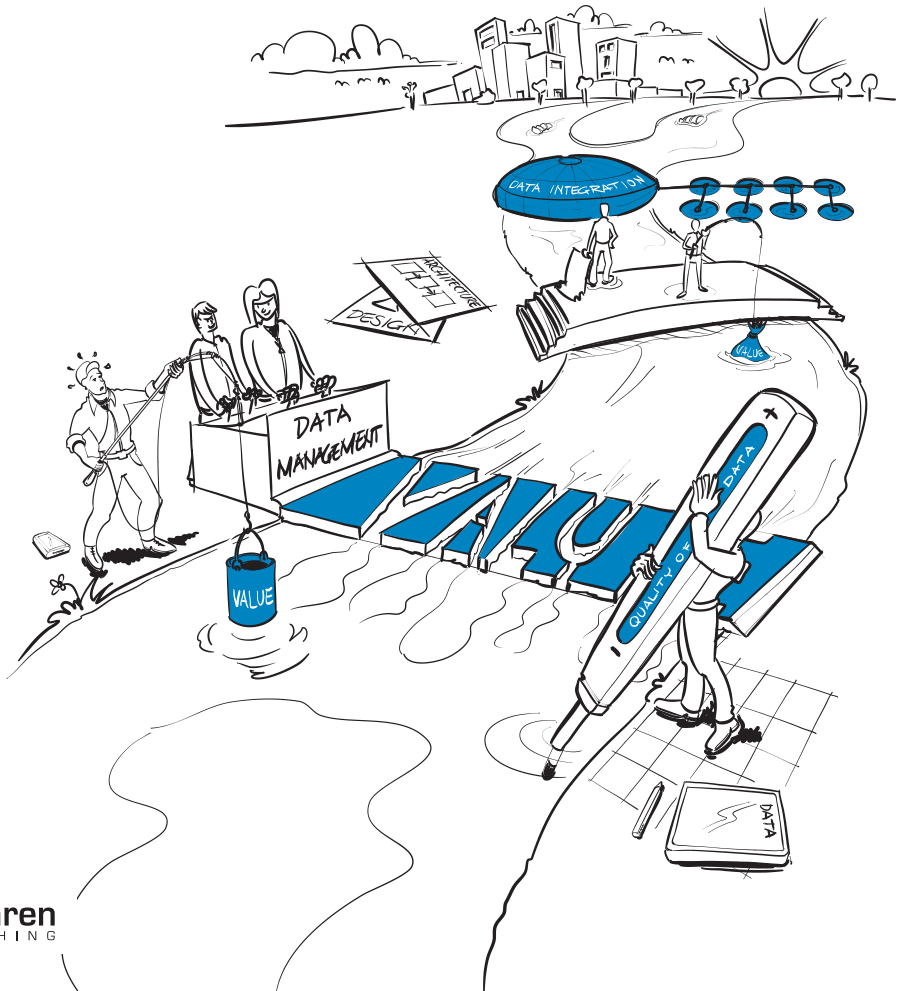
DATA MANAGEMENT
A GENTLE INTRODUCTION

2nd Edition

Balancing theory and practice

Bas van Gils



Van Haren PUBLISHING

Data Management: a gentle introduction - 2nd edition

Other publications by Van Haren Publishing

Van Haren Publishing (VHP) specializes in titles on Best Practices, methods, frameworks and standards within four domains:
- IT Management
- Architecture (Enterprise and IT)
- Business Management and
- Project Management

Van Haren Publishing is publishing on behalf of leading organizations and companies: Agile Consortium, World Commerce and Contracting, IAOP, IPMA World, KNVI, PMI-NL, NLAIC and The Open Group.

Van Haren Publishing is part of the Van Haren Group and additional to the book publishing also provides the following services: accredited training materials and e-learning through Van Haren Learning Solutions, as well as independent professional certification via examination through Van Haren Certify.

