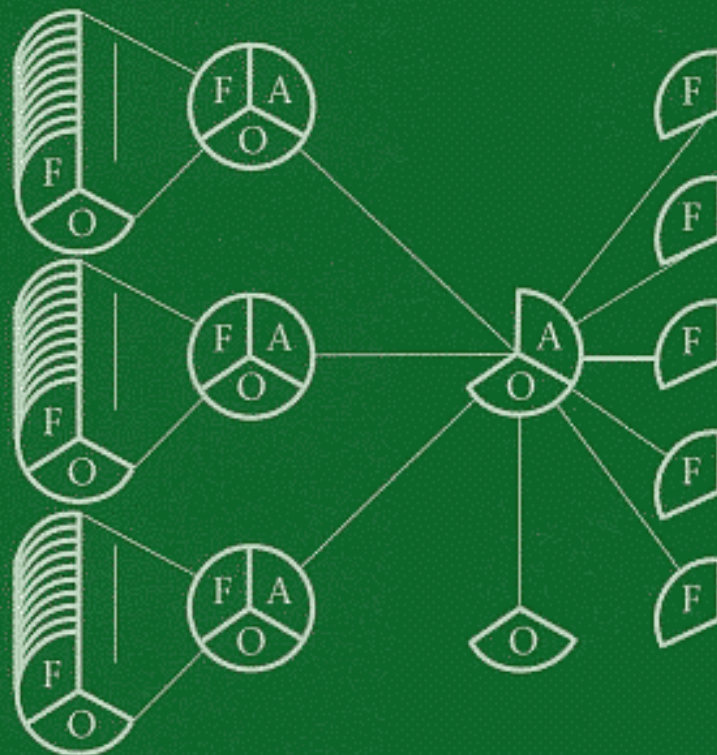


Management of distributed data in distributed environments



R.J. Mersel

Management of distributed data in distributed environments

Management of distributed data in distributed environments

Proefschrift

Ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft
op gezag van de Rector Magnificus,
prof.ir. K.F. Wakker,
in het openbaar te verdedigen
ten overstaan van een commissie
aangewezen door het College van Dekanen
op 28 november 1995 te 10.30 uur

door

Robertus Johannes MERSEL

informatica ingenieur
geboren te Rotterdam

Dit proefschrift is goedgekeurd door de promotor:

Prof.dr.ir. M. Looijen

Samenstelling promotiecommissie:

Rector Magnificus

Prof.dr.ir. M. Looijen

Prof.drs. B.K. Brussaard

Prof.Dr.-Ing.habil. W. Gerhardt-Häckl

Prof.drs. J.A.M. Oonincx R.A.

Prof.dr.ir. P. van der Veer

Prof.dr.ir. J.C. Wortmann

voorzitter

Technische Universiteit Delft, promotor

Technische Universiteit Delft

Technische Universiteit Delft

Koninklijke Militaire Academie

Technische Universiteit Delft

Technische Universiteit Eindhoven

For Shehrazade

Published and distributed by:

R.J. Mersel
Delft University of Technology
Faculty of Technical Mathematics and Informatics
Department of Information Systems
Julianalaan 132
P.O. Box 356
2600 AJ Delft, The Netherlands
Tel: +31 15 278 44 75
Fax: +31 15 278 66 32
Email: secr@is.twi.tudelft.nl

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Mersel, Robertus Johannes

Management of distributed data in distributed environments /
Robertus Johannes Mersel. - [S.l.: s.n.]. - III.
Proefschrift Technische Universiteit Delft. - Met index, lit. opg.
ISBN 90-9008870-9
NUGI 855/684
Trefw.: gegevensbeheer, gedistribueerde gegevens.

Omslagontwerp: R.J. Mersel / J. Schonewille

© 1995, R.J. Mersel

All rights reserved. No part of this publication may be reproduced in any form or by any means without prior written permission of the publisher.

Preface

Over the years much attention has been paid to methods and techniques for the development of information systems, however, methods and techniques for information systems management have not received much attention. Consequently, the practice of information systems management is characterized by the ad hoc carrying out of management tasks and a lack of coherence. In the late 1980s the attention of organizations began to shift from information systems development towards information systems management. At the same time and as a result of this a new research group ‘management of information systems’ was established at the Delft University of Technology.

This thesis describes a research project that was conducted within the group ‘management of information systems’ during the period of 1991 through 1995. The research aimed at providing an instrumentation to be used by personnel at the operational level of an organization for the management (i.e. the operation, control and maintenance, see definition 2) of data to enable them to perform their management tasks for data in the states utilization (U) and exploitation (E) of the state model for information systems. The instrumentation also provides data management personnel at the tactical level with a means to diagnosis and to provide directions for the operational level of data management. The position of this research project in the research programme on management of information systems and its aims are discussed in chapter 1.

The lack of methods and techniques has lead, over the years, to a large number of unrelated concepts and to a confusion of concepts regarding data management, as is demonstrated in chapter 2. The title of this thesis could therefore be the subject of misunderstanding. Distribution is defined in this thesis in two distinct types: deconcentration which concerns the geographical locations, and decentralization which concerns the organizational units, see section 1.1. The environment is defined in this thesis in terms of the interested parties (the stakeholders) that have an interest in the functionalities that are to be performed by the data. The data are defined in this thesis in terms of their characteristics, and data management is defined in this thesis in terms of management tasks, see chapter 4.

The lack of methods and techniques has also lead to an inability of data management organizations to relate all the aspects of data management, as is demonstrated in chapter 3. The instrumentation specified in chapters 4 and 5 of this thesis consists

viii

of two components: a framework that provides a comprehensive definition of all the concepts relevant for data management; and several methods and techniques to be used by personnel at the operational level of the data management organization.

This instrumentation is tested for practical applicability in three case studies where it is used as a diagnostic instrument to assess the data management organization. The conclusions show practical applicability of the instrumentation in present data management organisation, however, they also show possible improvements and extensions that can be made to the instrumentation.

In realizing of this thesis I am indebted to many people. In the first place I wish to thank Maarten Looijen who never ceased to spur me on to achieve the utmost, and who helped me whenever I reached an impasse in my research. Many thanks to my colleagues from the Department of Information Systems who provided a willing ear whenever I needed one, and to Cees de Wijs and Louis van Hemmen for the many nights we spent in exploring the uncharted regions of information systems management. I would like to thank Miranda Aldham-Breary in particular for making this thesis more readable. Without her help this thesis would have been written in a language that was almost but not completely unlike English.

The participation of organizations is essential when performing empirical research to assure the practical applicability of a theory. Without the support of the Joint Office for Social Security Administration (GAK) there would not have been a thesis. I would like to thank in particular Wim Kuipers, Simon Bosma and Jacob de Boer, Thiel Chang and all the others at the Automation and Information department who helped. Edi de Vries from the Dutch Ministry of Defence gave insight into practical difficulties of data management during the early stages of the research. Thanks also to Shell Moerdijk who were willing to serve as testing ground for my theories.

