established to share knowledge and experience, eliminate functional silos, and provide horizontal support services using innovative analytical tools. Each organization must be able to exploit the analytical capabilities

provided by the Data Platform and may require specialized support personnel initially or on an ad hoc basis. The different data profiles must have in-depth knowledge of the activity area and be closely linked to the business.

Knowledge about available algorithms, use cases, data sets, coding notebooks, vocabularies, and semantics should be easily accessible. The Technology Transfer Center (CTT) of the e-Government Portal and the Semantic Interoperability Center (CISE) play a key role, and their content and use should be promoted.

Adequate promotion of the data culture makes it necessary to design data training itineraries for administration personnel, both for management, technical, and generalist profiles. The existence of multidisciplinary profiles should be encouraged, combining knowledge of economics, sociology, data analysis, and information technologies, among others. More generally, emphasis should be placed on the necessary personnel training to enable them to obtain and conduct necessary data processing in self-management mode.

Externally, although with a clear internal projection, the focus should be on the dissemination and communication of the data culture, constituting a true community of knowledge. The objective is for the datos.gob.es platform, beyond publishing open data, to become a real showcase for data-related initiatives, a focus of knowledge, and a generator of community around it.

9.4 Conclusions

Data has become the incredible transforming power of society. Its capacity to generate knowledge, drive innovation, and empower individuals and communities is undeniable. The role of administrations in facilitating the collectivization of the value generated and properly governing the data can offer a better, critical service to citizens.

The European Data Strategy aims to strengthen and boost the Digital Single Market, fostering the creation of federated data ecosystems that promote collaboration and avoid the concentration of market power. The strategy focuses on developing innovative capabilities, harnessing the potential of data, making the right connections between data and cloud services, and protecting European principles and digital rights.

One of the main novelties of this strategy is the focus on the collectivization of the value generated in data ecosystems. These ecosystems are based on community, transparency, innovation, and the ability to scale and generate shared benefits. Unlike platform models, where much of the value is retained in intermediation, the European Union is committed to an ecosystem model that allows participants to maintain autonomy and simultaneously collaborate in point-to-point transactions. In short, the strategy highlights the power of data to transform the economy and society.

Data, understood as a public asset, is a critical element in the digital transformation of public administrations; it is their true transforming power. Public

administrations are vital agents in the collectivization of their value. Achieving a citizen-centered, open, transparent, inclusive, participatory, and egalitarian administration requires a data-driven approach. Such administration will leave no one behind.

The values, principles, and strategic lines proposed allow us to continue advancing in the construction of a data-oriented administration capable of taking advantage of the potential of data through innovative technological means, improving its efficiency and effectiveness by relying on transparent and informed decision-making.

Moving forward jointly and harmoniously on all aspects, as mentioned earlier, will unleash the transformative power of data in society. This will enable the effective deployment of an innovative data-driven administration, allowing the design, implementation, and evaluation of citizen-centric public policies, empowering a data-driven, sustainable, inclusive, and social value-generating economy.

References

1. Kumpula-Natri, M. Building a European data strategy. The Parliament Magazine. https://www.theparliamentmagazine.eu/news/article/building-a-european-data-strategy (2021)
2. Dion, O., Pons, A.: Data de Confiance: Le partage des données, clé de notre autonomie stratégique. Digital New Deal (2023)
3. Edler, J., Blind, K., Kroll, H., Schubert, T.: Technology sovereignty as an emerging frame for innovation policy. Defining rationales, ends and means. Res. Policy. **52**(6), 104765 (2021)

Chapter 10
Data Governance in the Insurance Industry

Juan Francisco Riesco

10.1 The Insurance Industry and Its Main Features in Terms of Data Governance

The insurance industry is one of the first sectors that started betting and investing in data governance, probably only after the banking and tech sectors. Yet it is also true that the approach adopted by some other insurance companies is significantly different from those in other sectors, with no standard pattern of design, deployment, or scaling of data governance in the insurance sector. The commonality across the industry is the regulated nature of the sector, with regulators placing a high value on data governance as a key practice to monitor, develop, and invest in.

The type of companies that do most of the insurance business is mature and stable with decades of existence, complemented by some digital start-ups and insurtech. Different company areas use data, usually with a vertical focus on their own business or processes and with varying degrees of depth. Consequently, and regardless of how many years of existence insurance companies may have, the use of data is part of their DNA: from the very beginnings of the business, they needed to assess the probability of occurrence, severity, and recurrence of the risk events they were insuring and the basis of the company.

In addition to DNA data, the insurance industry is considered traditional, so it is not the most attractive for data professionals. Companies must deal with their capability to attract data workers, and data governance must be in charge of this challenge.

In this chapter, we analyze data governance in the insurance sector based on these six characteristics that define and describe this industry:

J. F. Riesco (✉)
Mutua Madrileña, Madrid, Spain

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_10

- Heterogeneous data governance strategies in the insurance sector
- Insurance: a regulated sector
- Mature and stable companies
- High data utilization and evolving data culture
- The traditional focus on operational excellence with a vertical approach
- Insurance attractiveness challenge

10.2 Heterogeneous Data Governance Strategies in the Insurance Industry

When looking at insurance companies, it is easy to realize that there is not a common approach in the development of data governance. Some central aspects reveal different ways of deploying data governance in the companies; some examples of these aspects are:

- Defensive vs. offensive strategy
- The role of the CDO
- Centralized vs. federated model
- Data strategy and value creation

10.2.1 Defensive vs. Offensive Strategy

Some insurance companies follow a more offensive strategy, linking data governance to analytical products and models, while others follow a more defensive strategy where data governance is associated with regulatory and corporate reporting. Neither strategy is easy to implement, and each has its challenges. In the case of an *offensive strategy*, it has the advantage of an easier way to measure the acquired value, e.g., by improving data quality, the predictive performance is increased, and the associated business process is also enhanced (measuring variation in KPIs related to underwriting, claims, combined ratio, cross-selling, retention). But it also has the challenge of involving teams that are pretty independent and autonomous (e.g., data scientists or business data experts), and these teams are often unwilling to delegate or lean on other teams for structuring data, defining variables, or improving the quality of the information they consume.

On the other hand, in the case of a *defensive strategy*, teams involved in reporting to the ExCo/board or the regulator (e.g., business controller, finance, or risk teams) tend to understand better the importance of managing shared and agree with KPIs with reachable, risk-friendly thresholds for the company. However, in this case, the challenge appears in measuring and showing stakeholders the value of trusting the data, avoiding further inspections, limiting discussion about figures, reducing

production time, or investing more time in analysis than understanding or understanding adapting figures.

The relevance of good communication is common to both strategies. It is crucial to explain where the value of data governance comes from, highlighting the importance for that specific case, other users, and the whole insurer.

In summary, compared with other sectors, the data governance approach in the insurance industry is more hybrid than in other sectors: In the banking industry, data governance pursues a more defensive strategy; in contrast, telco or retail industries come from a more offensive strategy.

10.2.2 The Role of the CDO

Another particularity of the heterogeneity of data governance in the insurance industry is how the role of CDO is deployed. Functions developed by CDOs vary from company to company. In some companies, the CDO position is more focused on key data management functions such as data governance and data quality (whether or not combined with data strategy); in others, it is more focused on technology disciplines such as data architecture and modeling; in some others, they are leading centers of excellence for BI or advanced analytics. It is even possible to find some examples of CDOs in business lines.

As a result, the reporting lines also change according to the selected combination of focus and the organizational structure. For instance, CDO reporting targets include from the CEO to CIO, maybe including CMO, CFO, and even middle management.

The functions assigned to the CDOs and their level of reporting are a good thermometer of the company's willingness to be data-driven. In any case, the most important thing is to be consistent. A CDO cannot be asked to generate direct revenue if its main objective is standardizing reporting to the board of directors. Similarly, a CDO cannot be asked to standardize board reports if their main objective is to generate value-added services for end users.

10.2.3 Centralized vs. Federated Model

Insurance companies must also choose the most suitable data governance model. This decision also involves different choices. Choosing a governance model is a trade-off between specialization and control on the one hand and autonomy and proximity to the business on the other. But it is also a decision that is influenced by the timing of the decision; the best alternative to start deploying data governance may be utterly different from evolving from scratch the company and scaling the data governance function across the enterprise. It is not uncommon to be clear about

the target model but to be faced with the dilemma of how to get started and take the first steps.

Some companies typically select a centralized model to create core teams with deep knowledge about data disciplines with skilled and specialized teams. Center of excellence might appear for data quality, modeling, architecture, governance, analytics, and reporting. Apparently, a centralized option seems easier to start, but insurance companies face some challenges with this model. The main challenge is how to create these centralized teams.

Grouping the most advanced data users to work together in the same data area will largely benefit the company. This grouping will be the basis of the center of excellence, which will manage the knowledge of the business in a more integrated way, and will enable better ways to ensure the best-trained professionals on technical and methodological data governance subjects. But unfortunately, achieving this group of people is not always straightforward; it is not frictionless, and there is no full warranty to assign the right people (in terms of skills and professionality). In addition, people's capacity might be limited by their backpack tasks and working methods. Their starting point might differ significantly from person to person. Consequently, companies must invest efforts to create a homogeneous team to be trained and work together. These efforts include reskilling people who have been working for the company for many years and with high average age, which is common in insurance companies.

Some companies hire highly qualified personnel, either because they do not see themselves as capable of tackling these efforts or because there are no data professionals of the required profile within them. If so, this qualified personnel must be selected based on a carefully defined profile to ensure the right mood and skills are met. Challenges are also present for this option: first, it is necessary to attract talent to the insurance company; second, it will be required to have the skills to gain knowledge about the insurance business, legacy systems, current data repositories, and data pipelines, something not straightforward that implies a steep learning curve. Therefore, what can be initially seen as the fastest way to launch and propel data governance in an insurance company might have similar or even looser maturity periods than gathering expert people from different business areas.

On some other occasions, insurance companies initially opt for a federated model. This alternative benefits from avoiding bottlenecks and prioritization dependencies from a central team. In these cases, establishing data governance from scratch based on business teams distributed across the organization requires a high degree of maturity and commitment to follow data management guidelines and best practices that implies coordination with other areas of the company. Despite the apparent difficulty, this model is preferred by the vast majority of prescribers (including data mesh promoters) even when it implies a strong belief and knowledge about data governance and a strong understanding of how the company has decided to implement it within all the federated areas. This issue could be overcome by chapter teams that advise, support, and accompany the creation of data products promoting and ensuring that corporate standards are met initially until the federated teams do it independently.

10.2.4 Data Strategy and Value Creation

Defining a sound data strategy consistent with the company's corporate strategy is the cornerstone for achieving the stated business goals. Consequently, to support data strategy goals, it is necessary to consolidate teams that can work aligned to achieve the maximum data value for the company. Again, different focuses appear in the insurance sector regarding how to achieve and capture value from data.

Business goals usually require creating different lines of business based on data, for example, leads generation based on data services or sale reporting based on aggregated and anonymized data.

Some companies' business strategies include reducing costs by enhancing their business processes. This cost reduction may involve the "datafication" of some parts of the business that were not previously observed. Examples in the insurance industry of these efforts can include the following:

- Increase contact rate by having good contact details or knowing the best moment and channel to interact with the user.
- Improve fraud detection through more information about the cases and better knowledge about the relations and patterns.
- Create data-driven claims processes that reduce the time and the cost of repairs at the same time, and increase customer satisfaction.
- Digitize invoicing processes using OCRs, RPAs, and data standardization.

In other cases, data is seen as a source of rising profits. So, insurance companies pay attention to increasing customer lifetime value based on data. Examples of this challenge are as follows:

- Increase conversion rate by minimizing data asked by the customer because it is already available or because it can be retrieved from external sourcing.
- Expand the coverage of the current policies, or promote the contract of new policies based on a better knowledge of the customer and their potential needs.
- Increase the company's lifetime and retention rate based on identifying timeliness moments of the truth and sharing required data among departments to manage and personalize customer offering to increase satisfaction proactively.