Topics are (per domain):

IT Management	IT Service Management	FitSM, ISM®, ISO/IEC20000, IT4IT®, ITIL®, VerISM®, SAF, TRIM, XLA®
	Data Management	Data literacy, Data visualization, DMBOK
	IT Asset Management	HAM, ITAM, SAM
	IT Security Management	BIO, ISO/IEC27001, NIS2
	Test Management	CTAP
	Application Management	ASL
	Other	eCF, IT-CMF, Scrum
Project Management	Project Management	Half Double, ICB, ISO/IEC21500, P3.express, PM^2, PMBOK Guide, Praxis, PRINCE2
	Agile	Agile, Agile PM
	Other	PMO
Business Management	Operations Management	Lean, Lean Six Sigma, OBM, OMC, RASCI
	Contract Management	CATS CM, CATS RVM, WorldCC
	Business Information Management	BiSL, DID
	Artificial Intelligence	AI, Generative AI
	Outsourcing	OPBOK
Enterprise Architecture	Enterprise Architecture	BIAN, TOGAF
	Modeling	ArchiMate, BPMN
	Software Architecture	ISAQB
	Other	Open Agile Architecture

For the latest information on VHP publications, visit our website: www.vanharen.net.

Data Management:
a gentle introduction
2nd edition

Balancing theory and practice

Bas van Gils

This book is dedicated to my two children:
Koen van Gils and Stijn van Gils.
You are my rock stars.

Van Haren
PUBLISHING

Colophon

Title:	Data Management: a gentle introduction - 2nd edition
Subtitle:	Balancing theory and practice
Author:	Bas van Gils, managing partner @ Strategy Alliance
Illustrations:	Andy Lo Tam Loi
Reviewers:	Mirjam Visscher and Tanja Glisin
Text editors:	Lisa Gaudette and Steve Newton (Galatea)
Publisher:	Van Haren Publishing, 's-Hertogenbosch
ISBN hard copy:	978 94 018 1312 9
ISBN eBook (pdf):	978 94 018 1313 6
ISBN ePUB:	978 94 018 1314 3
Editions:	First edition, first impression, February 2020
	Second edition, first impression, September 2025
Lay out and DTP:	Coco Bookmedia, Amersfoort – NL
Copyright:	© 2020, 2025 Van Haren Publishing

Trademark notices
TOGAF® and ArchiMate® are registered trademarks of The Open Group. All rights reserved.
DMBOK® is a registered trademark of DAMA International

Disclaimer
No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of the copyright owners

Specifically, without such written permission, the use or incorporation of this publication, in whole or in part, is **not permitted** for the purposes of training or developing large language models (LLM) or any other generative artificial intelligence systems. It is also not permitted for use in, or in connection with, such technologies, tools, or models to generate any data or content and/or to synthesize or combine it with any other data or content.

Foreword by Tony Shaw

I wonder if Bas van Gils had in mind the quote by Albert Einstein, that "everything should be made as simple as possible, but not simpler", because in this book you are about to read, he has created a "gentle introduction" which truly serves the purpose of explaining data and data management. Personally, when I first got into the world of data 20+ years ago, and coming from a background in marketing and business development, I had to learn about data management through the gradual osmosis of interacting with data professionals. While this is useful in understanding the "what" of real-world practice, it doesn't fill in the theoretical foundations of "how" and "why" which are necessary to understand why that real-world practice works the way it does. I know I would have come up to speed a whole lot faster, if I'd had access to this book.

One of the big themes in corporate data today is data literacy, and as organizations strive to become more data-driven, then it's a theme that will only grow in relevance. Data is not a trend that's going to flame out in a few years, so just like financial literacy and human capital management, it is now obvious that data literacy is going to be a critical knowledge requirement for all managers and executives in the future. As such, we should be thinking about data education in the same way we think about financial and HR education, building the foundations in schools and universities, then continuing to apply those foundations to practical experience through employee onboarding programs, and broader corporate training.

This book serves these objectives well. All the important enterprise-level data management topics are included. It serves as a valuable curriculum for someone just starting out in a professional data career, or indeed for someone who like me, who picked up bits and pieces without much structure to my learning. Bas's explanations are clear, and build upon each other systematically. I personally appreciate the research that has gone into identifying the clearest definitions available, even when that means quoting other sources. Bas has effectively curated the "best of" from

existing industry literature, and tied everything together into a consistent whole, through his own lucid insight, analysis and explanations.