I would further like to thank the students I had the pleasure to work with and who have helped me during this research project. Peter Zeldenrust was there in the beginning and performed his graduation project at the Dutch Ministry of Defence. Wilma Notermans, from the Hogeschool Heerlen, applied the Analysis of Change of the ISAC method to data management of the (legislative) information system AAW-conveniences. She proved all academic misconceptions of B.Sc.-students to be wrong. Anton van den Bogaart from Nijmegen University provided a good discussion on the consequences of decentralization for the organization of data management. Bart Zonneveld did a '*superagrooalisticprosifunkstication*' graduation project for replacement of the access management information system (TBS) at GAK. Arjen Groen helped me with the unemployment benefit (WWO) case during his graduation project in which he mastered my definitions of data management, and my methods and techniques before they had the concise definitions they have in this thesis.

Most of all I want to thank my wife Shehrazade, I don't need to tell her why.

Delft, 27th February 1997

Rob Mersel

Contents

1	Introduction	19
1.1	Background	19
1.2	Research programme	20
1.2.1	Defining management of information systems	20
1.2.2	State model	22
1.2.3	Three forms of management	24
1.3	Placing this thesis within the research programme	25
1.3.1	Placement within the definition	25
1.3.2	Placement within the state model	26
1.3.3	Placement within the three forms of management	27
1.4	Research approach	27
1.4.1	Research Objectives	28
1.4.2	Literature	31
1.4.3	Practice	31
1.4.4	Research questions	31
1.4.5	Management instrumentation	32
1.4.6	Case study design	33
1.4.7	Case execution	34
1.4.8	Case results	36
1.4.9	Research results	36
1.5	Outline of this thesis	37
2	Environment, data and management: literature	39
2.1	Environment	40
2.1.1	Relationship between data and the real system	40
2.1.2	Administrative integration	42
2.1.3	Owner and custodian of data	44
2.1.4	Distribution of the environment	45
2.2	Data	47
2.2.1	Data viewed as an information system component	47
2.2.2	Data infrastructures	48

2.2.3	Requirements on data	50
2.2.4	Distribution of data	56
2.3	Management	65
2.3.1	Strategic level	66
2.3.2	Tactical level	66
2.3.3	Operational level	67
2.3.4	Distribution of management	75
2.4	Conclusions	79
2.4.1	Conclusions environment	80
2.4.2	Conclusions data	80
2.4.3	Conclusions management	80
2.4.4	Current literature on management of distributed data in distributed environments	81
2.4.5	Directions for this thesis	81
3	Environment, data and management: practice	83
3.1	Environment	84
3.1.1	Ministry of Defence (CO)	84
3.1.2	GAK	88
3.2	Data	91
3.2.1	Ministry of Defence (CO)	91
3.2.2	GAK	92
3.3	Management	98
3.3.1	Ministry of Defence (CO)	98
3.3.2	GAK	99
3.4	Conclusions	101
3.4.1	Conclusions Ministry of Defence (CO)	102
3.4.2	Conclusions GAK	103
3.4.3	Current practice of management of distributed data of distributed environments	104
3.4.4	Directions for this thesis	104
4	The FATO framework	107
4.1	Functionalities	109
4.1.1	Pragmatics	109
4.1.2	Semantics	111
4.2	Automation facilities	114
4.2.1	Syntactics	114
4.2.2	Empirics	118
4.3	Task fields	122
4.3.1	Actual data management	122
4.3.2	Data administration	123
4.3.3	Database administration	124

CONTENTS

xi

4.3.4	Storage management	125
4.4	Organization	126
4.4.1	External influences	126
4.4.2	Tactical and strategic task fields	129
4.4.3	Functional management and technical management	131
4.5	Positioning distribution within the FATO framework	132
4.5.1	Functional distribution	133
4.5.2	Technical distribution	133
4.5.3	Organizational distribution	134
5	Methods and techniques	135
5.1	Adaptations required for management	136
5.2	Environment	137
5.2.1	List of functionalities	137
5.2.2	List of stakeholders	138
5.2.3	Functionality matrix	138
5.2.4	Create/Use matrix	138
5.3	Data	138
5.3.1	Semantic model	139
5.3.2	Automation facilities matrix	139
5.3.3	Table of properties	139
5.3.4	Table of stakeholders	140
5.4	Management	140
5.4.1	Activity model	140
5.4.2	Table of RAWE distribution	141
5.4.3	Coordinating mechanisms	141
5.5	Positioning within the FATO framework	142
5.6	Testing the management instrumentation	143
6	Personal Data Register Case	145
6.1	Positioning the case	146
6.1.1	Deconcentration	146
6.1.2	Decentralization	148
6.2	Environment	149
6.2.1	Stakeholders and functionalities	149
6.3	Data	159
6.3.1	Pragmatic level	159
6.3.2	Semantic level	163
6.3.3	Distribution principle	167
6.3.4	Syntactic level and empiric level	172
6.4	Management	174
6.4.1	Management tasks	174
6.4.2	Management units	176

6.5	Case results	177
6.5.1	Conclusions: Personal Data Register	178
6.5.2	Conclusions: Management Instrumentation	181
7	Unemployment Benefit Case	189
7.1	Positioning the case	190
7.1.1	Deconcentration	190
7.1.2	Decentralization	191
7.2	Environment	192
7.2.1	Stakeholders and functionalities	193
7.3	Data	197
7.3.1	Pragmatic level	197
7.3.2	Semantic level	199
7.3.3	Syntactic level	201
7.3.4	Empiric level	203
7.4	Management	204
7.4.1	Management tasks	204
7.4.2	Management units	208
7.5	Case results	209
7.5.1	Conclusions: Unemployment Benefit	210
7.5.2	Conclusions: Management Instrumentation	213
8	Moerdijk Process Database Case	221
8.1	Positioning the case	222
8.1.1	Deconcentration	222
8.1.2	Decentralization	223
8.2	Environment	224
8.2.1	Moerdijk systems architecture	225
8.2.2	Stakeholders and functionalities	226
8.3	Data	230
8.3.1	Pragmatic level	231
8.3.2	Semantic level	233
8.3.3	Syntactic level	234
8.3.4	Empiric level	235
8.4	Management	236
8.4.1	Management tasks	236
8.4.2	Management units	239
8.5	Case study results	240
8.5.1	Conclusions: Management Instrumentation	241

CONTENTS

xiii

9 Conclusion	249
9.1 Positioning the case studies	250
9.1.1 Deconcentration	251
9.1.2 Decentralization	252
9.2 Research findings	253
9.2.1 The FATO Framework	253
9.2.2 Methods and techniques	256
9.2.3 Applicability of the instrumentation	256
9.3 Answers to the research questions	257
9.3.1 Characteristics of data	257
9.3.2 Methods and techniques	259
9.4 Future research	260
9.4.1 Improving the instrumentation	260
9.4.2 Beyond the steady state	260
9.4.3 Internal influences	261
9.4.4 External influences	261
Samenvatting	281