The insurance industry also has scenarios of "bancassurance": a business case where the channel is a banking entity and the factory an insurance company; this is usually funneled through a joint venture between the insurer and the financial institution. In these situations, data strategy and value creation are based on increasing the sale of insurance products in the banking network. The financial entity's workforce and customer data are essential to do this. The most significant benefits are achieved when combined with the insurance company's knowledge, expertise, and product personalization capabilities. Consequently, integrating and coordinating banks' and insurance company's data governance efforts are critical. The combination and coordination require sharing the knowledge and data-based know-how of both banks and insurance companies without sharing data (subject to GDPR and

other privacy laws). In addition, the bank team must understand and translate that data knowledge into commercial and retention actions for their portfolio of customers. However, regardless of the particular case, data strategy, and monetization goals, both companies must agree to be consistent and to put in place the resources, operating model, and functions required to achieve the established strategy. Unfortunately, as the situation may vary from insurer to insurer, there is no unique recipe for deploying this combined data governance.

10.3 Insurance: A Regulated Sector

The insurance sector, like finance, utilities, or telco industries, is regulated. This fact has some clear implications for data governance. This subsection outlines the impact of being a regulated sector on data governance initiatives.

An insurance company sells a product and receives money (premiums) from the customer to take on risks to which the customers are exposed. So, the customer expects that if something under the insurance coverage happens in the specific covered timeframe, the insurance company will compensate for the consequences of that event. Therefore, insurance companies receive money up front that might be used in the future to pay customers; thus, there is a relevant component of required solvency for insurance. On the other hand, insurers establish criteria to decide whether to underwrite a policy depending on a particular risk. These criteria are usually stated considering the market's offering and demand. So, there is also a business component related to market behavior which conditions business decisions to assume or not the coverage or a specific risk over the company's capability to cover all customers' premiums. In this sense, the insurance company's target can be only as large as its capability to compensate all customers' premiums in the worst scenario without becoming financially insolvent. Supervisory authorities must watch that insurance companies remain permanently solvent to prevent customers from losing their hired rights. These supervisory authorities will focus on the safety and stability of insurance companies' investments, especially in difficult times, and on the fairness and protection of policyholders and users while dealing with insurers.

Consequently, insurance companies must submit much more information to the market, authorities, and regulators than nonregulated enterprises. Besides, this information is used to compare the company to other companies, monitor their own evolution, and assess its solvency and conduct. The company must use standard and stable definitions of the business concepts and ensure the quality of the provided data to support these operations. In summary, data governance practices should be in place for the information shared with the market and with authorities. Additionally, supervisory authorities encourage insurance companies to have in place data policies, data committees, and data functions that ensure that good practices in data management are available in the companies. These practices include the continual inspection of business processes to assess and verify that (1) policyholders and users are treated fairly; (2) there are no discriminations in the underwriting or claims

process; (3) advertising, marketing, and commercial practices met expected standards; (4) internal processes to calculate premiums, taxes, and claim payments; and (5) all the reporting is working as defined in the internal procedures.

In this context, it is easy to understand that data governance must be more promoted and implemented to cover all data used in the reporting activities (internal management reporting for decision bodies and regulatory reporting for authorities) and critical business processes than in other nonregulated companies.

Despite this claim, insurance companies must still face where and how to locate the data governance function better. When the main reason to promote data governance within business areas is to meet regulatory needs, data management functions can be seen as a second line of defense (after business and before audit teams). In this case, these management actions are focused on reporting or ensuring that inspections are in place, but they are not involved in the business's daily activities. This focus reveals a defensive strategy, but it might be a burden for an offensive strategy, where data management is genuinely embedded in developing new business products and processes.

Therefore, being regulated can help many insurance companies start certain data governance functions; however, the company requires more profound thoughts to align the chosen business and regulatory data governance goals. For example, consider the case that a company has decided to follow an offensive strategy based on significant teams to generate a holistic view of the customer to increase knowledge about its potential needs and likes to increase profits per customer. In this case, it might not be a good idea to have the vast majority of the central team focused on regulatory inspections and reporting, not giving excellent service to business areas that depend on the holistic view of the client to meet their business goals.

10.4 Mature and Stable Companies

We can outline the insurance sector as a group of mainly mature and stable companies; of course, there are some new companies, start-ups, and insurtechs, but the ratio of the business they have gained is not representative of the industry as a whole. Deploying and scaling data governance in mature and stable insurance companies has advantages and challenges to bear.

Among the advantages, developing data governance in stable and solvent insurance companies provides a known and steady environment where the course of action for data governance can be maintained by making only minor adjustments to things that might not be working as expected. Additionally, in contrast to other types of companies, insurance companies usually have an investment capacity to allocate data governance programs in the short and medium term. Insurance companies are used to more extended maturity periods than TMT or retail companies, which help to combine short-term initiatives with medium- and long-term ones, creating a good foundation for the future. And at the same time, insurance companies usually develop data governance programs that answer daily needs.

This stable environment is also risk averse (as part of the DNA of an insurance company), which, jointly with the perception of not having significant threats from outside the industry, projects a feeling of security. Therefore, there is no pressure to transform the companies; more than straightforward strategies, compromise solutions are frequently met. This means the term would be the *progressive change* instead of *transformation*. That usually also applies to the data governance operating models, where data functions are not consolidated in one team with the autonomy to define, create, put into production, and evolve data products. Organizational structure tends to be more traditional, maintaining part of the teams where they traditionally used to be. Sometimes, the split of data teams tries to be compensated with agile formulas or functional reporting. This structure might work better in multinational groups with matrix reporting cultures than in companies with hierarchical traditions. Another aspect to bear is the effort and time required in change management. Change management is always time- and effort-consuming, but it can be even higher when the average age of the staff is high, the average tenure is also high, and changes are seen as progressive with long maturity periods. As a result, the learned lesson is that adapting the plans' horizon to the companies' reality is crucial.

10.5 High Data Usage with Data Culture in Progress

First, it is vital to understand the difference between data usage and data culture. Data governance pursues creating a data culture in companies. Unfortunately, the number of insurance companies that have arrived at the point of having an extensive data culture company-wide is scarce. However, data usage is ample in most insurance companies since it is linked to the insurance business, assessing the probability, severity, and recurrence of specific events associated with the insured risks.

When discussing data usage, we describe situations where the company's main areas use their data as a part of key business processes. Typically, management reporting is used for monitoring business performance, including different levels of the organization in different ways. But in general terms, it is possible to state that people use their data, based on their solutions, without any need to coordinate with other departments because they feel self-satisfied enough with their data for their purpose.

In contrast, when discussing data culture, there is an understanding that data is a corporate asset. Thus, data might be necessary for other areas, and every employee should look after the data that is maintained, cleaned, improved, and so on. Likewise, employees understand the value of using data from other areas, which can help to improve the information managed by the business process and the ability to enhance its performance. Data governance is appreciated as a "must" because data need to be understood, trusted, and structured to avoid misunderstanding, lack of confidence, and inefficiencies. Therefore, data should be validated in origin while captured, and the department which produces this data is responsible for defining, controlling quality, and offering them to other people in the company.

Having defined both terms, let us have a look at how this works in insurance companies. In the first place, data usage is inherent to the insurance business. Actuaries use data daily to determine underwriting policies, fix premiums, and negotiate reassurance; this happens from the first day insurance company is created. The same happens with claims, accounting, controlling, or people in charge of the different businesses to monitor and improve the performance of the various business processes.

In the second place, when talking about data culture, it is not so common to have reference people appointed and playing a governance role, projects and data producers following data management standards and best practices, having in place a program for transmitting the relevance of data to the whole company, training the targeted people to increase their data skills or for evolving the information and solutions to be self-sufficient for using corporate data for their daily tasks. We can find some companies that have appointed data governance roles with greater or lesser activity in the maintenance and evolution of data products. We can also find some companies that have focused on self-service and have trained certain people or areas to use particular data consumption tools. And we can find a few insurance companies that have in place an extensive communication and training program for the whole company. But finding insurance companies with a data culture in business is hard, and very few can be positioned as data-driven companies with a comprehensive data culture.

10.6 Traditional Focus on Operational Excellence with a Vertical Approach

Insurance companies are traditional, meaning they have many years of history (not greenfield). Additionally, some have grown through several acquisitions and several integration processes. From the earliest, the search for efficiency in each process with a very vertical focus has been a mantra to gain profitability and be competitive in the market. Consequently, the mandate to the heads of the different areas was to optimize each part of the value chain separately for many years.

When we look at the use of data in each department, we find several characteristics that might be linked to this guideline of optimizing each part of the process in a legacy company. Firstly, we can see that the grade of data used to optimize the processes varies among departments, with limited use of data from other departments. Secondly, the sophistication of data usage and analysis depends mainly on the knowledge or conviction of the department head or on another specific person who promoted, at some point in time, more intense use of data inside the area. Thus, it is straightforward to identify which sites are the most advanced and who was the promoter of that situation. Particular areas might vary from company to company, but the pattern is typical across organizations. Thirdly, there are asymmetries in the maturity levels of data management and data consumption among areas and

employees in similar positions. Let us analyze how the data governance function in the insurance industry must consider these three aspects:

- Traditional optimization focus on departmental data
- Grade of sophistication dependent on particular data promoters
- Asymmetries among end data users

10.6.1 Traditional Optimization Focus on Departmental Data

Having departmental data available for analysis in a legacy company was even an important accomplishment in many areas. Therefore, much of the efforts made by some areas were focused on gathering and making (as better as they could) the detailed data of the area available. Thus, when talking about the data environment and looking at the company's different departments, it is not unusual to find data silos, different architectures depending on the area, spaghetti data flows, and various analytical tools for the same purpose. In this context, there has been low reuse of data, KPIs, and pipelines during the years. Likewise, specific cross-tasks involving coordination by different areas or various business units were found difficult to implement in some companies. However, it is usually possible to find more advanced data innovations, e.g., master data management, 360° view of the customer, or standard corporate data models or data repositories (e.g., corporate data warehouses, corporate data lakes). Of course, in the last years, more and more cross-functional initiatives are arising and being demanded in companies to gain a holistic view of data initiatives like customer journeys, promotion of seamless omnichannel personalization, or increase of customer satisfaction in processes that involve several areas. However, it is also imperative to remember the traditional working method that is usually still in place.

We should understand this history when deploying or evolving data governance in insurance companies. First of all, we need to tear down the barrier of using vastly only departmental data. Creating forums where departments share and explain the available data that other areas can use is vital. Promoting data exchange can start from existing data and continue later by including regular communication about the new data made available with every data product put into production by each different area. Through that, people will have a broader knowledge of data that can help in their daily tasks.

Secondly, from the data governance perspective, it is necessary to promote the creation of corporate structures that generate an efficient, unique source of truth, as well as simplify and make more accessible the exploitation of data. Usually, it is more straightforward to have more data (since insufficient data affects each user) than to understand the relevance of structured data with a corporate view and standard definitions (agreed upon by the different stakeholders). But this is the pillar of reusing data. The main reason for this roadblock is human: promoting standard definitions and structures across the board implies to make involved and including in

"my projects or my tasks" some other areas, which will probably have their vision and which also will have to say something about "my data and how I should organize them." Consequently, it is required to change how projects are done, involving new functions and roles but minimizing the potential overhead to achieve certain maturity for these disciplines. From the beginning, it is necessary to feel that data governance helps to create better and faster products since more business knowledge and data expertise are allocated to the project.

Thirdly, it is required to be very cautious about using new technologies. Stakeholders should consider that moving to the cloud, creating data lakes, or implementing data fabric architectures might not solve, per se, the existence of data silos. Technology is only technology and, of course, can help make some projects more manageable; but the lack of technology did not cause spaghetti data flows and data silos. To create a holistic data ecosystem, where data can be consumed in a self-service mode by the different areas, much more is needed than technology. Data must be understood and organized corporately, and capabilities must be in place (tech and people). Technology supports part of the data responsibilities, but it is very much important creating cross-area initiatives sponsored by top management with regular follow-up at Executive Committee. This way, it is more probable that different areas are on board and will be surfed together when difficulties emerge. Corporate structures also create technology interdependencies among other business systems, but once again, the answer is not only technology. Sound synchronization between the legacy operating systems and the analytical systems and vice versa is fundamental. So, it also requires establishing new procedures and bodies to coordinate the unique situation.