I wish you, the reader, well whether this is the start of your data management journey, or like me, you are finding structure for your fragmented knowledge. You have found an excellent resource to help you fulfill your objectives.

Tony Shaw, CEO & Founder of Dataversity
October 2019

Foreword by Hans Weigand

"Language (die Sprache) is always a mediator", the famous Von Humboldt wrote 200 years ago. "It is between the finite and the infinite", he continues, "and at the same time between one individual and the other". In traditional philosophical categories: as a subject-object relator and a subject-subject relator. That Von Humboldt spoke using the terms finite and infinite says something about his view of the human subject (its finiteness, in several respects). It is important to note that when Von Humboldt calls language a mediator, he explicitly wants to say that the two things that get mediated do not exist independently of each other, but that in a way they come into existence through the mediation. The mediator is more than a formal relationship. That is why for him language is not a coding system where an (arbitrary) sign is determined for something that already exists for us. Such a coding system does not *make* language, it *presupposes* language.

To some extent, the characterization of Von Humboldt for language can also be applied to data, the subject of this book. Yes, the formal data structures in a computer have been designed, so as such they are not language in the Von Humboldt sense. Still, they draw on language and so take over some of its characteristics. Data also mediates between subjects. This is one reason why data needs to be protected, as identified in chapters 17 and 21 of this book, and why "shared understanding" is a fundamental goal. It is also mediating with an infinite world around us. To use a phrase of Bas, "data codifies what we know about the world". At another place, data is defined as the combination of fact and meaning. If this is true (and who am I am to question Bas?), it means that managing data has two rather different faces. Because managing facts, as stored in files on a disk, is quite different from managing such an intangible thing as "meaning". I don't want to push this point too much, but I think here is one reason why data management is not simple and not comparable to the management of physical assets such as vehicles or library books, in spite of some similarities.

When data is a mediator, it also runs the risks of the fate of the mediator: always to fall in between. So that neither the IT department nor the business unit cares for it; that there is no budget for it. That it is seen as instrumental only and so is not a genuine concern in its own right. In the short history of IT so far we have learned that this would a big mistake. Data needs to be recognized as an asset and needs to be managed. Not as a goal in its own of course – a point that is stressed by Bas several times in this book. It remains a mediator, but still, it needs to be managed properly. Therefore, I am glad with this book that takes data management seriously. A book that tries to integrate insights on data management from theory and practice. A book that can not only serve practitioners and companies that struggle with data management but that can also be a good reference text for academic courses in the field of Information Management or Data Science. I wish it all the best!

Dr. Hans Weigand, Associate Professor Information Systems, Tilburg University
October 2019

Preface

When I started my studies at Tilburg University in 1998, one of the first things that I learned was an appreciation for the 'golden triangle' of processes, data, and systems. Only through careful alignment of these three can organizations function well. It was interesting to see that so many people – academics and professionals alike – worried mostly about either *systems* or *processes*, while *data* appeared to take the back seat.

After my studies, I started working on my dissertation at Nijmegen University. The focus of my research was *Web information retrieval*. The main idea behind my research was based on economic principles: if you have *demand* and *supply* of data, then all you have to do is "match" the two. How hard can that be? After all, the topic of information retrieval had been studied for decades. Let's just say that I learned a lot in those days, not just about the *information needs* of people surfing the Internet, but also about semantics, data modeling, data structures, etc.

Since then, I have worked in many different roles, from IT professional to strategy consultant and pretty much every role in between. Over the years, I noticed that *data* was becoming an increasingly important topic. People started to recognize that mishandling data was costing the organization in missed opportunities, rework, reputational damage, etc. and that products and services could be greatly enhanced when enriched with data. Around this time, people started talking about data as "the new oil" and recognized it for the valuable asset that it really was. This was further strengthened by the apparent rise of topics such as *artificial intelligence*, *data science*, and *big data*.