List of Figures

1.1	The information systems management paradigm	21
1.2	State model	22
1.3	Extended state model	23
1.4	Three forms of Information Systems Management	24
1.5	Specialization of the management paradigm for management of data .	26
1.6	Overview of the research approach	29
1.7	Diagram for distribution of the environment (decentralization)	35
1.8	Diagram for distribution of data (deconcentration)	36
2.1	The value cycle	43
2.2	Linking value cycles into a value chain	44
2.3	Relationship between process databases and subject databases	50
2.4	Distributed database management systems implementation alternatives	63
2.5	Three schema and five schema architecture for databases	64
2.6	Typical organization chart of a data management department	77
2.7	Triangular relationship between environment, data and management .	79
3.1	Organizational diagram of Ministry of Defence	85
3.2	Organizational diagram of CO	86
3.3	Organizational diagram of GAK	89
3.4	Organizational diagram of the Social Insurance directorate	90
3.5	The implementation of FINA	93
3.6	Technical-infrastructure of the Social Insurance directorate	94
3.7	Data architecture of Social Insurances directorate	96
3.8	Implement of the data architecture	97
3.9	Triangular relationship between environment, data and management .	102
4.1	Storage hierarchy	119
4.2	External influences	127
4.3	Organizational positions of task fields	132
4.4	Diagram for distribution	133

6.1	Deconcentration of BRP	148
6.2	Decentralization of BRP	148
6.3	Semantic model of BRP	165
6.4	Distribution principle	171
6.5	Method used to obtain the case study results for the BRP	179
7.1	Deconcentration of WWO	192
7.2	Decentralization of WWO	192
7.3	Activity model: failure recovery	206
7.4	Revised activity model: failure recovery	207
7.5	Method used to obtain case results WWO	210
8.1	Deconcentration of MPD	224
8.2	Decentralization of MPD	224
8.3	Systems architecture at Moerdijk	226
8.4	Method used to obtain case study results MPD	242

List of Tables

2.1	Five types of decentralization	47
3.1	Overview of GAK	88
3.2	Use factor subject register BRP	98
4.1	The FATO framework	108
4.2	Characteristics of data	108
5.1	Methods and techniques positioned within FATO	143
6.1	List of stakeholders BRP	152
6.2	List of functionalities BRP	154
6.3	Functionality matrix BRP	157
6.4	Create/Use matrix BRP	158
6.5	Summary pragmatic characteristics	160
6.6	Table of properties auditability	160
6.7	Table of stakeholders pragmatics BRP	161
6.8	Summary semantic characteristics	164
6.9	Amount of data (without verification and justification data)	168
6.10	Allocated DASD storage capacity	173
6.11	Table of RAWE distribution BRP	178
7.1	List of stakeholders WWO	194
7.2	List of functionalities WWO	195
7.3	Functionality matrix WWO	196
7.4	Summary pragmatic characteristics	198
7.5	Summary semantic characteristics	200
7.6	Summary syntactic characteristics	202
7.7	Summary empiric characteristics	203
7.8	Table of RAWE distribution WWO	209
8.1	List of stakeholders MPD	228
8.2	List of functionalities MPD	229

8.3	Functionality matrix MPD	230
8.4	Table of RAWI distribution MPD	240

Chapter 1

Introduction

1.1 Background

Automation and information technology penetrate corporate enterprises and government ever deeper while automated information systems are used increasingly to control and improve products and services rendered by private corporations and public administration. This makes these organizations ever more dependent on their automated information systems. As a consequence of this dependency governments and corporate enterprises have increasingly higher requirements and preconditions for the proper operation of automated information systems to assure aspects such as availability, reliability and security.

Data are one of the components of information systems upon which requirements and preconditions are made in the light of efficient and effective management of an organization. The other components of automated information systems are hardware, software, people and procedures, each component has its own particular characteristics, chapter 4 describes the data characteristics that are important in this thesis. Data characteristics are determined initially during the policy and planning and development stages of the life cycle of an information system. When an information system is utilized and exploited, data will exhibit dynamic characteristics rather than static characteristics. Due to the variable character of data, management activities must be performed to make sure that at any time the data characteristics meet requirements and preconditions. Deconcentration and decentralization are two types of data distribution that influence data characteristics, as well as requirements and preconditions for data.

Deconcentration of data is defined within this thesis as the placement of data at several geographical locations within an organization. Deconcentration of data over several locations can be based on two principles [23, p. 293]. The first principle is called 'partitioning' where any given logical data object (say, a personnel record or a customer record) has exactly one physically stored representative, stored at

precisely one location. Thus leading to a situation where all the personnel records are stored at one location and all the customer records at another. The second principle is called 'replication' where any given logical data object can have several physical representatives that can be stored at several sites. In both types of deconcentration data characteristics will differ for every location. The characteristics will differ because the automation facilities that are used at locations differ, and because utilization at the locations also differs.

Decentralization of data is defined within this thesis as the placement of data at several units within an organization. In public administration and corporate enterprise organizational units are obtaining increasingly more autonomy in defining their own management and automated information systems. Local autonomy implies local responsibilities and authorizations over data. Local autonomy also implies that organizational units make specific requirements and preconditions for data suited to their own needs. In many cases data utilization goes beyond the borders of the organizational unit, and data are used throughout the whole organization, perhaps even outside the organization. A decentralized organization therefore makes a variety of requirements and preconditions for data that may be in conflict with each other.

The two types of distribution, when combined, lead to a complex situation where it is difficult to maintain data in accordance to such a variety of requirements, preconditions and characteristics. Management of data aims to maintain the different relationships between requirements, preconditions and characteristics.

1.2 Research programme on the management of information systems

The research described in this thesis is part of the ongoing research programme on the management of information systems at Delft University. The research programme is aimed at methods and techniques for management of information systems. These methods and techniques are based on the state model for information systems and the three forms of management described below. The methods and techniques are required to be general in such a way that they are application independent and can be employed in a practical manner. The research programme is itself part of a larger programme on complex distributed information systems in which issues of planning, development and management of information systems and database systems are combined.

1.2.1 Defining management of information systems

Each set of coherent, dynamic phenomena (each system) can be abstracted into a real system and an information system that determines the behaviour of the real system [11]. The information system contains and creates representations of the real system, to control the real system. In turn both the real system and the information

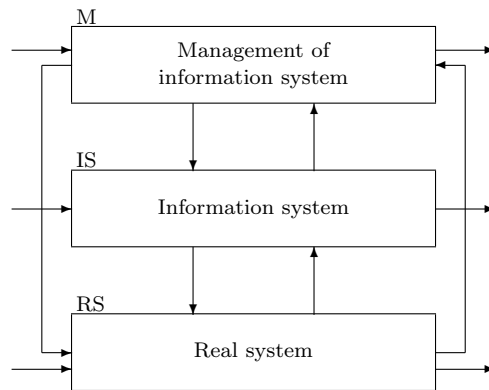


Figure 1.1: The information systems management paradigm

system can be abstracted into a real system and an information system (the recursion principle).