In this context, many things must be done that are not straightforward and that need to change ways of working: for instance, the first decision is delimiting scope and deciding where to start, setting clear goals, communicating properly to all involved teams, and regularly monitoring the status with top management to make those initiatives, shared initiatives and, if possible, with shared incentives.

10.6.2 Grade of Sophistication Dependent on Particular Data Promoters

It has already been stated that data usage in insurance companies is relatively high at the different levels of the organizations and in different areas. But, of course, some areas, such as actuarial or commercial, have historically been more data-intensive. Apart from them, there are other areas (that vary from company to company) with extended use and management of data (on some occasions complaints, in other operations, but also it might appear some business lines like life or health, or even support functions like finance and risk). There is a common root in the sophistication of a department using data; as already highlighted, it depends on the department head or any other skilled data employee who has the opportunity and autonomy to create

data products for the department. Therefore, the most sophisticated areas using data in each company will be determined by a combination of the functions of the area and the exceptional team compounding it.

Several factors can determine the adequate level of sophistication in data usage: the use of interdepartmental data, the use of external data, the existence of standard definitions, the monitoring of validations, the improvement of data quality, the creation of data structures avoiding data replication, the type of data products and analysis performed, or the kind of analytical models developed.

Once these advanced areas are located, key data-skilled people (let us call them *data promoters*) are also identified very soon. These data promoters have valuable knowledge about source systems, existing repositories, products, KPIs, and tools to get the most out of data. Additionally, these data promoters can create departmental repositories and be asked to create them. Most likely, these people are the reference people in providing data to the area.

From a data governance perspective, advanced areas and data promoters are a gift to the organization but are also challenging to manage. This good breeding group might be turned into a defiance position since they are essential when anyone wants to know more about data (definition, logic, origin, usages). As data governance looks or should look after the democratization of access, knowledge, and use of data, it is vital to give a relevant and structural role to these people and areas.

Talking about roles, another challenge appears: one of the leading hypotheses when naming roles is their capacity to make decisions related to the business data domain. And in all cases, as data promoters are not usually department heads, they might not be data decision-makers. To solve this situation (and some similar others), there are different roles to be appointed, like data owners and stewards. These types of appointments can happen in both business and IT areas. Therefore, roles must be thought to seat these situations.

Data promoters are usually critical in resolving any data incident in BAU processes and any relevant data project for the area. So, they are generally busy with little time for additional tasks that the new role might require. Freeing up these people's time is also a key challenge for sharing knowledge, propelling cross-projects, and supporting change management. To achieve this time-freeing, official recognition of the role and new functions, together with a transitioning plan, is overriding to make it a reality.

In summary, due to the relevant number of data promoters or data power users in insurance companies, it is essential to define a strategy for how the governance model will take advantage of this situation and accommodate the role map.

10.6.3 Asymmetries Among End Data Users

In previous paragraphs, it has been outlined that some departments in insurance companies are usually more advanced in the use of data than others. Therefore, on average, employees in more developed areas should have more extensive expertise

in the use and management of data than in other areas. But even within a particular area, there are people more skilled than others in the use of data. The reason why those areas do not have more qualified, experienced people can be due to the lack of time, the lack of ability to execute data actions, or the availability of other teams providing that service. Summing up, it is possible to find very different starting points for employees in similar positions that would be willing to use and be autonomous in the application of data in their daily tasks, but it is also possible to find similar people expecting to take advantage of data in very different ways. Asymmetries are evident, and it is necessary to deal with them.

It is also possible to detect situations where end users try to analyze data or develop data products that other users have already done. Not sharing this knowledge about existing products leads to the feeling of needing to create everything, every time from scratch. In general terms, this is a symptom of a poor data culture in which best practices have not been shared between departments or even inside a particular division. If possible, insurance companies should have skilled and powerful workers in specific data disciplines to better support the area. However, if this knowledge is not shared adequately with others and conveniently extended, the knowledge, the know-how, and the related capabilities will abandon the company as the worker leaves out.

In this situation, data governance should tackle three points: creating data communities, defining training tracks, and designing walk alone programs.

Firstly, *data communities* are crucial to sharing the acquired know-how about developing available data products, tips, and best practices, as well as locating reference people in mind in case somebody needs help. The community should work in a decentralized manner, where central teams should not be intermediaries and only the promoters of content and activity. They can provide the community with videos, papers, updates, templates, and other artifacts seen as accelerators in the use and management of data. In addition, they can encourage community members to create helpful content identifying best practices for the company.

Secondly, not all employees want to manage their data in the same way, and neither they part from similar starting points when it comes to using the data. Therefore, when defining *training tracks* is important to create different modules which suitably combined can support several training paths. On the one hand, people should be able to choose the training track that best suits and contributes to the target scenario they want to achieve; it is important to realize that different visions and knowledge can be required to execute data tasks depending on the position and the person playing the same or similar position. On the other hand, modular training gives the flexibility to self-adapt the content based on current status, goals, and available time.

Thirdly, skilling people based on a "one-size-fits-all" approach can lead to many people having finished specific courses but not acquiring new abilities to apply in their daily tasks autonomously. It might be required to create personalized "walk alone programs" to achieve that objective, where the user learning how to deal with data as part of their functions has the support of a specifically better-trained or experienced leading person. This leading person is in charge of (1) promoting the

learner's self-assessment to decide and customize upon the results, which is the best training track, including modules that best fit the learner's needs, (2) supporting the first steps of the learner toward the way to being autonomous, and (3) helping the learner to overcome any stopper that may rise when exploiting data in their functions.

In conclusion, in this context of asymmetries, there is a patent need for a critical role in data governance that makes sense to be central; let us call it *data culture promoter*. These data culture promoters should be focused on creating data communities and propelling the activity and quality of content, communications, and interactions in those communities. They have also to depict the training programs, adapted to the reality of the insurance company, being important to create different itineraries and a syllabus as modular as possible to fit different needs in other forms. Training people is not enough, and they should also create a plan to support people on their first, second, and third data up/reskilling steps that are much more than a typical change management action plan. And finally, they must monitor the results of all these activities and evolve and change what might be necessary.

10.7 The Insurance Companies' Challenge of Attracting Talented People

In any data governance deployment, which ultimately implies a certain level of transformation for the company, talented people are a crucial factor. These talented people are the basis for implementing or scaling up several data functions requiring more capacity to better serve the company's needs. Consequently, insurance companies need talented people.

The first option should always consist in trying to upskill or reskill the current base of employees to make the company evolve. But as previously talked about in this chapter, the average age and average tenure of employees in insurance companies are pretty high compared to other industries, and skilling existing employees might be challenging. This factor does not mean that evolving data governance with current employees would not be possible: it only means that the time required for that evolution/transformation could be higher when trying to do it only with the existing base of employees. Therefore, if an insurance company wants to start the transformation in the short term or accelerate it, it will probably have to hire people.

The problem is that currently, there is a high demand for data professionals in the market. These days, new, attractive positions are regularly offered to data professionals, even when they are not actively searching for new opportunities, leading to frequent job changes.

When considering hiring talented data people, there are typically two options: hiring young people with data and tech backgrounds and training them in the company and hiring experienced people with deep knowledge, expertise, and working years in data disciplines. On the one hand, the insurance industry is not usually a

sector that could appeal to the youngest generations. These new types of professionals value features typically related to the insurance industry, for example, collaborative working, new agile methodologies, cutting-edge technologies, sharing external data, using sophisticated data analytics techniques, or working in open and dynamic environments with no hierarchies. These features are linked to other more contemporary and trendy sectors like technology, media, retail, or telecom. But on the other hand, experienced people look more for job stability and security, typology of projects to develop, and a pleasant working mood where it could be easier to perform daily tasks. They also scout each company deeper regarding the managerial team, growth possibilities, dependents and autonomy, and level of dialogue.

Insurance companies require very likely both types of talent, young and experienced. Therefore, offered positions must be attractive for two kinds of profiles. Providing both types of positions is an important challenge because transforming a company is impossible without having balanced talent. But fortunately, insurance companies have specific valuable tools to attract and convince young and experienced data professionals. First, insurance companies offer the possibility of finding a balance between personal and professional life—the reader is encouraged to compare job offers with those in the consultancy sector or other highly demanding industries like online retailers or media companies. Second, insurance companies usually offer competitive benefits, including reasonable salaries, pension plans, health insurance, and bonuses linked to stable and secure companies. Third, transformation plans are in place with relevant investment capacity in data governance, so projects and challenges await new joiners. These features are not usually recognized in the market for professionals, requiring explanation and carefully showing the value of each one of these aspects in each recruiting process.

As discussed, data professionals—even those who are not actively searching for new positions—receive offers quite often; therefore, hiring is only the beginning, and insurance companies have to continue being attractive to data employees day by day. It is imperative to invest in training and innovation and create collaborative and productive work environments to achieve this goal because, in the end, these employees want to develop their careers by doing exciting things in a pleasant mood, maintaining employability with potential growth options.

10.8 Insurance Trends and Their Impact on Data Governance

This chapter began by highlighting that "the insurance sector is one of the first industries that started betting and investing in data governance." However, it must be noted that due to the maturity and stability of the sector, other industries that started investing and deploying data governance later have already surpassed the data governance global state of the art in the insurance sector.

Trends around the insurance world encourage us to think that data governance in the insurance industry will receive a new impetus. To explain this statement, let's outline what kinds of things are changing, what insurers do need to face the unique situation, and how data and data governance can contribute.

Firstly, let us understand what kinds of things are changing in the insurance industry. New demands from new generations and older people that extend their life expectancy can be observed. Some demands are related to the way of interacting with the insurer: it is a more direct, digital, and mobile-based relationship. Other demands are driven toward the product offering: they must be more modular, flexible, and customizable but also ensure new risks (e.g., mobility, climate change, cybersecurity, social media, retirement funds, and elderly care). Besides, users are going to be more demanding, in terms of autonomy, immediacy, and data disclosure, and they will look at insurers to solve their daily needs not only for having a policy (frontiers among industries are blurring).

Secondly, let's see how insurance companies can face this situation. They need to be more customer-centric, more natively omnichannel, with more hyper-personalization capabilities (in terms of relationships and products). But also, when companies have learned much more about their customers, they need to put data and information available for the end users to make their own decisions, as well as the need to foresee future needs and offer solutions to cover them in a broad and structured way (more than just traditional policies).

Thirdly, let's guess how data and data governance can contribute to providing users with the best service. Digital and omnichannel processes are intense in data, so there is a need to retrieve more information, structure it, and make it available for all interactions. When talking about new risks to be insured, many of them are intensive in data (e.g., new ways of mobility—like autonomous cars, drones, or car sharing fleets—cybersecurity threats, social media activity, and climate change risks, among others). The disclosure of more information to end users requires high data governance standards. But the necessary evolution from internal processes also needs more and faster data available to meet emerging and future demands (e.g., continuous underwriting, personalized payments, premiums adapted to changing contexts, or new methods of assessing provisions).

In summary, combining new trends, new entrants, and other industries' inertia, together with market speed, will favor the relevance of data governance in the insurance industry.