I started studying *data management* in earnest around 2008. A few years later, Tanja Glisin suggested I study the DAMA DMBOK® [MBEH09] which really opened my eyes to the depth and breadth of the field. I found that the DMBOK was *the* reference within our field at the time, especially when complemented with other – more in-depth – publications. The second version of the DMBOK was published in

2017 and showed the significant improvement of our knowledge of the field [Hen17]. I have used both versions of the DMBOK over the years, both as a reference during consultancy assignments and teaching.

The DMBOK is a great reference, but may practitioners find it too theoretical to be of practical use. A more *pragmatic* book that combines theory with practical recommendations is missing. After much debate and discussions with friends, many of whom I have interviewed for this book, I decided to attempt to fill this gap.

The decision to actually move forward with the writing project was made in March of 2019, while visiting the Enterprise Data World conference in Boston, Massachusetts. I wrote the first version of the book during the summer months of 2019 and am forever grateful for all the support and help I received. A few years later, I wrote my second book on data management [Gil23]. That publication picked up where this book leaves off. It also takes the DMBOK as a basis but goes much deeper. One could say that the *Gentle Introduction* is more pragmatic whereas *Data in Context* is more theoretical in nature. In the fall of 2024, I decided it was time for an update of the *Gentle Introduction*. Life happened (several challenges in the family) and caused the update to take a bit longer than expected. Still, we got it done and this new edition will provide the reader with more up-to-date insights.

For the update, I adopted the following strategy. I went through each of the chapters individually and asked myself two questions: (1) In teaching/speaking about this topic, have I received any feedback that I should process? (2) Have I learned something new that requires me to update the material? Somewhat surprisingly, most of the material still seems very relevant and up-to-date. This is the result of the choice to stay away from specific technologies and focus on core concepts. All in all, I did feel the need to add several topics, include some new interviews, and make some (small) changes.

There are so many people to thank, and I sincerely hope I am not forgetting anyone. First of all, I would like to thank my colleagues at Strategy Alliance for their patience and help in preparing the manuscript. I would also like to thank Maurits van der Plas, Ivo van Haren, and Bart Verbrugge of Van Haren Publishing: I know that I have strong opinions on how/what I want with the book - and I have probably tried your patience over and over.

The book wouldn't have been nearly as good without the help of Lisa Gaudette. She is my rock and "language hero". Thank you so much for your patience, hard work, and grammar/punctuation lessons. Whenever I thought we had cleaned up a piece of text, you always found more ways to make it better. I would also like to thank Mirjam Visser for her extensive review of the first version of the manuscript as well as the pleasant discussions we had on data management. My colleagues at

both Antwerp Management School, Strategy Alliance, and DAMA Netherlands also deserve a big thank you: writing is an intensive process, and I know I have been busier than normal over the last few months. So, thank you for your patience and help! Last but not least, I would like to thank my family for their support. I know I have been hiding behind my computer to finish the manuscript and wouldn't have been able to make so much progress without your flexibility and support.

Regarding the interviews and intermezzos in this book, I want to mention that some respondents have changed jobs since the first version of this book. On the one hand, I was tempted to change the roles to reflect their new positions. On the other hand, it seemed better to keep the original roles since those capture the context of the interviews best. I decided to go with the latter.

As a final remark, I would like to point out that a lot of time and effort went into checking the material. Any errors that remain are my own. I hope you find the book interesting and useful. Enjoy the read!

Bas van Gils
August 2025

Contents

1 INTRODUCTION 1

1.1 Goals for this book 4
1.2 Intended audience 4
1.3 Approach 5

2 DATA AS AN ASSET 6

2.1 Data 6
2.2 Asset 8
2.3 Data and process 9
2.4 Visual summary 11

3 DATA MANAGEMENT: WHY BOTHER? 12

3.1 A definition of data management 12
3.2 Value of DM 13
3.3 Key challenges for DM 15
3.4 Visual summary 16

4 POSITIONING DATA MANAGEMENT 17

4.1 The center of the universe 17
4.2 DM and business process management 18
4.3 DM and IT management 19
4.4 Information/data analysis 20
4.5 Database management 21
4.6 DM and enterprise architecture management 22
4.7 Philosophical considerations 24
4.8 Visual summary 27