Definition 1 *An information system is a structured composition of hardware and associated system software and application software, data, procedures and people, for the knowledge, control and support of real systems.*

The research programme on the management of information systems is based on the information systems management paradigm where the recursion principle is used to relate management to information systems and real systems. In figure 1.1 it is shown that management maintains mutual relations with the information system and the real system. In addition to these relationships, management, the information system and the real system have links to the organization that they are a part of.

In section 1.1 it was stated that every component of an information system has its own particular characteristics. These characteristics are determined initially from requirements and preconditions set during the planning and development stages of the information system. During the utilization and exploitation of the information system (see section 1.2.2) characteristics do not remain the same but will change over time, and management activities must be performed to make sure that at any time the characteristics of information system components meet their respective requirements and precondition. Further definitions of requirements, preconditions and characteristics, can be found in chapter 4 but for now the following definition can be used to delineate the research programme on the management of information systems:

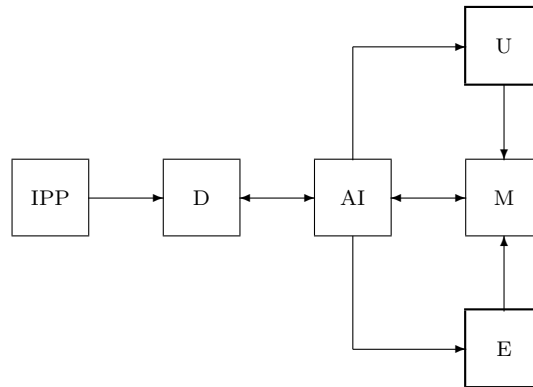


Figure 1.2: State model

Definition 2 *Management of information systems is the operation, control and maintenance of implemented information systems in accordance with user requirements and preconditions and the characteristics of hardware, software, data, procedures and people.*

Information systems management contains tasks on behalf of control, operation and maintenance of information systems. Based on their nature, structure and cohesion, the tasks of information systems management are clustered into task fields, and task fields are clustered into task areas. A complete overview of task fields and the tasks that are involved in information systems management, is given by Looijen in ‘Een Beheer Methodiek’ (A Management Method) [59].

1.2.2 State model

The task fields and the task areas of information systems management are placed in a state model that describes the states of the life cycle of an information system and the relationships between these states. Figure 1.2 shows the state model for management of information systems. Information systems management plays a part in all states of the life cycle of information systems. Information plans and information policies of the organization that lead to the development of automated information systems are defined in the state ‘information policy and planning’ (IPP). The information system is developed and built in the state ‘development’ (D) after which it is presented for acceptance. In the state ‘acceptance and implementation’ (AI) the information system is tested by the organizations that will utilize and exploit the information system. If the information system is not accepted it is rejected and sent back. If the

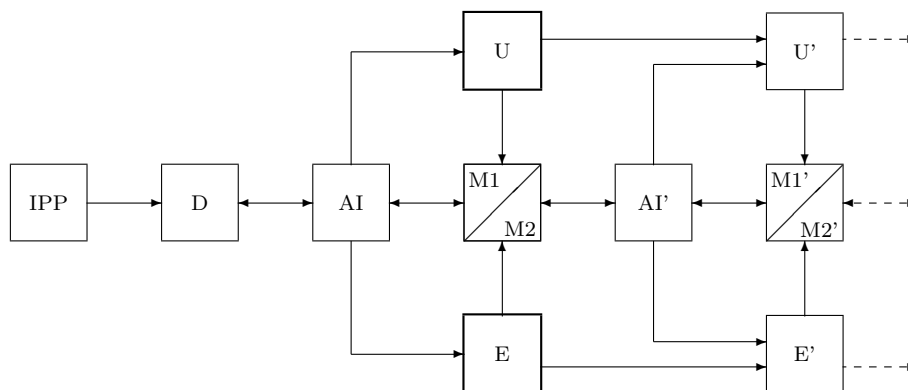


Figure 1.3: Extended state model

The state model can be extended by dividing the state ‘maintenance’ (M) into two sub-states ‘M1’ and ‘M2’. The extended state model is shown in figure 1.3. M1-maintenance are such that the management tasks in states U and E do not change. After M1-maintenance, and acceptance and implementation of the modifications, the information system returns to the original U and E states. M2-maintenance is such that the information system progresses to new utilization or exploitation states and the management tasks within the states U or E change. Progress of the information system also occurs due to the dynamics of the states utilization and exploitation but unlike progress with M2-maintenance, it does not lead to changes in management tasks. M2-maintenance can be made at the level of re-development of the information system, or at the level of re-definition of information plans and policies. The information system does not return to the original states U or E but moves to the new states U’ and E’ respectively where new management tasks are placed. The state model and the extended state model are described in further detail in ‘Een Beheer Methodiek’

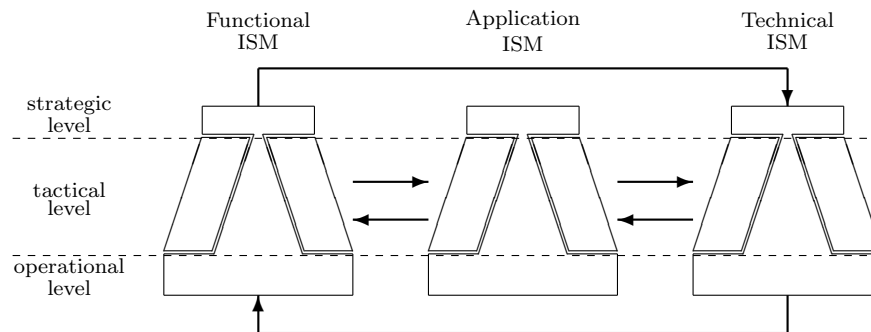


Figure 1.4: Three forms of Information Systems Management

(A Management Method) [59]. The same publication further defines relationships between tasks fields, states, and the three forms of management.

1.2.3 Three forms of management

The state model differentiates between the task fields of information systems management. The state ‘utilization’ contains task fields aimed at supporting utilization and maintaining the functionalities of an information system. The state ‘maintenance’ contains tasks fields aimed at changing the functionalities and automation facilities of an information system. The state ‘exploitation’ contains task fields aimed at supporting the exploitation of the automation facilities and technical operations of an information system. Each task field is assigned to one of three organizational units of information systems management: functional management, application management and technical management [60].

Functional management is a function directed at maintaining the functionalities of the information system. This includes support of utilization as well as evaluation of utilization and modifying functionalities of information systems. Functional management includes all task fields in the state ‘utilization’, and task fields in the state ‘maintenance’ that relate to changes in functionalities and utilization of information systems.

Application management is a function directed at maintaining algorithms and data structures of application specific software. This software is determined by specific application of the information system for control of a real system in contrast to base software that is used for general purposes. Application management includes modifying and testing application software, and contains those task fields in the state ‘maintenance’ that relate to changes in the application software.

1.3 PLACING THIS THESIS WITHIN THE RESEARCH PROGRAMME

25

Technical management is a function directed at maintaining the technical operations of an information system, and includes support of exploitation of automation facilities as well as modifying technical realization of information systems with the exception of application specific software. Technical management includes all task fields in the state ‘exploitation’, and task fields in the state ‘maintenance’ that relate to changes in automation facilities and technical operations of information systems.