Chapter 11
Data Governance in the Health Sector

Alberto Freitas, Julio Souza, and Ismael Caballero

11.1 Importance and Implementation of Data Governance in Healthcare

Technological advances in healthcare, namely, the introduction of electronic health records (EHRs) and the increasing adoption of emerging technologies related to digital health, have created challenging environments for health organizations, as the amount of data has grown exponentially, and thus a radical change in the scale, methods, and capabilities for data gathering, aggregation, and analysis is required [1]. In this sense, every aspect of healthcare, from management to daily clinical practice, has been more and more underpinned by data and information, making those strategic assets in the healthcare sector. Overall, health data are crucial to enhance the quality of the delivered care, to support scientific innovation, to ensure patients' safety, and to support efforts aimed at shifting the classical model of care, reactive and focused on the disease, to a more preventive, personalized model of care, centered on the patients. Furthermore, many health organizations are currently dealing with the need of Big Data technologies, which in turn have created high expectations in terms of healthcare revolution by taking advantage of the enhanced computing power to process large and broad ranges of health data in real time. These technologies can result in gains to the public interest in several health-related areas,

A. Freitas · J. Souza
Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS)/ Center for Health Technology and Services Research (CINTESIS), Faculty of Medicine, University of Porto, Porto, Portugal
e-mail: alberto@med.pt; juliobsouza@med.up.pt

I. Caballero (✉)
DQTeam/Alarcos Research Group, University of Castilla-La Mancha (UCLM), Ciudad Real, Spain
e-mail: Ismael.Caballero@uclm.es

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_11

such as diagnostics, treatment selection, personalized and preventive medicine, telemedicine, health population support (e.g., to capture and monitor disease trends and outbreaks), medical research, and cost reductions (e.g., fraud detection, insights into better patient care resulting in long-term savings) [2].

Although the access and analysis of Big Data have the potential to transform healthcare practices and their outcomes, stakeholders and decision-makers first need to understand what is required to make the best possible use of all the data generated, gathered, and stored and avoid the associated shortcomings [3]. According to the classical four V's definition, Big Data means high volume of data, generated at a high velocity and coming at different types and formats (variety), which in turn brings issues on the accuracy and reliability of the data (veracity) [4]. The so-called Big Data revolution in healthcare has been on hold due to failures to deal with this scenario, as many of the health data are of poor quality and are available in the form of small, incompatible datasets. Moreover, the added value extracted from Big Data relies on advanced analytical techniques such as those from artificial intelligence (AI) and machine learning, which can be highly susceptible to poor data quality. Current practices concerning data collection, curation, and sharing make it difficult to analyze health data on a large scale. Modern health data standards that assure adequate levels of data quality and ensure greater data compatibility for pooling and timely access to data by researchers and other stakeholders are basic requisites to fulfill the potential of Big Data and data-driven healthcare. Effective data management practices are clearly needed across health organizations, and these organizations need to invest in specialist training in data science and information technology. This points to a growing role of data management specialists and knowledge engineers at the organizational level who can pool and curate datasets in healthcare settings [5].

Nevertheless, data governance in the healthcare sector is typically less mature than in other industries, and health organizations tend to require more time to improve their data and analytics maturity levels [6]. Given this scenario, especially in the Big Data context, numerous challenges and constraints exist to deal with health data, namely:

1. Data Complexity Healthcare-related data, including from health research, is typically complex and more heterogeneous when compared to other fields. This complexity of health data is mostly due to its highly unstructured nature, relying on natural language processing, apart from being dispersed, fragmented, and typically non-standardized. EHR sharing across organizations and even within the organizations' lines is a constant problem. The several pieces of information to populate the EHRs are often generated through specific systems, such as magnetic resonance imaging (MRI) scanners and pathology applications, creating further constraints to aggregate and analyze clinical data [7]. Additionally, healthcare data has been increasingly collected outside clinical encounters, such as pharmacy transactions, claims data, and other emerging communication technologies, such as innovative wearable systems and Internet of Things (IoT) [8], usually involving big instant data

gathered from multiple sensors or systems, and that must be capable of providing continuous and autonomous services [9].

Because of the unique complexity of health data, traditional approaches to managing data will not work in the healthcare sector. Instead, different approaches are needed, focusing on handling the multiple sources, the unstructured and structured data, the lack of consistency, variability, and other issues arising from data complexity, within a constantly changing regulatory sector. Therefore, to cope with these unpredictable changes and inherent complexity, organizations must invest in data governance programs, specifically tailored for healthcare, design, and implemented, passing through reevaluations, making corrections and adjustments whenever necessary. Therefore, to tackle the complexity of healthcare data, data governance frameworks must be flexible enough to be extended to as many healthcare settings as possible while facilitating the adjustment and incorporation of environment-specific data requirements, characteristics, and processes involved in the data life cycle.

2. Data Privacy and Security Significant concerns regarding privacy and confidentiality exist in health research and the healthcare sector due to the high sensitivity of health data. During data collection, especially in clinical trials and healthcare surveys, obtaining patient consent is a critical and challenging step. In this sense, healthcare organizations expect the data to be stored and held in secure databases, where only authorized individuals are allowed access. On the other hand, a considerable sharing of the information is centralized and thus vulnerable to external attacks [10].

In April 2016, the European Commission agreed to replace the Directive 95/46/EC [11] with the General Data Protection Regulation (GDPR) [12], which entered into force in May 2018. The GDPR is a key component of the European Union (EU) privacy law, addressing concerns regarding data access and security and giving EU citizens increased control over their personal data. Moreover, the GDPR also intended to simplify the regulatory environment for business in the digital health area, introducing the concept of data protection by design and per default, in which all services and products for the EU market must include data protection in their design, throughout all stages of development [13]. The GDPR has become a model for national laws worldwide, with an estimate of 10% of the world's population having its personal data covered by the GDPR in 2019 [14]. In the United States, focusing on for-profit organizations, the California Consumer Privacy Act (CCPA) is a regulation similar to GDPR, signed into law in June 2018, in which several consumer privacy rights and business obligations were defined on the collection and sale of personal data [15].

In this sense, data governance programs will need to address the existing national and international regulations on data privacy and security, having to balance the plethora of opportunities and value brought by health data, especially in the context of Big Data, to improve healthcare management, practices, and outcomes, while preserving the right of citizens to control their own data. As mentioned earlier, the modern sources of personally generated health data, coming from emerging

communication technologies such as wearable devices, may fall outside of existing regulations and policies on privacy [7]. Different organizations using health data for different purposes, such as research centers, companies, and hospitals, should ideally have data protection offices that need to deal with regulation issues on privacy and security, being key actors for the proper implementation of data governance programs.

3. Traceability of Patient Data Clinical practice is substantially impacted by how well medical information is gathered, processed, accessed, and communicated between healthcare professionals and clinicians [16]. In the digital age, healthcare organizations should ideally ensure that professionals can access clinical data in optimal conditions and all patient data should be utterly traceable across the entire health system. Throughout the years, significant advances have been achieved, but yet on-demand access to medical information is still far from being adequate in several settings, resulting in increased efforts, costs, undesirable outcomes, and decreased efficiency [16]. Despite the increasing adoption of EHRs and the evolution of information infrastructure supporting healthcare provision, not all health data sources are effectively connected, and information systems deployed in healthcare facilities are mainly devoted to support local operational tasks, being implemented without an integrated perspective, resulting in significant data heterogeneity data duplication.

The traceability of patient data is an essential aspect that urgently needs to be addressed in the healthcare sector. In fact, the GDPR itself has introduced specific articles concerning the importance of establishing activity recording and how to operate over it to ensure data privacy and security. Article 30 of GDPR [17] requires organizations to maintain a complete record regarding all personal data processing activities, whereas article 32 [18] states the need for organizations to implement measures that lead to adequate levels of security across data processing operations. Therefore, there is increased pressure on healthcare organizations and software producers to implement auditable traceability approaches tailored for their current systems [19].

4. Interoperability and Standardization Overall, healthcare data, especially in the context of Big Data, deal with a wide range of different standards, language barriers, and clinical terminologies. EHR systems themselves, even at the organizational level, are usually fragmented, and patient data is maintained in formats that are not compatible with all the technologies and software applications required to process them, causing further issues regarding data acquisition, transferring, cleansing, analysis, and sharing [10]. Inconsistent variable definitions and the speed at which new evidence-based practice and research emerge are key constraints to implementing standardization. The idea of standardization is directly linked to the concept of interoperability, which is defined by the Healthcare Information and Management Systems Society (HIMSS) as "the ability of different information systems, devices and applications (systems) to access, exchange, integrate and cooperatively use data in a coordinated manner, within and across organizational,

regional and national boundaries, to provide timely and seamless portability of information and optimize the health of individuals and populations globally" [20].

Healthcare data, especially regarding Big Data, comes in different formats across several small and incompatible datasets. Interoperability governance must ensure that interoperability occurs at four fundamental levels: (a) foundational interoperability, which refers to the ability of different systems to exchange data between each other; (b) structural interoperability, which is the ability of the system receiving the data to interpret the information at the level of data fields; (c) semantic interoperability, which refers to the ability of a system to exchange, interpret, and actively use the exchanged information, where healthcare professionals and authorized personnel are able to share patient information (this level of interoperability allows the improvement of the quality and efficiency of the delivered care and promotes patient safety); and (d) organizational interoperability, which can be understood as the goal of most healthcare organizations, facilitating the safe, clean-cut, and timely use and communication of the data between and within organizations and people [21].

Data governance programs should thus define the most appropriate standards to be adopted in specific contexts. There are currently a variety of health standards and initiatives, such as openEHR [22], Fast Health Interoperability Resources (FHIR) [23], Digital Imaging and Communications in Medicine (DICOM), Health Level Seven International (HL7), SNOMED CT, the Unified Code for Units of Measure (UCUM), and Continua Design Guidelines (CDGs) [24], each of which with their own specificities, advantages, and disadvantages. Thus, it is up to the data management team, in alignment with healthcare professionals and other stakeholders, to decide which standard is suitable for the data needs, which is a decision that is heavily linked to the specific setting and the underlying clinical scenario.

5. Timely Data Access [5] Accessing and sharing clinical research data is a highly efficient way to foster scientific knowledge. Big Data, which is a combination of several datasets, can bring even more advances and benefits for healthcare and society, being the reason for which several international consortia are investing efforts to build Big Data-driven translational research platforms to produce high-quality scientific evidence on disease-specific causes and risk factors, diagnosis, prognosis, and medical treatments [25]. Translational research aims to transform scientific discoveries produced in laboratories and clinical trials into novel interventions and treatments, with the utmost goal of disseminating these discoveries to improve healthcare and the population's health. Considering the anticipated benefits related to large-scale sharing of health data, ethical issues arise, forcing stakeholders to address and manage multiple privacy and confidentiality aspects, ensure that valid informed consent is provided in the context of clinical research, and determine the people who will make decisions regarding the access to the data. To find a balance between ethical issues and potential benefits, data sharing platforms need support concerning the compliance with regulations such as the GDPR, in the EU, and the CCPA, in the United States, as norms on personal data sharing for health research remain open to researchers' interpretation and only limited practical guidance is provided. Moreover, timely access to health data for research is a major bottleneck,

as higher benefits are obtained if the patients' data are shared as soon as possible. Still, even publicly available datasets are usually shared only after the completion of studies, when results have been published, meaning that data analysis by other researchers can occur with a delay of months or years [25]. Ethical guidance and governance are critically needed to boost fair and sustainable data sharing for health research, especially amid the efforts to build Big Data translational research platforms. Data governance programs should provide clearly defined data sharing policies specifying how data requests from internal and external actors will be registered, tracked, and managed and how data sharing will occur in a secure and efficient way.

All the challenges mentioned above have clearly introduced an urgent need for improved data culture within health organizations. As mentioned, health data is particularly complex, requiring huge efforts to link, aggregate, clean, and transform data obtained from multiple systems and sources. Healthcare organizations need to prioritize the implementation of frameworks addressing aspects of data quality (DQ), data management (DM), and data governance (DG). DG can be generally understood as the process of managing data assets throughout their entire life cycle to ensure they meet the quality standards of an organization. Health-related DG programs must include the people, processes, and systems used to manage data throughout the entire data life cycle, ensuring greater data quality and allowing data to benefit the organization, its users, and even the society as a whole [26].