PART I: THEORY

5 INTRODUCTION 31

6 TERMINOLOGY 32

6.1 Introduction 32
6.2 Data codifies what we know about the world 33
6.3 Storing data in systems 34
6.4 Data in processes 35
6.5 Connecting the business and IT perspective 37
6.6 Outlook 38
6.7 Visual summary 39

7 DATA MANAGEMENT: A DEFINITION 40

7.1 Introduction 40
7.2 Managing the lifecycle of data 42
7.3 Deconstructing DM 43
7.4 Visual summary 46

8 TYPES OF DATA 47

8.1 Classifying data 47
8.2 Five fundamentally different types of data 48
8.3 Transaction data 48
8.4 Master data 49
8.5 Business intelligence data 50
8.6 Reference data 50
8.7 Metadata 51
8.8 Visual summary 52

9 DATA GOVERNANCE 53

9.1 Introduction 53
9.2 Data governance and data management 54
9.3 Data governance activities in DMBOK® 55
9.4 A modern approach to data governance 56
9.5 Position of data governance 58
9.6 Visual summary 59

10 METADATA 60

10.1 Types of metadata 60
10.1.1 Business metadata 60
10.1.2 Technical metadata 62
10.1.3 Operational metadata 62
10.2 Metadata is the foundation 64
10.3 Metadata repositories 64
10.4 Visual summary 66

11 MODELING 67

11.1 Scope 68
11.2 Abstraction levels 68
11.3 Modeling languages 72
11.3.1 Fact-based modeling 72
11.3.2 Entity relationship modeling 73
11.3.3 Architecture modeling with ArchiMate 73
11.4 Relationship to other DM capabilities 75
11.5 Visual summary 76

12 ARCHITECTURE 77

12.1 Architecture 77
12.2 Data architecture 80
12.3 Relationship to other (data management) capabilities 83
12.4 Visual summary 84

13 INTEGRATION 85

13.1 Introduction to data integration 85
13.2 Common integration patterns 87
13.2.1 Batch integration 87
13.2.2 Accessing data through services 87
13.2.3 Change data capture 88
13.2.4 Streaming data integration 88
13.2.5 Data virtualization 89
13.3 Integration from an architecture perspective 90
13.3.1 Dealing with the number of potential connections 91
13.3.2 Dealing with different names and structures 91
13.3.3 Dealing with different patterns 92
13.4 Data mesh 92
13.5 Visual summary 95

14 REFERENCE DATA 96

14.1 Definition 96
14.2 Using reference data to harmonize the meaning of data 98
14.3 Historic versions of reference data sets 98
14.4 Reference data and governance 99
14.5 Visual summary 100

15 MASTER DATA 101

15.1 Multiple versions of the truth 101
15.2 Basic MDM concepts 103
15.3 Relationship to other data management capabilities 106
15.4 Visual summary 107

16 QUALITY 108

16.1 Introduction 108
16.2 The notion of quality 108
16.3 Data quality 109
16.4 Data quality management 112
16.5 Critical data elements 114
16.6 Relationship to other capabilities 115
16.7 Visual summary 116

17 DOCUMENT AND CONTENT MANAGEMENT 117

17.1 Characteristics of documents 118
17.2 Lifecycle and archives 119
17.2.1 Documents, originals and copies 119
17.2.3 Archives: authenticity and proof 120
17.2.4 Records continuum model 120
17.2.5 Implications 121
17.3 Other document collections 122
17.3 Visual summary 123

18 RISK AND SECURITY 124

18.1 Risks and risk mitigating measures 124
18.2 ISO standards 126
18.3 Data security management 127
18.4 Training and certification 130
18.5 Relationship to other capabilities 130
18.6 Visual summary 133

19 BUSINESS INTELLIGENCE & ANALYTICS 134

19.1 Defining business intelligence and analytics 134
19.2 Common system types 136
19.3 Structuring data 137
19.4 Self-service BI 140
19.5 Relationship to other capabilities 142
19.6 Visual summary 143