In each of the three forms of information systems management, Looijen distinguishes a strategic level, a tactical level and an operational level [60]. This arrangement is independent of the position of information systems management, as a whole or in part, at the strategic, tactical or operational level of a company or government. The strategic level of information systems management is concerned with directives regarding the contents of information systems management, the locations within the organization where it is established and the relations it needs to maintain with other management units that are within or outside the organization. The tactical level of information systems management is concerned with the technical means, the personal means, and the general business support directed towards information systems management at the operational level. The operational level of information systems management is concerned with the direct support of utilization of the information system provided by functional management, the direct support of modification of application software and databases provided by application management, and the direct support of exploitation provided by technical management.

1.3 Placing this thesis within the research programme

1.3.1 Placement within the definition

Research into the management of distributed data in distributed environments is placed within the research programme on the management of information systems by adapting the information systems management paradigm for management of data (shown in figure 1.5). The research described in this thesis concentrates on the data component of information systems, it does not consider the information system as a whole. Tasks that involve the management of information systems components, people, procedures, hardware and software will not be considered in this thesis. When necessary; relationships with data management tasks will be described.

A precise description of management of distributed data in distributed environments is provided in chapter 4. For now, the following operational definition of management of distributed data in distributed environments will be used:

Definition 3 *Management of distributed data in distributed environments is the operation, control and maintenance of distributed data in accordance to the requirements and preconditions of the distributed users and the characteristics of the distributed data.*

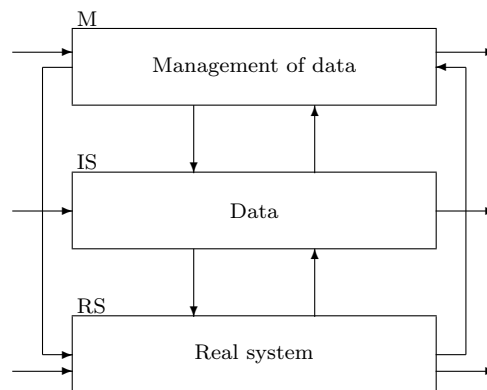


Figure 1.5: Specialization of the management paradigm for management of data

The difference between users and stakeholders from a management perspective and how stakeholders define the environment of data will be described in chapter 4, a description is also given of the ways that the environment of data and the data can be distributed.

It was stated in sections 1.1 and 1.2 that components of information systems exhibit characteristics. Research on the management of distributed data in distributed environments also involves data characteristics. The definition of management of distributed data in distributed environments can be used to describe the aim of the research described in this thesis; it is aimed at creating methods and techniques to operate, to control and to maintain distributed data in accordance with requirements and preconditions of distributed environments.

1.3.2 Placement within the state model

Various aspects of management of distributed data in distributed environments play a role in each state of the state model. The dynamic nature of data characteristics is expressed in the utilization and exploitation of data and the methods and techniques that are to be used in the states 'utilization' (U) and 'exploitation' (E). These two states are emphasized in figure 1.2. Aspects of data management in all states of the state model are discussed in 'Een Beheer Methodiek' (A Management Method) [59].

1.3.3 Placement within the three forms of management

The research provides methods and techniques to support management of distributed data in distributed environments in the states 'utilization' (U) and 'exploitation' (E). Task fields within the state U are placed in functional management and task fields within the state E are placed in technical management. The state 'maintenance' is not considered in this thesis, and the tasks fields contained within the state 'maintenance' are therefore not described. As a consequence the whole of application management remains outside this thesis, and those task fields of functional and technical management pertaining to maintenance will also remain outside this thesis. Only the functional management of utilization of data and the technical management of exploitation of data are considered in this thesis.

The methods and techniques for data management are positioned on the operational level of functional data management and technical data management, and their adequacy is determined at this operational level. Introducing the methods and techniques proposed in this thesis in a real data management organization is a decision which needs to be taken at the strategic and tactical levels of data management. It is however, not the intention of this research to involve strategic and tactical aspects of data management, but to consider such organizational considerations a given fact for the operational level.

A complete placement of this thesis within the research programme on the management of information systems can now be given:

Research into management of distributed data in distributed environments is aimed at supporting the operational level of technical management and functional management with methods and techniques for operating, controlling and maintaining distributed data in accordance with the requirements and preconditions of distributed environments within the states 'utilization' and 'exploitation'.

1.4 Research approach

The position of the research within the research programme on the management of information systems is described in section 1.3 of this thesis. A short overview of the steps of the research approach shown in figure 1.6 given below, followed by an in depth description of each step.

The first step of the research approach was to determine the aims of the research and the objectives of the research that defined what results needed to be achieved within the four years of the research period. These research objectives determined the literature and practice to be studied. The research objectives lead to an overview of the literature concerning distribution of environments, distribution of data, and management of data. The practical research was of a more general nature and provided an insight into the ways in which practice deals with matters of the management of

distributed data in distributed environments.

Conclusions drawn from the study of the literature and from practical experiences were used to formulate two research questions (see page 31). The research questions were used to specify and delimit the research topics. A framework was devised into which concepts of management of distributed data in distributed environments were placed. Methods and techniques were used to provide support and supplement data management. The framework was then combined with the methods and techniques to provide an instrumentation for the management of distributed data in distributed environments.

The research described in this thesis is characterized by three properties:

- Emphasis was placed on *how* management of data functions and *how* effective support can be provided to management of data.
- Establishing the adequacy of the instrumentation for support of data management, required application of the instrumentation in real situations.
- A lack of previous (scientific) research and of theories, concerning the complete overview of management of distributed data in distributed environments.

Wierda [108, p. 34] indicates that in this situation case-study research can be applied successfully. Yin attributes clear advantages to case-study approach when:

“A ‘how’ or ‘when’ question is being asked about a contemporary set of events, over which the investigator has little or no control.” [111, p.20]

As multiple case studies are considered to be more convincing than a single study, the testing of the practical adequacy of the instrumentation for data management involved three cases. The case study design provided for identical execution of the three case studies so that the results obtained were as unambiguous as possible. Each case was positioned by the distribution of the environment and the distribution of data, and the case study results were analysed within the context of that particular type of distribution. The case study results led to identification of the elements of the management instrumentation that were adequate or needed adaptation.

The research was finished with a comparison of the case study results and general statements providing answers to the research questions. The dotted line in figure 1.6 indicates that the research results can be used to improve the instrumentation. Further research might use the general statements as a starting point to improve the management instrumentation. Each step of the research approach is described below.

1.4.1 Research Objectives

The subject of the thesis is management of data in general and the management of distributed data in distributed environments in particular. This thesis considers distribution as an additional feature of data and the environment in which it is placed.