The remainder of this chapter will present a case study of Portugal to illustrate part of a data governance effort in the hospital sector through a framework denominated CODE.CLINIC, which includes a Process Reference Model (PRM) for governing and managing hospital administrative data, with emphasis on data produced through clinical coding. Basic definitions and concepts regarding the PRM and their contribution for implementing data governance programs will also be further provided in this chapter.

11.2 A Case Study of Portugal

11.2.1 Clinical Coding and the Hospital Information Structure in Portugal

In Portugal, there is an extensive healthcare data structure across nearly all levels of care, supporting the collection and storage of data constantly used to drive quality improvements across different healthcare settings. Much of this rich data infrastructure is a consequence of the increasing use of EHRs over the last years, paired with unique patient identifiers. Data sources in the Portuguese health system include setting-specific information structures, disease-specific registers, and individual-level data sources [27].

The information infrastructure in the hospital sector is as extensive as those deriving from primary care, and a high level of standardization already exists in terms of discharge summaries, clinical reports, and surgical checklists that are under nationwide guidelines, facilitating planning and quality monitoring for all hospitals within the National Health Service (NHS). In fact, standard monitoring indicators are computed and collected across different dimensions from multiple hospitals at national level (e.g., access, performance, quality, and financing/costs), and the reported data is publicly available through an online platform (https://benchmarking-acss.min-saude.pt/) on a monthly basis [27]. Behind this rich healthcare data structure in the Portuguese hospital sector, there is a comprehensive nationwide hospitalization database, the National Hospital Morbidity Database (HMD), which maintains a wide range of data on inpatient and outpatient episodes occurring in all public hospitals and public–private partnerships within the Portuguese NHS. This database is regularly updated following the collection of administrative, demographics, and clinical data resulting from the several routine processes in hospitals [28].

The collection of clinical data begins with the documentation of all clinical information and services provided during hospital encounters through a variety of data collection instruments, in paper and/or digital formats, namely, narrative discharge notes and pathology and surgical reports. Once the patient is discharged from the hospital, this information is then accessed by physicians that have been trained and licensed as medical coders. Information access by medical coders is based on a standard software application called SClinico, which allows the retrieval of the different pieces of information from the patients' EHR. All these primary data are then assessed by medical coders, who should evaluate and assign codes to each diagnosis and symptom according to the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) codes and procedures according to the International Classification of Disease, 10th Revision, Procedure Coding System (ICD-10-PCS) [29]. This process is most of the times manual, labor-intensive, and time-consuming. All hospitalization data, including the clinically coded information, is stored in the national database (HMD) through another standard application implemented in all public hospitals at a national level, the Hospital Morbidity Information System (SIMH, from its acronym in Portuguese).

In this sense, clinical coding is the primary source of clinical information behind hospital administrative datasets. These datasets have been mainly used for managerial purposes, such as financing, resource monitoring, resource allocation, and decision-making. These data are also a major nationwide epidemiological information source, apart from supporting the conduction of health research and the benchmarking of hospital providers by means of monitoring indicators. Considering the increased use and reuse of these data, it is paramount that clinical coding processes at hospitals produce reliable, accurate, comprehensive, up-to-date, and consistent data. The set of processes to produce these data is complex. It involves a diverse workforce, from healthcare providers to IT and administrative staff, as well as resources and protocols, which often vary across hospitals. Furthermore, there are several points in the data life cycle at which barriers to high-quality data may be

introduced, including quality issues related to the patient's original documentation, available resources, training and support for coding, interpretation of the documented clinical information, and the level of adoption of official guidelines. Consequently, various data quality issues arise, affecting the usefulness and trust level of the produced data for proper use and reuse.

Several of these barriers related to clinical coding, with potential impact on data quality, have been identified in the literature, with many of them focusing on medical coder-related barriers on training, knowledge, and evaluation standards [30–36]. In Portugal, several specific barriers were highlighted by medical coders during focus groups and interviews [37–39], namely:

- Lack of awareness on the importance of health records
- Subjective nature of medical language, often presenting a nonstandard syntax, unclear abbreviations, and heterogeneity regarding diagnoses descriptions
- Poor communication between healthcare workers, namely, between medical coders and the healthcare providers responsible for the original information reporting
- Lack of relevant information for audits
- Variability in adopting, accepting, and interpreting official guidelines to standardize health records
- Lack of precise patient documentation
- Incomplete or unclear discharge notes, missing discharge notes for specific services (e.g., outpatient surgery services), and incomplete or missing surgery reports
- Delays regarding coding and delivering the coded information
- Lack of supporting tools designed to help medical coders during their activities
- Decreased productivity, with many medical coders performing different activities besides clinical coding
- Variability between hospitals in terms of frequency and processes for clinical coding auditing

Additionally, numerous common coding errors have been reported in the literature, highlighting the wrong selection of principal diagnosis codes, missing additional and comorbidity-related codes, and choosing nonspecific codes, resulting in loss of clinical information [40]. Other coding errors frequently described in the literature are misspecification, miscoding and resequencing, and deliberate coding errors aimed at obtaining financial compensation or avoiding administrative penalizations (e.g., upcoding, down-coding) [41, 42]. In Portugal, some studies have also identified significant inter-hospital variability in coding comorbidities and nonoperating room procedures, indicating potential issues concerning coding accuracy and credibility [43, 44].

Given the potential impact of the data quality issues on hospital financing, management, and research, it is therefore essential to provide ways of governing data and thus ensure increased data quality in a systematic fashion. However, managing data quality in hospital settings is characterized by high complexity, often involving multiple information systems, stakeholder groups, workers, rules,

and processes. A useful nationwide data governance framework should be carefully designed to ease the existing issues and barriers and strengthen the value obtained from the data.

In this section, a PRM denominated CODE.CLINIC will be described, which in turn can be understood as part of an effort to implement data governance frameworks across the Portuguese hospital sector, targeting the improvement of hospital data quality, namely, those generated by means of clinical coding. The hypothesis behind this initiative is that the existing problems can be mitigated by gathering and grouping a set of processes, or best practices, to govern the entire data life cycle, seeking a more homogeneous and high-quality clinical coding across hospitals, including during the phases of use and reuse of the data, either within the organization or externally by researchers, health authorities, policymakers, and other healthcare stakeholders. This set of best practices or processes should ideally cover all aspects of clinical coding, data quality management, and data governance. Furthermore, this PRM can also serve as a body of knowledge and guidance for the several clinical coding processes, including the identification of relevant stakeholders, information systems and software applications that are employed to support the processes, and key performance indicators to monitor the implementation of the PRM processes within the hospitals.

Before presenting CODE.CLINIC, the main concepts and purposes behind a PRM and how it can support health organizations in implementing effective data governance programs are explained.

11.2.2 CODE.CLINIC PRM

A PRM can be understood as a set of processes supporting the organizational process model, comprising processes addressing aspects of data management, data governance, and data quality management [45]. The process-based approach behind the CODE.CLINIC PRM was developed in alignment with ISO/IEC 8000-61 [46] for data quality management and ISO 12207 [47] for software processes, also meeting the data governance, data management, and data quality management requirements defined in the ISO 8000-61 framework-compliant Alarcos' Model for Data Maturity, MAMD (from its acronym in Spanish) [48].

The MAMD framework is a model developed by experts in data governance and data management from the University of Castilla-La Mancha, Spain. MAMD, currently in its fourth version, includes a PRM with 22 processes grouped into 5 organizational maturity levels. It is important to highlight that the third version of MAMD was the one in which CODE.CLINIC was based. Thus, MAMD encompasses, in a joint and coordinated way, a set of processes for data management (DM), data quality management (DQM), and data governance (DG). The DM component defines good practices regarding the technological infrastructure management required to meet the organizations' business requirements. The DG component defines good practices related to the design of organizational data strategies aligned

with the organizations' business strategies. The DQM component refers to good practices to optimize business data quality requirements.

Additionally, the MAMD also provides a mechanism to evaluate and improve the capacity of the organization's processes regarding these three components (DM, DG, and DQM). This mechanism is referred to as Process Assessment Model (PAM). The PAM presents the elements organizations need to evaluate and improve their activities following the defined PRM. The PAM was designed so that the requirements of ISO/IEC 33003 and other parts of the ISO/IEC 33000 series are met [49]. Furthermore, the PAM comprises a key component, the Maturity Model, which links the processes defined in the PRM to distinct maturity levels and sorts these processes in an increasing level of difficulty, according to the organizations' capabilities. There are six maturity levels defined in the MAMD: maturity level 0 or immature; maturity level 1 or basic; maturity level 2 or managed; maturity level 3 or established; maturity level 4 or predictable; and maturity level 5 or innovating (for further details on the different maturity levels, see Chap. 7). It is up to each organization, based on its own capabilities and business requirements, to establish the targeted maturity level they intend to reach and which processes from the PRM shall be included in the different levels.

Overall, to implement the DM, DQM, and DG components, as defined in the MAMD's PRM, it is important first to identify the most relevant and needed processes according to the different levels of maturity. The processes are typically tailored according to the organizations' reality. Moreover, organizations need to adapt the definition of the MAMD processes according to their own characteristics so that the results of the processes can be accomplished. Finally, the definition of the MAMD processes needs to be adapted to the degree of capacity that the organization aims for.

As said, the specification of CODE.CLINIC used the MAMDv.3 to define tailored processes that comprehensively address several aspects of clinical coding and all data life cycle phases, comprising the DM, DG, and DQM components. The processes characterize the formal pathways of the coded data and can be used as a source of knowledge to guide specific activities during clinical coding. All the information structured by the PRM can be used to outline clinical coding processes when designed from scratch or to review and improve existing processes by identifying barriers and the underlying root causes. Therefore, every process defined in the PRM can be understood as a "knowledge box" where different stakeholders can find the necessary knowledge, including activities and work products, communication schemes, and related key performance indicators to be monitored. Additionally, processes can be reviewed from time to time to enrich the existing model and include new activities and/or work products, accompanying changes in guidelines, rules, new data, and business requirements, and changing technologies.

The design of CODE.CLINIC PRM[1] was initiated with the description of the entire life cycle of coded data by identifying all processes and actors involved in the clinical coding production in a Portuguese public hospital considered a reference in clinical coding. In this sense, the formal pathways and processes regarding clinical coding were traced at the hospital level. To collect this information, a series of interviews with an experienced clinical coder at the reference hospital, who presented a more complete view on the entire data life cycle, were performed. Information collected included documentation sources and instruments used for clinical coding, information systems and software applications involved, coders' education and training, guidelines and reference instruments, how clinical information is collected in routine processes, quality control procedures (e.g., internal or external audits), people and institutions involved, available tools to support coders, current norms and regulations at hospital and government levels, how of the produced data is used and reused, who are the users, and how data storage, curation, access, and sharing are processed.

A total of 16 processes distributed across 4 broad categories were defined in the first version of CODE.CLINIC, using the concept of Primary, Support and Organizational processes specified in the ISO/IEC/IEEE 12207. This structure enables a better understanding of the processes' purposes and their contribution to the general aim of clinical coding. The four categories of processes are:

1. **The Strategic Processes—"G Processes":** This category of processes addresses key DG processes involved in clinical coding, mainly those related to the definition and identification of standards at the organizational level, best practices, guidelines, rules, and policies behind the several stages of the coded data life cycle, with emphasis on the organizational structure and human resources. Strategic processes also define the people involved in the several activities and how to enable the communication between the different parts. Additionally, G processes also address how health organizations should provide the necessary personnel's specific competences and skills.
2. **The Main Processes—"M Processes":** Main processes cover all the aspects related to the adequate clinical coding itself, describing the several activities within the coded data life cycle, from data acquisition to the use and reuse of the coded data.
3. **Support Processes—"S Processes":** In this category of four processes, the specificities of quality management of the data used as input (patient documentation) and output (coded data) of the coding clinical are covered. In addition, the concerns related to technological infrastructure management along with the maintenance of the reference data standards are also covered.
4. **Other Processes—"O Processes":** Finally, the O processes group includes other processes that do not fit into the previous categories but are part of the data life cycle and thus directly or indirectly impact DM, DG, and DQM processes. In the

[1] The full PRM of CODE.CLINIC can be downloaded from https://medcids.med.up.pt/wp-content/uploads/sites/730/2023/04/Modelo-Referencia-Processo_CODE-Clinic.pdf.

context of clinical coding, these processes are those related to the hospital encounter itself and the underlying care provided, which in turn will be the origin of all clinical information.

