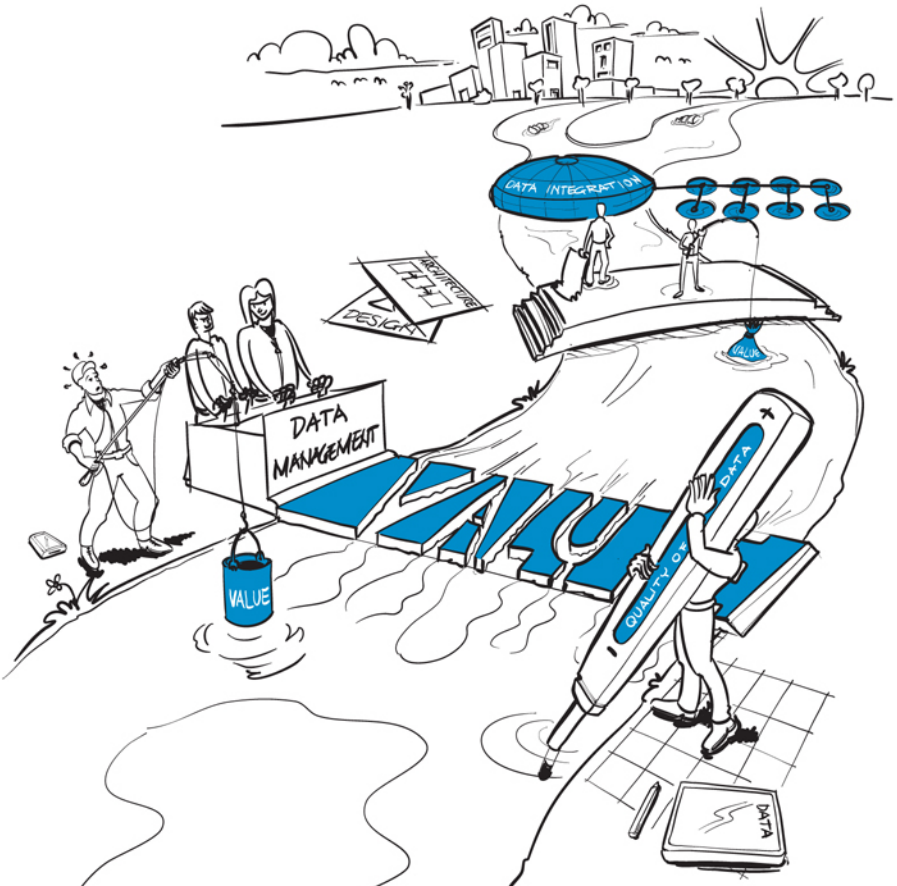
# DATA MANAGEMENT
## A GENTLE INTRODUCTION

Balancing theory and practice

Bas van Gils




Van Haren PUBLISHING

Data Management: a gentle introduction

## Other publications by Van Haren Publishing

Van Haren Publishing (VHP) specializes in titles on Best Practices, methods and standards within four domains:

- IT and IT Management
- Architecture (Enterprise and IT)
- Business Management and
- Project Management

Van Haren Publishing is also publishing on behalf of leading organizations and companies: ASLBiSL Foundation, BRMI, CA, Centre Henri Tudor, CM Partners, Gaming Works, IACCM, IAOP, IFDC, Innovation Value Institute, IPMA-NL, ITSqc, NAF, KNVI, PMI-NL, PON, The Open Group, The SOX Institute.

Topics are (per domain):

### IT and IT Management

ABC of ICT
ASL®
CMMI®
COBIT®
e-CF
ISO/IEC 20000
ISO/IEC 27001/27002
ISPL
IT4IT®
IT-CMF™
IT Service CMM
ITIL®
MOF
MSF
SABSA
SAF
SIAM™
TRIM
VeriSM™

### Enterprise Architecture

ArchiMate®
GEA®
Novius Architectuur
Methode
TOGAF®

### Business Management

*BABOK® Guide*
BiSL® and BiSL® Next
BRMBOK™
BTF
CATS CM®
EFQM
eSCM
IACCM
ISA-95
ISO 9000/9001
OPBOK
SixSigma
SOX
SqEME®

### Project Management

A4-Projectmanagement
DSDM/Atern
ICB / NCB
ISO 21500
MINCE®
M_o_R®
MSP®
P3O®
*PMBOK® Guide*
Praxis®
PRINCE2®