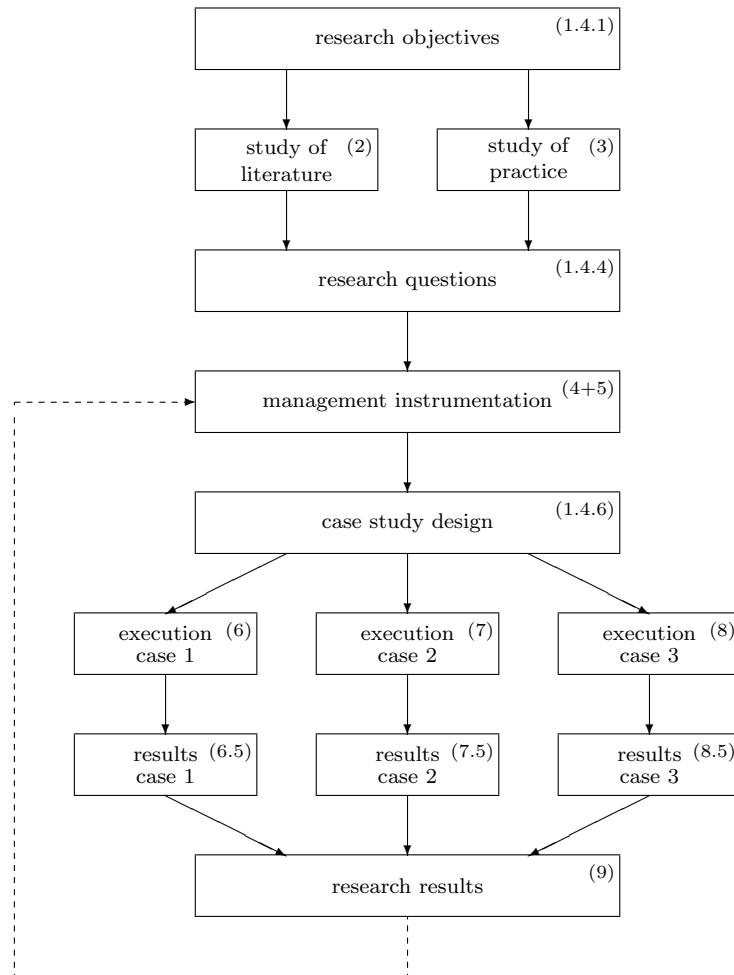


Figure 1.6: Overview of the research approach

It is shown in section 1.1 that distribution makes increasing demands on management and further study into the effects of distribution can be justified. Research into the management of distributed data in distributed environments such as is described by this thesis aims to give a systematic approach to management of distributed data in distributed environments. A theoretical foundation for data management is given

rather than a collection of descriptions of *success* stories intended to incite others to follow. This thesis is not concerned with any specific tool that can be used for data management, such as distributed database systems, but rather with the purpose for which such tools are used by data management. The intention of this thesis is to provide a framework to prevent a fragmented perception of data management and to give a complete overview of all aspects of data management in general and of management of distributed data in distributed environments in particular. The position of this research within the research programme on the management of information systems, given at the end of section 1.3.3, aims to create methods and techniques to operate, to control, and to maintain distributed data in accordance with requirements and preconditions of distributed environments. This aim of the research described in this thesis can be specified as follows:

The aim of the research was to construct an instrumentation consisting of a framework of concepts and of methods and techniques placed within this framework. The research was further aimed at providing a well-tested instrumentation to support management to operate, control and maintain distributed data in distributed environments.

Three objectives were defined to reach the aim of the research:

1. To specify a framework containing all concepts applicable to management of distributed data in distributed environments.
2. To develop methods and techniques to support management of distributed data in distributed environments.
3. To test the adequacy of the complete management instrumentation

The criteria for accepting the results of the research described in this thesis is not their theoretical soundness or theoretical completeness, rather the results are accepted on their empirical adequacy [35]. This pragmatic view is summed up by Van Fraassen as follows:

“In so far as they (pragmatic views) go beyond consistency, empirical adequacy, and empirical strength, they do not concern the relation between the theory and the world, but rather the use and usefulness of the theory; they provide reasons to prefer the theory independently of questions of truth.” [35, p.88]

The theoretical foundation of management of distributed data in distributed environments is provided by theories described in the literature and incorporating suitable theories into the management instrumentation. The first step of the research approach was therefore to study available theories.

1.4 RESEARCH APPROACH

31

1.4.2 Literature

The research objectives and the research programme on the management of information systems determined what approaches and theories were selected from the literature available on the subject of this thesis. The approaches and theories must concern the three subjects of research reported on in this thesis: environments, data and management. The description of management of distributed data in distributed environments offered by literature may lead to the conclusion that certain aspects haven't been studied sufficiently. The literature must show that research into management of distributed data in distributed environments is meaningful and necessary. Theoretical descriptions from the literature are not sufficient, however, ones own observations are necessary to gain insight into the research matter.

1.4.3 Practice

It was necessary to gain further insight into current practices in management of distributed data in distributed environments before research into the field could progress. Two surveys into current practice were performed: one at the Dutch 'Joint Office for Social Security Administration'¹ (GAK) and one at the 'central organization' (CO) of the Dutch Ministry of Defence. These surveys provided further insights into the difficulties faced in current practice of management of distributed data in distributed environments. The surveys of current practice were performed concurrently with the theoretical research. As a consequence the research questions and the management instrumentation were based upon observations of important theories and recent practices in management of distributed data in distributed environments.

1.4.4 Research questions

The topics of interest of this thesis are expressed by two research questions that are based on the studies of literature and practice.

The role of characterizations of data in data management were described in section 1.3. Literature provides many characteristics of data that are eligible for use in management. It is possible that some of the characteristics that are part of the conceptual framework, will be found to be of no use within a particular situation. It is also possible that within a particular situation additional characteristics play an important part. This thesis shows to what extent the characteristics within the conceptual framework are adequate for managing distributed data in distributed environments. The adequacy of characteristics can be determined by answering the following research question:

Research question 1 *What characteristics of data are of interest for performing tasks of managing distributed data in distributed environments?*

¹In Dutch: Gemeenschappelijk Administratie Kantoor (GAK)

The answers to the first research question determine what are the relevant characteristics of data for performing management. Consequently it is possible to determine what methods and techniques can be used to let these data characteristics meet the requirements and preconditions made from the utilization of data. Taking into account the aim of this research project the methods and techniques used should at least be able to model the following items:

- The environment of data
- The data
- The distribution of the environment and of data
- Activities within management tasks
- Organization of management of data

The above considerations lead to the following question:

Research question 2 *What methods and techniques can be made available to management to operate, to control, and to maintain distributed data in accordance to the requirements and preconditions of a distributed environment and the characteristics of the distributed data?*

The answers to this questions will indicate the possibilities and limitations of the instrumentation for managing distributed data in distributed environments.

1.4.5 Management instrumentation

Both literature and practice provide elements for the foundations of the instrumentation for management of distributed data in distributed environments. Approaches described in literature form the basis for the instrumentation, and practice points out particular aspects of management of distributed data in distributed environments that are part of the instrumentation. The instrumentation for management of distributed data in distributed environments can be divided into two parts. The first part is a general framework of concepts that are important for management of distributed data in distributed environments. The second part consists of methods and techniques, placed within the framework, that can be applied in management of distributed data in distributed environments.