Furthermore, each process within the CODE.CLINIC PRM was defined in compliance with ISO/IEC/TR 24774 [50], which characterizes the processes according to the following components:

- Title: consists of a descriptive heading for the processes
- Purpose: description of the main goal of the health organization when executing a given process
- Outcomes: represent the expected results from the successful execution of a given process
- Activities: a concrete list of actions, or best practices, required to achieve the expected outcomes

The CODE.CLINIC PRM was designed to be comprehensive and flexible enough to be adapted to different hospitals. The outcomes and activities should be properly selected and reinterpreted according to the specific context. The involved actors and stakeholders that are relevant for the customization of CODE.CLINIC have been identified and categorized into three distinct groups:

1. **Consultive Roles:** This group includes policymakers in the health sector, typically external to the organization, usually at the regional or national level. These actors provide general concerns and recommendations concerning clinical coding in technical support, management, and interoperability support. In the context of clinical coding in Portugal, those actors include the Central Administration of the Health System (ACSS, from its acronym in Portuguese), the Shared Services Ministry of Health (SPMS, from its acronym in Portuguese), the Order of Physicians of Portugal, and their branch to assign certifications on clinical coding, the Portuguese Association of Medical Coders and Auditors (AMACC, from its acronym in Portuguese).
2. **Active Roles:** This group includes personnel directly or indirectly involved with clinical coding at the hospital level, thereby being the people required to implement the strategic, main, and support processes. Those include hospital managers at department and service levels, healthcare providers, IT (information technology) workers, clinical coding office managers, and medical coders.
3. **Benefited Roles:** This group includes actors that use or reuse the data for various purposes, such as public health authorities, healthcare managers, and researchers.

Table 11.1 lists the CODE.CLINIC PRM processes, by category. The full definition of each process, including their respective activities, outcomes, and work products, which can be understood as key resources to execute that process, as well as involved actors, can be found in Annex A.

Table 11.1 List of CODE.CLINIC PRM processes, per category

Strategic processes—"G processes"	Process G.01. Creation or selection, implementation, and maintenance of standards, best practices, norms, guidelines, and policies Process G.02. Development of policies Process G.03. Organizational structure management Process G.04. Stakeholders' skills and competences management
Main processes—"M processes"	Process M.01. Data acquisition Process M.02. Data integration (internal) Process M.03. Data coding Process M.04. Submission of clinically coded data to the national repository Process M.05. Incorporation of coded data to APR-DRG (DRG grouper software) Process M.06. Data exploitation for hospital management, financing (billing), and public health Process M.07. Data exploitation for clinical and epidemiologic research
Support processes—"S processes"	Process S.01. Data quality management of patient documentation Process S.02. Data quality management in coded data Process S.03. Reference data management Process S.04. Technological infrastructure management
Other processes—"O processes"	Process O1. Healthcare taking process

11.3 Summary and Conclusions

In the current scenario of increased generation and availability of health data within and across health organizations, the importance of governing these data's access, sharing, usage, storage, retention, analysis, and disposition is becoming paramount at an exponential rate.

To address the challenges mentioned earlier in this chapter, key aspects should be tackled for the implementation of data governance programs in healthcare, including: (a) to ensure that all support for an integrated foundation for data governance will be provided by the management/board team of the organization; (b) to allocate all needed resources to form a data governance committee, which requires a significant staff enlargement, involving data owners, data stewards, data analysts, and data architects; (c) to promote the integration between data owners with the operations and activities within the data life cycle in order to reach an effective solution; (d) to invest on staff training, defining robust strategies to ensure that the necessary skills and training of the healthcare workers are achieved, including efforts to ensure that changing technologies, novel approaches, and standards of care are kept up to date; (e) to define consistent data protection measures and appropriate procedures for data access and restriction, complying with national regulations (e.g., GDPR), which include the definition of clear data retention and usage policies; (f) to achieve the adequate levels of data quality and trust, addressing sources of inaccurate, incomplete, inconsistent, and unstandardized data, by means of data integrity policies;

(g) to deal with data complexity by defining data dictionaries, the specification of individual data elements, the relationship with other data about the individual, the way data is represented, and how clinical entities and concepts are represented, recurring to adequate health standards; (h) to access data and share policies that are paramount in DG programs to increase the value of data (appropriate access should be defined, ensuring that people within and outside the organization have appropriate access to the data; these policies include the security measures to protect data and ensure proper use of data whenever accessed and shared); and (i) finally, to tackle the lack of standardization and interoperability issues—a comprehensive data governance program for healthcare organizations should identify rules on how to relate health data to clinical concepts, requiring the use of adequate standards, and how to systematically integrate health data assets to produce high-quality information to be used for safe decision-making and ensure that data is useful, up-to-date, and relevant to fulfill its purposes [3].

A data governance program must address the existing challenges regarding health data more pragmatically. The presented case study in Portugal proposes a PRM that tackles the current challenges in the context of hospital administrative datasets and clinical coding. Yet, these challenges only represent a small constituency of those affected by the lack of data governance in the health sector. The implementation of a framework for clinical coding such as CODE.CLINIC will promote greater harmonization of clinical coding processes across hospitals and increase interoperability between organizations, enabling actions such as benchmarking and increased patient traceability. The institutionalization of the CODE.CLINIC aims to enhance the efficiency of clinical coding, promote interoperability, and improve data quality by facing the barriers discussed in Subsection 11.2.1. The PRM tackles these by means of governing solutions in a unified and controlled fashion and from an organizational perspective. In this sense, CODE.CLINIC provides a road map toward more harmonized approaches to data governance across hospitals.

Clinicians, healthcare managers, researchers, patients, and the general public are aware that health data have enormous value and are the key to driving future advances in medicine while ensuring that confidentiality and data privacy protection norms mandated in official regulations are fully complied. An effective governance of health data will contribute to the boost of scientific innovation and further improve populations' health and healthcare systems' quality. Healthcare organizations urgently need to bring together up-to-date data management practices and invest in specialists that can maximize health data's usability and quality, encouraging new policy frameworks that promote appropriate data sharing for research.

References

1. OECD: Health data governance for the digital age: implementing the OECD recommendation on health data governance. Organisation for Economic Co-operation and Development, Paris (2022)

2. Batko, K., Ślęzak, A.: The use of big data analytics in healthcare. J. Big Data. **9**(1), 3 (2022)
3. Hovenga, E.J.S., Grain, H.: Health data and data governance. Stud. Health Technol. Inform. **193**, 67–92 (2013)
4. Russom, P.: Big Data Analytics. The Data Warehousing Institute, Fourth Quarter, Seattle (2011)
5. Dhindsa, K., et al.: What's holding up the big data revolution in healthcare? BMJ. **363** (2018)
6. Tse, D. et al.: The challenges of big data governance in healthcare. Presented at the 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) (2018)
7. Winter, J.S.: AI in healthcare: data governance challenges. J. Hosp. Manage. Health Policy. **5**, 8 (2021)
8. Surantha, N., et al.: A review of wearable internet-of-things device for healthcare. Proc. Comp. Sci. **179**, 936–943 (2021)
9. Jóźwiak, L.: Advanced mobile and wearable systems. Microprocess. Microsyst. **50**, 202–221 (2017). https://doi.org/10.1016/j.micpro.2017.03.008
10. Kruse, C.S., et al.: Challenges and opportunities of big data in health care: a systematic review. JMIR Med. Inform. **4**(4), e5359 (2016). https://doi.org/10.2196/medinform.5359
11. Parlement européen et du Conseil: Directive 95/46/CE du Parlement européen et du Conseil, du 24 octobre 1995, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données. (1995)
12. General Data Protection Regulation (GDPR) Compliance Guidelines. https://gdpr.eu/. Accessed 2 May 2022
13. Santos-Pereira, C. et al.: Are the healthcare institutions ready to comply with data traceability required by GDPR? A case study in a Portuguese healthcare organization. Presented at the International Conference on Health Informatics February 24 (2020). https://doi.org/10.5220/0009000405550562.
14. Hulsen, T.: Sharing is caring—data sharing initiatives in healthcare. Int. J. Environ. Res. Public Health. **17**(9), 3046 (2020). https://doi.org/10.3390/ijerph17093046
15. State of California: The California Consumer Privacy Act of 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375 (2018)
16. Cruz-Correia, R., et al.: Traceability of patient records usage: barriers and opportunities for improving user interface design and data management. Stud. Health Technol. Inform. **169**, 275–279 (2011)
17. GDPR: Art. 30 – Records of processing activities. https://gdpr-info.eu/art-30-gdpr/. Accessed 13 Mar 2023
18. GDPR: Art. 32 – Security of processing. https://gdpr-info.eu/art-32-gdpr/. Accessed 13 Mar 2023
19. Gonçalves-Ferreira, D., et al.: HS.Register - an audit-trail tool to respond to the general data protection regulation (GDPR). Stud. Health Technol. Inform. **247**, 81–85 (2018)
20. EHRIntelligence: How health data standards support healthcare interoperability. https://ehrintelligence.com/features/how-health-data-standards-support-healthcare-interoperability. Accessed 13 Mar 2023
21. HIMSS: Interoperability in healthcare. https://www.himss.org/resources/interoperability-healthcare. Accessed 13 Mar 2023
22. Frexia, F., et al.: openEHR is FAIR-enabling by design. Public Health Inform. **113–117** (2021). https://doi.org/10.3233/SHTI210131
23. Ayaz, M., et al.: The Fast Health Interoperability Resources (FHIR) Standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med. Informatics. **9**(7), e21929 (2021). https://doi.org/10.2196/21929
24. COCIR: Interoperability standards in digital health – A White Paper from the medical technology industry. http://www.cocir.org/media-centre/publications/article/interoperability-