20 DATA SCIENCE & AI 144

20.1 Algorithms 145
20.2 Data science 146
20.3 ARtificial intelligence 148
20.4 Offense and defense 149
20.5 Visual summary 150

21 TECHNOLOGY 151

21.1 People are key 151
21.2 Observations about technology 153
21.3 Technology and the functional areas of DMBOK® 155
21.3.1 Data governance and stewardship 155
21.3.2 Metadata 155
21.3.3 Modeling 156
21.3.4 Architecture 156
21.3.5 Integration 157
21.3.6 Reference and master data 157
21.3.7 Quality 158
21.3.8 Security 158
21.3.9 Business intelligence 159
21.3.10 Big data 159
21.4 Technology adoption 159

22 DATA (HANDLING) ETHICS & COMPLIANCE 161

22.1 Ethics in data 161
22.2 Ethical handling of data 162
22.2.1 Ethical principles behind data protection 163
22.2.2 The data lifecycle 164
22.2.3 Using ethical principles in the data lifecycle 165
22.3 The relationship between ethics and governance 166
22.4 Visual summary 168

PART II: PRACTICE

23 INTRODUCTION 171

24 BUILDING THE BUSINESS CASE FOR DATA MANAGEMENT 173

24.1 The need for a business case 173
24.2 Qualitative and quantitative business case 175
24.3 Incremental approach to building a business case 178

25 KICK-STARTING DATA QUALITY MANAGEMENT 180

25.1 Top-down approach 180
25.2 A motivation for starting small 181
25.3 Setting up your first experiments with data quality management 181
25.4 Scaling up after successful experimentation 184

26 FINDING DATA OWNERS AND DATA STEWARDS 186

26.1 Top-down and bottom-up 186
26.2 Ownership/stewardship models 188
26.3 Finding owners and stewards 190

27 THE ROLE OF TRAINING 192

27.1 People first, and the need for training 192
27.2 Types of training 194
27.3 How to design a training program 195

28 SETTING UP A DATA MANAGEMENT POLICY 198

28.1 Data management policy 198
28.2 Typical structure for a data management policy 200
28.3 Setting up a data management policy 202
28.3.1 Top-down 202
28.3.2 Bottom-up 204
28.4 Recommendations 205

29 BUSINESS CONCEPTS AND THE CONCEPTUAL DATA MODEL 206

29.1 Freezing language 206
29.2 Definitions and conceptual data models 208
29.3 Definitions in a context 210
29.4 Recommendations 212

30 SETTING UP A METADATA REPOSITORY 214

30.1 The importance of metadata 214
30.2 Metadata repository architectures 215
30.3 Implementation strategies 218
30.3.1 Top-down metadata strategy 218
30.3.2 Bottom-up metadata strategy 218
30.3.3 Matching the strategy to the situation 219
30.4 Recommendations 220

31 LEVERAGING ENTERPRISE ARCHITECTURE 221

31.1 EA as a source of information 221
31.2 EA models and visualizations 223
31.3 Building effective solutions 225
31.4 Recommendations 227

32 INTEGRATION ARCHITECTURE 228

32.1 Data is everywhere 228
32.2 Start simple 229
32.3 Keep it simple 232
32.4 Recommendations 234

33 A PRAGMATIC APPROACH TO DATA SECURITY 235

33.1 Motivation for a security framework 235
33.2 Security use cases 237
33.3 Security levels in business terms 238
33.4 The link to security measures and controls 240
33.5 Tying it together 240

34 ROLES IN DATA MANAGEMENT 242

34.1 Change and run 242
34.2 Roles in the DMBOK 244
34.3 Skills in the SFIA framework 244

34.4 Definition of roles 246
34.4.1 Architect 246
34.4.2 Business management 247
34.4.3 Data owner, data steward 248
34.4.4 Project management 248
34.4.5 Chief data officer 249
34.4.6 Business analyst, process analyst, and system analyst 249
34.5 Reflection and recommendation 250