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

# Data Management:
# a gentle introduction

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 |
| 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 editor: | Lisa Gaudette and Steve Newton (Galatea) |
| Publisher: | Van Haren Publishing, 's-Hertogenbosch |
| | |
| ISBN hard copy: | 978 94 018 0550 6 |
| ISBN eBook (pdf): | 978 94 018 0552 0 |
| ISBN ePUB: | 978 94 018 0555 1 |
| | |
| Edition: | First edition, first impression, February 2020 |
| Lay out and DTP | Coco Bookmedia, Amersfoort – NL |
| Copyright: | © 2020 Van Haren Publishing |

# 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 dataliteracy, 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. 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. Then, of course, there are the people who graciously granted me interviews to use in this book – you are all heroes:

- *Marco van der Winden is manager of the corporate data management office at* PGGM*, a Dutch pension provider.*
- *Marc van den Berg is managing director of* IT *and Innovation at* PGGM*, a Dutch pension provider.*
- *Frank Harmsen is managing director at* PNA *and professor at Maastricht University.*
- *Lisa Gaudette is director in the Office of Sponsored Programs and Research of Clark University.*
- *Jan Robat is head of data quality management at* ABN AMRO.
- *Fanny Vuillemin is senior data manager at* AXA.
- *Céline Lescop is lead data architect at* AXA.
- *Piethein Strengholt is principle data architect at* ABN AMRO.
- *Eric D. Schabell is global technology evangelist and portfolio architect director at Red Hat.*
- *Tanja Glisin is an experienced data management professional and frequent collaborator of the author of this book.*
- *Norbert van de Ven is data governance consultant at Hot ITem.*
- *Stijn Hoppenbrouwers is professor of Data & Knowledge Engineering at* HAN *University of Applied Sciences, Arnhem and assistant professor at Radboud University Nijmegen.*
- *Jeroen Cloo is partner at Novius Adviesgroep.*
- *Kiean Bitaraf is data management consultant at Deloitte.*
- *Raymond Slot is managing partner at Strategy Alliance.*

- *Paul Heisen is senior enterprise architect at De Lage Landen (*DLL*).*
- *Robin Vuyk is head of business architecture and design at* PGGM*, a Dutch pension provider.*
- *Daan Riepma is a smart data consultant at Axians.*
- *Ronald Damhof, "just a data-guy", self-employed, often in the role of enterprise (data) architect in large (mostly public) organizations.*

The book wouldn't have been nearly as good without the help of Lisa Gaudette. 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 manuscript as well as the pleasant discussions we had on data management. 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.

As a last 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
October 2019

# Contents

## PART I: THEORY

## PART II: PRACTICE

# List of figures

# 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 be 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].

To illustrate this point, I will borrow a slightly altered example from [Soa11] in example 1.

**Example 1. Data management benefits**
Assume you are working for a large global company with approximately 10 million customers. On average each customer purchases 1.2 products every year. Your strategy is to attempt to get more revenue from the existing customer base, rather than try to capture a bigger market share. To that end, a global *customer 360* initiative is considered. The data management team and marketing have worked together to compile a business case.

First, it is expected that a better overview of each customer will increase the number of purchases from 1.2 to 1.4, which is expected to raise an extra 8 million dollars in revenues over three years. Furthermore, it is estimated that the direct cost of wading through duplicated/ inconsistent data about customers by customer service representatives adds up to about half a million dollars over three years. The direct cost of the IT department around data integration issues is expected to be reduced by another half a million dollars over three years. This adds up to nine million dollars in benefits. Would that justify a significant investment in data management?

## 1.1 GOALS FOR THIS BOOK

One of the best ways to make progress in our field is to put knowledge in the public domain such that everyone can benefit from it. There are many ways to do this: scientific studies provide academic rigor but tend to be low on practical relevance. Handbooks such as the DMBOK[2] are the inverse: there is a lot of practical value but they tend to be low on the academic rigor [Hen17]. Balancing rigor and relevance is tricky to say the least. This book leans towards the practical relevance side and provides academic rigor whenever possible. The unique selling point of this book will lie in the fact that it offers (1) an up-to-date overview of the field, (2) with practical guidance in the form of a capability-based framework, and (3) is supported by real-world evidence through mini case studies.

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.
2 The DMBOK is the Data Management Body of Knowledge. It is a reference book by DAMA, the Data Management Association. The DMBOK compiles data management principles and best practices.