standards-in-digital-health-a-white-paper-from-the-medical-technology-industry.html. Accessed 13 Mar 2023
25. Waithira, N., et al.: Data management and sharing policy: the first step towards promoting data sharing. BMC Med. **17**(1), 80 (2019). https://doi.org/10.1186/s12916-019-1315-8
26. AHIMA: Healthcare Data Governance. https://www.ahima.org/media/pmcb0fr5/healthcare-data-governance-practice-brief-final.pdf (2022)
27. OECD: OECD reviews of health care quality: Portugal 2015: Raising standards. https://www.oecd.org/publications/oecd-reviews-of-health-care-quality-portugal-2015-9789264225985-en.htm. Accessed 13 Mar 2023
28. Souza, J., et al.: Multisource and temporal variability in Portuguese hospital administrative datasets: data quality implications. J. Biomed. Inform. **136**, 104242 (2022). https://doi.org/10.1016/j.jbi.2022.104242
29. Santos, J.V., et al.: Transition from ICD-9-CM to ICD-10-CM/PCS in Portugal: an heterogeneous implementation with potential data implications. HIM J. **18333583211027240** (2021). https://doi.org/10.1177/18333583211027241
30. Bramley, M., Reid, B.: Evaluation standards for clinical coder training programs. HIM. J. **36**(3), 21–30 (2007). https://doi.org/10.1177/183335830703600304
31. Hennessy, D.A., et al.: Do coder characteristics influence validity of ICD-10 hospital discharge data? BMC Health Serv. Res. **10**(1), 99 (2010). https://doi.org/10.1186/1472-6963-10-99
32. Lorenzoni, L., et al.: Continuous training as a key to increase the accuracy of administrative data. J. Eval. Clin. Pract. **6**(4), 371–377 (2000). https://doi.org/10.1046/j.1365-2753.2000.00265.x
33. Lorenzoni, L., et al.: The quality of abstracting medical information from the medical record: the impact of training programmes. Int. J. Qual. Health Care. **11**(3), 209–213 (1999). https://doi.org/10.1093/intqhc/11.3.209
34. Santos, S., et al.: Organisational factors affecting the quality of hospital clinical coding. Health Inf. Manage. **37**(1), 25–37 (2008). https://doi.org/10.1177/183335830803700103
35. Tang, K.L., et al.: Coder perspectives on physician-related barriers to producing high-quality administrative data: a qualitative study. CMAJ Open. **5**(3), E617–E622 (2017). https://doi.org/10.9778/cmajo.20170036
36. Walker, R.L., et al.: Implementation of ICD-10 in Canada: how has it impacted coded hospital discharge data? BMC Health Serv. Res. **12**(1), 149 (2012). https://doi.org/10.1186/1472-6963-12-149
37. Alonso, V., et al.: Health records as the basis of clinical coding: is the quality adequate? A qualitative study of medical coders' perceptions. Health Inf. Manage. J. **49**(1), 28–37 (2020)
38. Alonso, V., et al.: Problems and barriers during the process of clinical coding: a focus group study of coders' perceptions. J. Med. Syst. **44**(3), 62 (2020). https://doi.org/10.1007/s10916-020-1532-x
39. Alonso, V., et al.: Problems and barriers in the transition to ICD-10-CM/PCS: a qualitative study of medical coders' perceptions. In: Rocha, Á., et al. (eds.) New Knowledge in Information Systems and Technologies (WorldCIST'19), pp. 72–82. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-16187-3_8
40. Reid, B., et al.: Under-coding in Australia limits the performance of DRG groupers. Health Inf. Manage. **29**(3), 113–117 (2000)
41. Aelvoet, W.H., et al.: Miscoding: a threat to the hospital care system. How to detect it? Rev. Epidemiol. Sante Publique. **57**(3), 169–177 (2009). https://doi.org/10.1016/j.respe.2009.02.206
42. Hsia, D.C., et al.: Medicare reimbursement accuracy under the prospective payment system, 1985 to 1988. JAMA. **268**(7), 896–899 (1992)
43. Souza, J., et al.: Importance of coding co-morbidities for APR-DRG assignment: focus on cardiovascular and respiratory diseases. Health Inf. Manage. J. **49**(1), 47–57 (2020)
44. Souza, J., et al.: Quality of coding within clinical datasets: a case-study using burn-related hospitalizations. Burns. **45**(7), 1571–1584 (2019). https://doi.org/10.1016/j.burns.2018.09.013

45. ISO: ISO/IEC 33004:2015: Information technology — process assessment — requirements for process reference, process assessment and maturity models. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/41/54178.html. Accessed 11 Apr 2022
46. ISO: ISO 8000-61:2016: Data quality — Part 61: Data quality management: process reference model. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/30/63086.html. Accessed 4 Aug 2021
47. ISO: ISO/IEC/IEEE 12207:2017 - Systems and software engineering — software life cycle processes. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/37/63712.html. Accessed 11 Apr 2022
48. DQTeam: MAMD: Modelo Alarcos Mejora Datos. https://mamd.dqteam.es. Accessed 11 Apr 2022
49. ISO: ISO/IEC 33003:2015: Information technology — process assessment — requirements for process measurement frameworks. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/41/54177.html. Accessed 11 Apr 2022
50. ISO: ISO/IEC/IEEE 24774:2021 Systems and software engineering — life cycle management — specification for process description. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/89/78981.html. Accessed 11 Apr 2022

Chapter 12
Data Governance in the Telco Sector

José Luis Sanzana

12.1 Introduction

The importance of telecommunications companies in our day to day is fundamental. At present, we live in a technologically advanced society in which high-speed internet access and quality voice and video calls have become almost a fundamental need to communicate, work, study, entertain ourselves, and satisfy our basic needs for information and connection with the rest of the world.

At world-class events like the Mobile World Congress (MWC), the importance of the telecommunications industry and how it shapes our lives has been emphasized. As José María Álvarez-Pallete, CEO and Chairman of Telefónica and GSMA (Global System for Mobile Communications), declared at the last event held in February 2023, "Without us, there is no digital future." He was alluding to technological support and high-speed networks to develop and promote revolutions like the metaverse, artificial intelligence, IoT, Web 3.0, and everything that is being developed through 5G and everything that will be developed when we have access to 6G.

Not only is it growing in technology and high-speed networks, but this means that we increasingly collect an enormous amount of data; as Álvarez-Pallete comments, "In the last 10 years, data traffic has multiplied for 27 over the world. The world. It's a vast number."

We do all this information capture with small devices that are supercomputers and that we use at all times. Related to this, Álvarez-Pallete adds, "15 years ago, the mobile device, basically designed to send and receive voice calls, became into something else. At the convergence of mobile devices and the Internet, mobile computing was born."

In this chapter, we take a very summarized tour of how a telecommunications company is structured at a functional level, the type of services it provides to its

J. L. Sanzana (✉)
Zurich-Santander, Santiago, Chile

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

I. Caballero, M. Piattini (eds.), *Data Governance*,
https://doi.org/10.1007/978-3-031-43773-1_12

customers, how they collect and what means all the avalanche of data that they must manage, how we should structure specialist areas to order and govern the data to get the most out of it, and finally some examples of problems that occur between teams of specialists when they do not understand the work and advantage that disciplines linked to data governance could provide.

12.2 How to Operate in General This Type of Company

Telecommunications companies have similar operating structures associated with fundamental areas, not only business but also technical, necessary to have the different services offered.

The general functional structure follows the typical schema shown in Fig. 12.1.

Among the services offered by the operators are those related to the fixed and mobile world. The first includes essential telephone services, internet through fiber optics or cable, cable, and digital or satellite television. In the mobile world, there are telephone services (prepaid and postpaid); internet through 2G, 3G, 4G, and 5G bands; and roaming services, among others.

12.3 How Is the Data Collected, and What Can Be Done with All the Data Managed by This Type of Company?

Basic information collected by telecommunications companies is found in files called CDRs (call data records). These files store data such as sending and receiving telephone numbers, date and time of the call, duration, cost, and countless attributes that allow for a complete log of each person's calls.

In addition, mobile phones connect to base stations (e.g., antennas installed on top of buildings, tower-type structures that we see on the roads) through low-power signals that allow the position of each mobile to be geolocated.

This immense amount of data is collected with technology that supports the ingestion, processing, and exploitation of large volumes of data in real time.

Telecommunications companies have had a competitive advantage compared to companies in other industries because the mobile is considered a mobility sensor that leaves its footprint from the moment we turn it on until we turn it off.

This footprint allows analytical studies to be carried out to obtain conclusions such as where to install a store according to the flow of people, how to plan traffic during peak hours, what are the age ranges of the customers who circulate through my store, and if they are foreigner customers, among others. It is worth mentioning that all these types of analysis are conclusions obtained en masse, but in no case individualizing people to protect the rights of protection of personal data, an ethical

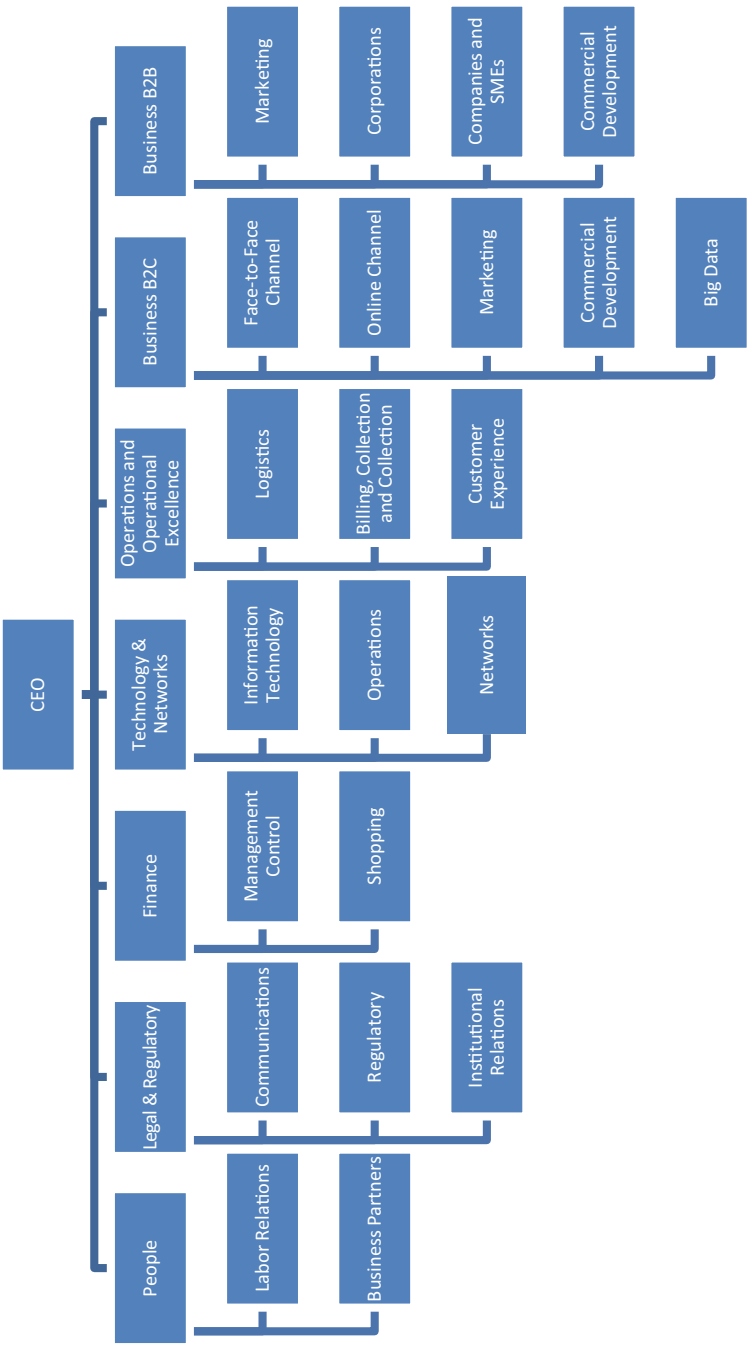

Fig. 12.1 Typical functional structure of a telecommunications company

requirement that must be applied by all companies that handle personal and sensitive customer data.

Therefore, the question we must ask ourselves is how can this type of company that obtains enormous amounts of data ensure order, classification, quality, security, and understanding of their data to get the most out of it, not only to improve their products but also to carry out studies that can be very useful for the government in power in implementing public policies that benefit people?

12.4 How Can You Govern the Data?

First, we must be clear about the functional roles the data and analytics area should have to govern the data and provide an excellent service within the organization.

For this, there can be several types of organizational structure depending on the size, priorities, and culture of the company. A typical example of the organizational structure covering data governance and other data management responsibilities is shown in Fig. 12.2.