35 BUILDING A DATA MANAGEMENT ROADMAP 251

35.1 To roadmap or not to roadmap 251
35.2 The steps towards an effective roadmap 252
35.3 Techniques 255
35.3.1 Vision phase 255
35.3.2 Analysis phase 256
35.3.3 Portfolio phase 257
35.3.4 Execution phase 257
35.4 Recommendations 258

PART III: CLOSING REMARKS

36 SYNTHESIS OF THE RECOMMENDATIONS 261

36.1 Data management 261
36.2 Antifragility and complexity 263
36.3 Expected benefits 265

37 CONCLUSION 268

37.1 Review 268
37.2 Outlook 269
37.3 Call to action 271

BIBLIOGRAPHY 273

INDEX 281

ABOUT THE AUTHOR 285

1 Introduction

It is often said that "data is the new oil". It is hard to figure out with any certainty who wrote about this metaphor first. A cursory search on Google suggests it was used originally in an article by The Economist [Par17] with many authors following suit by describing why, for all practical reasons, data is *not* the new oil (e.g. [Mar18]). Whatever the practical implications, the metaphor at least illustrates that data is an important business asset that deserves to be managed as such. This is the field of data management (or DM for short). See also sidebar 1.

Sidebar 1. Interview with Marco van der Winden (Summer 2019)

My experience is that the importance of data is underestimated in the way that there was/is no primary focus on it. Living in the low countries where there is an abundance of water, data is mostly seen as something that can be easily obtained, just like water. To continue the comparison, the Dutch are very good with containing the water streams and keeping the seawater outside with dikes. But with data we are less experienced. We let data sometimes uncontrollably flow though our fields without knowing where it goes or even why we are doing it.

We are not in the Middle Ages (when we became increasingly proficient at water management) and it should be clear that data must be governed in a way that we are more in control and that we can profit more from it. By the way, I think that a comparison with oil is not a smart one. Sooner or later there will be a shortage of oil. Above that, there are also some environmental disadvantages with oil. Data is more like water. It's the source of all living things. You can't live without it and there will always be water.

Marco van der Winden is manager of the corporate data management office at PGGM, a Dutch pension provider.

A key question that needs answering is: what does that entail? In other words: what is data management (DM) and how do you make it work? These are hard questions. Data is often seen as an abstract "thing" that sits in the realm of the IT department. This isn't helped by the fact that a lot of technology is so closely related to data that it is easy to confuse one for the other. Worse, data management professionals are prone to using complicated terminology such as *metadata*, *master data*, *lineage* and so on, which makes it hard for outsiders to truly understand what is going on. This is not a good thing: DM is an important capability that organizations must master[1].

Years later, I interviewed Marco again. I asked him for his latest insights on the same question as a few years earlier: what are your views on data/introducing data management in the organization. His responses can be found in sidebar 2.

Sidebar 2. Interview with Marco van der Winden (2)

The importance of good data that is made available quickly (at the right time for the right stakeholders) has increased further in recent years. The comparison with water, which is essential for human life, is apt: the same thing applies to a healthy business operation, where data is the water that flows through the organization. Establishing a solid data infrastructure is no easy task as it involves a combination of introducing new data technologies and the agility (of employees and management) within an organization to adapt and change.

Investing in your data infrastructure requires a long-term vision—not only in terms of what you want to achieve with your data to meet your business goals but also on what needs to be done to structure your data infrastructure accordingly. For the first goal, the offensive side, you can usually gain broad support. After all, for part of the organization, it is quite exciting to imagine all the things that could be done with data in the future. It stimulates the entrepreneurial spirit within the company. However, the fact that this also requires significant investment in technology and employees is less appealing.

This investment in the defensive side of data management often demands more effort than the offensive side. It means spending large sums on new data applications and, more importantly, dedicating a lot of time to changing the "way of working" or, in other words, increasing data literacy within the organization. The realization that substantial investment in the defensive side is necessary to enable the offensive side — in other words, to achieve your business strategy — is essential. You cannot harvest fruit from a tree that you do not water.

1 Throughout this book, I will use the term *capability* to signify an ability/ discipline that an organization may have. The simple formula capability = capacity × ability further signifies that the organization not only has to master the ability, but also have sufficient resources with the right abilities available in order to be successful.