The management instrumentation was tested by performing three case studies in different practical situations, and it can be corrected based on judgements of its adequacy and practical use in these three case studies. An identical way of performing the case studies was required to compare the results of the three different case studies and to guarantee that all the case studies contribute collectively to the answers to the research questions.

1.4 RESEARCH APPROACH

33

1.4.6 Case study design

The area of application for the management instrumentation should be as large as possible. It is important to perform the case studies in a wide variety of situations. Testing the application of the management instrumentation in a variety of practical situations gives a large as possible application domain to which it can be applied. Such a large application domain enables general statements concerning the management of distributed data in distributed environments. Despite the diversity of the practical situations found, it must be possible to relate the results of individual case studies. This can be done by performing each case study in an identical manner.

Identical execution of the case studies is guaranteed by using the same modelling techniques and the same phases. The same methods and techniques are used for modelling management of distributed data in distributed environments. The phases in which the case studies are performed are based upon the FATO framework that is developed as part the management instrumentation. It relates four basic aspects of data management: functionalities (F), automation facilities (A), tasks of management (T) and organization of management (O), and it is described in detail in chapter 4.

Phase-F descriptions of the environment of data and of functionalities and requirements for data:

1. descriptions of stakeholders and functionalities
2. requirements and preconditions at the pragmatic and semantic data levels

Phase-A descriptions of technical implementation of data:

1. descriptions of automation facilities
2. requirements and preconditions at the syntactic and empiric data levels

Phase-T descriptions of the management of data:

1. activities within the task field actual data management
2. activities within the task field data administration
3. activities within the task field database administration
4. activities within the task field storage management

Phase-O descriptions of support and organization of management of data

1. supporting aids
2. organization

The results of the various practical situations can be judged when each case study is performed in this manner. The criteria by which the case results are judged is given by the following question:

Is it possible to find bottle necks in data management in specific situations, and does the instrumentation developed for management of distributed data in distributed environments provide suggestions for improvement?

Every case study must determine if it is possible to use the instrumentation to analyse data management in a practical situation and to show improvements when necessary. The case studies must be positioned with respect to each other, so it is possible to relate the results.

1.4.7 Case execution

Each case study is an experiment that tests the management instrumentation in a different situation. Selection of a particular situation is based upon two aspects:

1. Distribution of the environment
2. Distribution of data

Different situations of management of distributed data in distributed environments can be positioned with respect to each other using these aspects.

Distribution of the environment

Distribution of the environment is described using decentralization of the organization where data management is performed. This implies that deconcentration of the environment is not considered when positioning the case studies. The environment of data consists of those who have a vested interest in the correct realization of the required data functionalities. Those with a vested interest are referred to in this thesis as *stakeholders* and stakeholders can be people or organizations or other information systems that require data to perform certain functionalities on their behalf. The *custodian* decides what functionalities the data perform, and what automation facilities are employed. The authority to make this decision is granted by the owner.

This thesis describes decentralization of the environment by placing the custodian within an organizational diagram first used by Mintzberg [73]. Mintzberg defines five types of decentralization where, in each case, one part of the organization takes control of the decision making process. For the sake of simplicity stakeholders are placed within the operating core of the organization though stakeholders may reside outside the organization. The five organizational configurations for decentralization defined by Mintzberg are shown in the following table:

Organization configuration	Location of custodian
type A: Simple structure	Strategic apex
type B: Machine bureaucracy	Technostructure
type C: Divisionalized form	Middle line
type D: Adhocracy	Support staff
type E: Professional bureaucracy	Operating Core

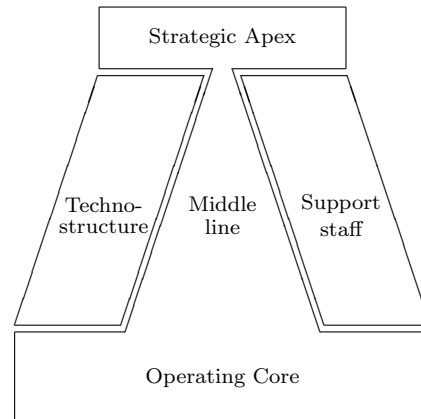


Figure 1.7: Diagram for distribution of the environment (decentralization)

The diagram used in this thesis to position distribution of the environment within each case study is shown in figure 1.7.

Distribution of data

Distribution of data is described using the geographical deconcentration (dispersion) of functionalities, automation facilities and the management organization. This implies that decentralization of data is not considered when positioning the case studies. This description of data distribution is based upon the FATO framework for data management defined in chapter 4, and should not be confused with the three forms of management defined in section 1.2.3.

Functional distribution of data is defined in this thesis as the distribution over several geographical locations of utilization of the actual contents and definitions of data for the knowledge, control and support of real systems.

Technical distribution of data is defined in this thesis as the distribution over several geographical locations of the exploitation of data structures and data storage that are implemented using available automation facilities.

Organizational distribution of data is defined in this thesis as the distribution over several geographical locations of the people and means employed in management of data, and also the distribution of responsibilities, authorizations, work and expertise required to perform management tasks over several geographical locations.

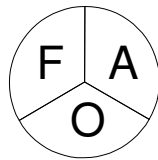


Figure 1.8: Diagram for distribution of data (deconcentration)

Aspects of distribution of the data are further elaborated in section 4.5. The diagram that is used in this thesis to position the distribution of data within each case study is shown in figure 1.8.

Three case studies are described in this thesis: two case studies from the “Joint Office for Social Security Administration” (GAK) and one case study undertaken at Shell Netherlands Chemicals. All three case studies were positioned using the diagrams for distribution of the environment and data. Positioning the case studies enables analysis of the management instrumentation for different kinds of distribution. Every case has specific characteristics that do not concern the distribution of the environment or the distribution of data, but that must be held in account when analyzing the results of the case study.

1.4.8 Case results

The results of each case study were judged by asking to what extent the management instrumentation is adequate for analyzing management of distributed data in distributed environments and providing suggestions for improvement (see page 34). Each case study puts the management instrumentation to the test by applying the methods and techniques offered by the instrumentation. The results of the application of the instrumentation were validated by interviews and by using other sources from the case studies. From the validation it became clear what elements of the management instrumentation are to what extent suitable for use within a particular situation. To obtain results for the research as a whole it was necessary to relate the results of each of the three case studies.

1.4.9 Research results

The relationships between the three case studies were shown by the contrasts and similarities between the results of the three cases. The contrasts and similarities showed the extent to which the management instrumentation presented in this thesis is of practical use for management of distributed data in distributed environments.

1.5 OUTLINE OF THIS THESIS

37

Relating the results of the three case studies also provided answers to the research questions given in section 1.4.4. Finally there are some recommendations for further research to extend and improve the usability of the management instrumentation.