We must be clear about how we organize our functional team and how we must order the data within our data lake or data warehouse. There may be various forms of classification, but we present two options that could give good results when ordering

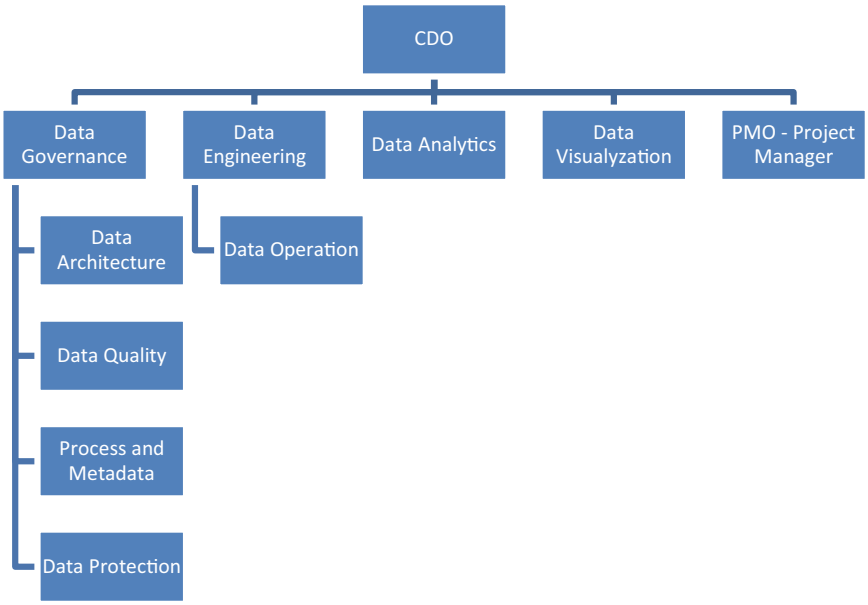

Fig. 12.2 Example of the functional structure of a data and analytics area

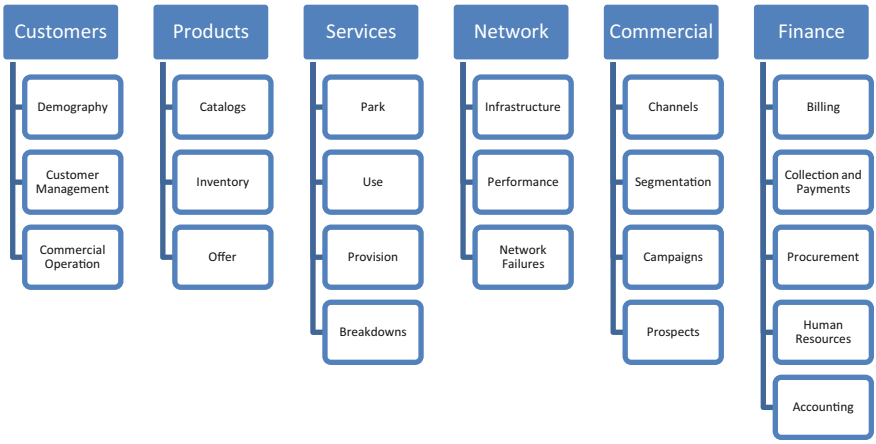

Fig. 12.3 Example 1 of data domains and subdomains for a telecommunications company

our house at the data level, which is structured into data domains and subdomains (see Figs. 12.3 and 12.4).

12.5 Problems That Can Occur in the Interaction Between Technical Teams and Specific Disciplines Associated with Data Governance

When we start a data governance program, which will involve closely interacting with other specialized areas, we must consider that it will be a process of change and continuous monitoring so that the technical teams are entirely aware of the work and deliverables of each role.

As an example, we will present part of the problems that occur in daily life between advanced analytics and data governance and how they can mutually support each other to optimize the development times of analytical models carried out by the data scientists.

If we talk about data governance, what are its primary purposes?

- Ensure that data is appropriately managed per policies and best practices.
- Support data and analytical projects in applying good practices associated with data architecture, data quality, metadata, and data protection, among others.
- Ensure that the information is updated, relevant, timely, reliable, and explicable.

On the other hand, what are the primary purposes of the analytics area?

- Analyze and exploit different sources of data.
- Obtain quality information to help make better strategic and business decisions.

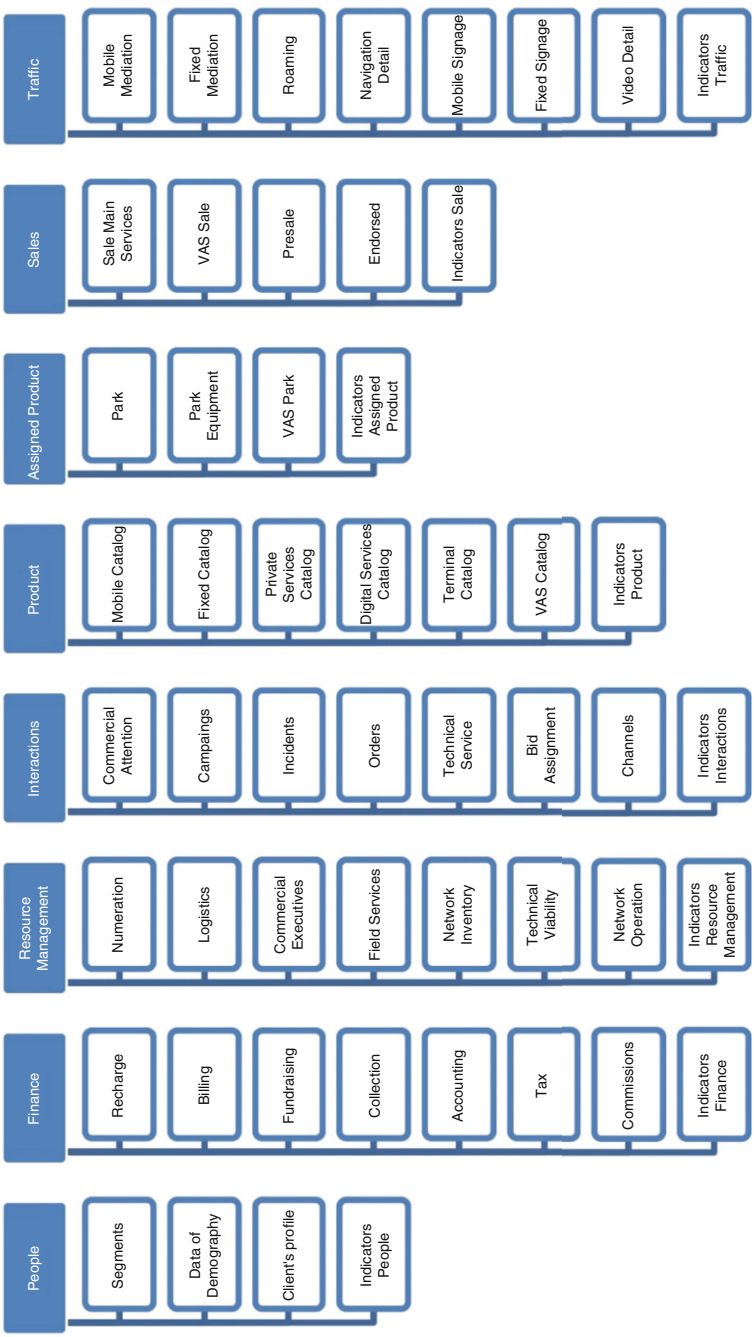

Fig. 12.4 Example 2 of data domains and subdomains for a telecommunications company

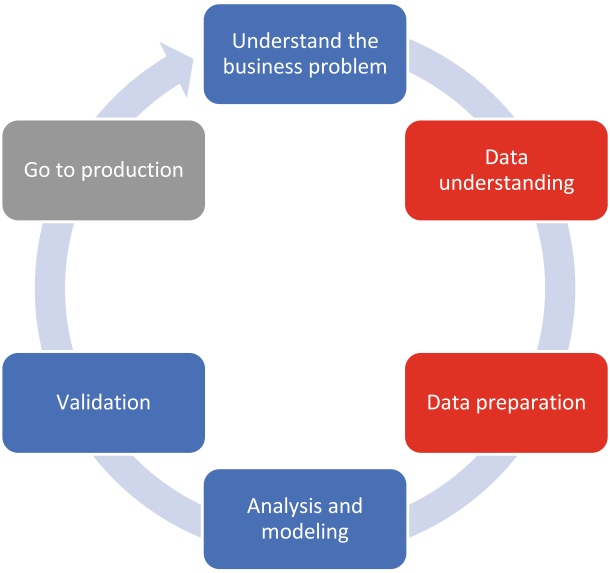

Fig. 12.5 Phases of the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining)

- Design analytical models (artificial intelligence and machine learning) and optimize decisions based on data.
- Find advanced, adaptable, and scalable analytics solutions.

In this context, how could these two disciplines work together?

Considering that one of the methodologies most used by the analytics areas is the so-called CRISP-DM, which considers six phases of the project development cycle (see Fig. 12.5).

Some data governance disciplines could support data scientists in the phases of data understanding and data preparation.

12.6 Data Understanding

The understanding of the data is directly related to understanding what metadata is (information that describes other data) and how to generate and store it so that it can be used at a transversal level in the organization.

The metadata must be stored in a data catalog, which can give us advantages such as the ones shown in Table 12.1.

Table 12.1 Benefits for data scientists of having a data catalog in the organization

Data catalog	Advantages for data scientists
Description of each data source and attribute	Agility in the search for data sources and owner of the same in case of doubts
Definition of business terms	
Data owner association to each data object	
Report the quality level of each data source and attributes (data health)	Minimize the use of unreliable data in analytical models (garbage in, garbage out)
Clarity in the traceability of the data (lineage)	Identification of the levels of data aggregation and end to end of the data flow

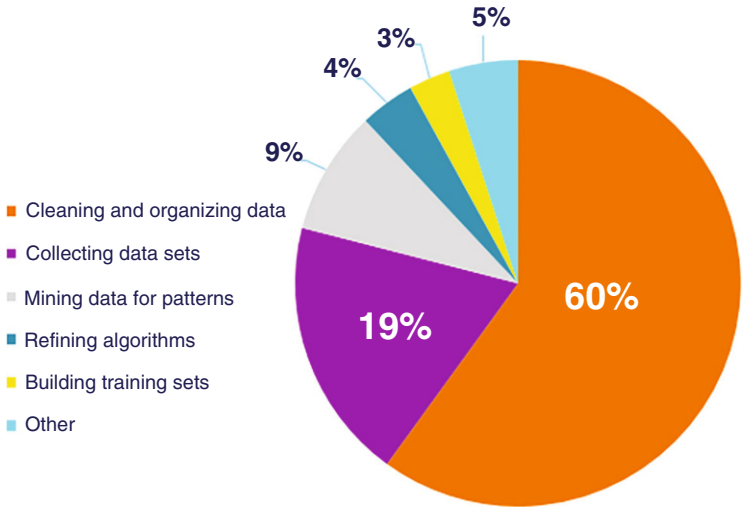

Fig. 12.6 Percentage of the dedication of a data scientist to the analytical process (https://towardsdatascience.com)

12.7 Data Preparation

As shown in Fig. 12.6, data scientists dedicate 79% of their time to analytical projects to investigate where the data sources they need are located and later clean them if the data arrives with errors from the source of origin or intermediate sources. Finally, they only dedicate 21% to constructing and creating analytical models.

As the ultimate goal is to reverse the percentages mentioned above, data quality specialists could contribute in the following way to prevent these tasks from being performed by data scientists:

- Identify and correct erroneous data by classifying it through different dimensions (% completeness, % duplication, etc.), which translates into providing analytical project teams with reliable information about the health of the data.

- Standardize the data format coming from different information sources (e.g., date format).
- Fluidly communicate between the data quality team and the data scientists to prevent the latter from implementing quality rules that remain encapsulated in the analytical models and are not transferred to the quality specialists so that they perform the remediation directly in the sources of origin.

These and other measures among the teams of specialists may drastically reduce the time in developing analytical models, but always be aware that this transition must be carefully monitored by a change management program that ensures the proper functioning of a work ecosystem that is not easy to achieve.

12.8 Main Conclusions

- We are in an unprecedented digital revolution that has led us to be at the technological forefront supported by the telecommunications industry, which provides us with fundamental support to continue promoting revolutions such as the metaverse, artificial intelligence, IoT, Web 3.0, 5G, and 6G.
- The data and analytics areas must have an operating structure that allows fluidity and agility in the development of data and analytics projects. This fluidity and agility minimize the time it takes to develop and put into production new products and offers due to the intense competition that exists in this type of industry, which is not only focused on delivering services at low prices but, above all, delivering services that improve the experience and quality of life of customers.
- Joint work and effective communication between analytics and data governance specialties generate maturity and speed in the teams and internal processes.
- Advanced analytics areas could reduce information search and exploration time by at least 50% by having a robust and updated data catalog.
- Not giving importance to metadata is not giving importance to your data. In short, it is similar to being blind at the data level.
- Data governance should be an enabler and streamliner, not a bureaucratic hindrance.