1.5 Outline of this thesis

This first chapter gave an introduction to research into management of distributed data in distributed environments. The contents of this thesis are positioned inside the research programme “management of information systems” and the research approach is described and motivated. The approaches concerning the environment of data and its distribution, data and their distribution and in the final section, management are described in chapter 2. The two surveys performed at the Ministry of Defence and at GAK are described in chapter 3. The results of these surveys are structured, in an identical manner to chapter 2, in three sections concerning the environment of data and its distribution, data and their distribution and in the final section, management. The instrumentation for management of data developed during the research is described in chapters 4 and 5. The FATO framework that relates all the aspects involved in managing distributed data in a distributed environment is described in chapter 4 and data distribution is positioned within the framework. The methods and techniques that can be used for managing distributed data in a distributed environment are described chapter 5, these methods and techniques are also positioned within the framework. Descriptions of the results of the three case studies are given in chapters 6, 7 and 8, and each of these three chapters ends with the conclusions concerning the applicability of the management instrumentation for that particular case study. In the final chapter of this thesis the results of the research are described and answers are given to the research questions of section 1.4.4. Recommendations for further research are given at the end of this thesis.

Chapter 2

Environment, data and management: literature

In this chapter a complete as possible overview is provided of all the various views in the field of management of distributed data in distributed environments. In this manner it provides insight into the research subject and outlines the underlying rationale for the research presented in this thesis.

The definition of management of information systems and the management paradigm (see paragraph 1.2) form the basis of the research programme on management of information systems. The research on management of distributed data on distributed environments is placed within these two definitions by specialization towards data as one of the five information system components. Requirements and preconditions from utilization by distributed users and data characteristics are the key elements of the definition of management of distributed data in distributed environments given on page 25. Data utilization is placed within the data environment where data perform functions on behalf of the users. As a consequence, the definition of data management presupposes a triangular relationship between the data environment, data, and data management.

This chapter is structured around the three elements environment, data and management, that are of interest in this thesis. Section 2.1 describes the data environment that produces requirements and preconditions as a result its utilization of data. Section 2.2 describes the characteristics that are exhibited by data, these characteristics have to meet the requirements and preconditions set by the environment. Due to the dynamics of the environment and of the data, the requirements, preconditions and characteristics are not automatically in agreement with one another. Section 2.3 describes the management required to operate, control and maintain data in such a way that its characteristics are in accordance with the requirements and preconditions set by the environment. Some conclusions regarding the views presented in this chapter are given in the final section 2.4.

2.1 Environment

The data environment is defined within the context of this thesis as all those parties who have an interest in the data, as a supplier or as a consumer of the data. These parties are the stakeholders who have a “stake” in the successful operation of the business functions that require the data [65]. In this thesis the stakeholders are placed at the level of the real system (see figure 1.1) where the business processes take place.

The information paradigm [11], from which the information systems management paradigm (see figure 1.1) is derived, defines that: “each set of interesting phenomena (i.e. also each dynamic system such as an organization) can be abstracted into a real system and an information system which determines the behaviour of the real system”. The representations of the real system created by the information system to obtain knowledge of, to control and to support the real system are laid down in the data contained in the information system. The following sections deal with the nature of this relationship between data and the real system, regarded in this thesis as the data environment.

The relationship between data and the real system is described in section 2.1.1, the effects of administrative integration in the data environment are given in section 2.1.2, ownership and custodian-ship of data are positioned within the environment in section 2.1.3, and the distribution of the environment is discussed section 2.1.4.

2.1.1 Relationship between data and the real system

The relationships between data and the real system are described using matrixes that show the business processes that create and use data. In Business System Planning (BSP) such ‘create/use’ matrixes are used to describe the relationships between data and business processes [50]. An identical approach is taken by James Martin [66]. In these approaches, data can be used or created within multiple business processes, and, in general, data cannot be attributed to a single process. Two categories of business processes are important in this thesis: administrative processes and industrial processes [32]. The differences in the real system may also lead different requirements and preconditions being posed upon data.

A second kind of distinction in the real system is the difference between the two approaches to the role of data: process oriented and data oriented. These two approaches may lead to different requirements and preconditions for data and are the subject of the second section.

Administrative and industrial processes

A commonly used subdivision of organizations is the division for organizations into those where the flow of goods is the primary component of the value cycle (see figure 2.1), and those without a predominant flow of goods [32, 53, 97]. A flow of goods can be found in production oriented organizations such as the chemical industry,

2.1 ENVIRONMENT

41

manufacturing, and trade [1]. Examples of organizations without a flow of goods are banks, insurance firms and public administration.

Industrial and logistical processes found in industry and trade require three types of data [101]: data required to make the product, data required to manage the production process, and data required to judge the condition of the production facilities. Within industry and trade, internal audits are less important due to the strong relationships between the flow of goods and the flow of money [1]. These strong relationships also mean that changes in industrial and logistical processes lead to changes in data definitions and data contents.

Administrative processes can be found in all types of organizations, they support the industrial and logistical processes, for example registration of inventory, but they are also an independent business function, for example payroll administration and accounts payable/receivable [7]. The administration function is concerned with data collection, data processing, data storage and data provision to the productive and controlling business functions that need particular data [3], and as such administrative processes are processes that have data processing as primary component. The indirect nature of administrative processes in the way they acquire data, requires different standards for correctness, completeness and reliability [52]. In financial institutions and government institutions internal audit are of high importance [1].

Due to these differences in relationships between data and the real system, it is likely that environments where administrative processes are important have a higher stake in the requirements and preconditions they impose on data, than environments where industrial processes are important, and consequently, more stringent measures are installed in data management.

Process oriented and data oriented approaches

Another way to look at the relationship between data and the real system, is the contrast between the process-centered and data-centered approach. These two approaches can be found in the functional view (process-oriented), the object view (data-oriented) and the communication view (that combines process and data) of information management [105]. Separation of processes and data is also applied in different types of systems modelling [85]. The process and data approach can easily be combined to produce an infrastructural approach to data, see section 2.2.2.

The process oriented approach to data defines, for individual business functions, a separate information system and data [61]. In this situation data are derived from the process or function, without considering the fact that this data may be used somewhere else in the organization. The data are defined and structured to provide for the needs of a specific department or organizational entity, or to support a particular manager. Such data can quickly become obsolete when the corporation is reorganized or management changes [66]. The requirements and preconditions for data are optimized for the specific process and may be less suited for other process that could benefit from the data. The process oriented approach is therefore best

suited when it is clear from the beginning that the data will not be used in other processes. The databases developed from a process oriented approach are known as ‘application databases’ [67].

The data oriented approach starts from a different perspective. It takes not one but all business processes into consideration and tries to find what business objects they have in common [61]. The data oriented approach emphasizes the creation of automated data bases for common use, before building individual applications. The data oriented approach avoids problems with proliferation of files, and maintenance, redundancy and inconsistency of data associated with the process oriented approach [67]. The data oriented approach implies sharing of data, this could become a problem because it can mean that users lose control of data they feel they ‘own’. Shifting control of data away from process managers can lead to a shift of power within an organization [61]. The requirements and preconditions for data are balanced for use by all current, and possibly future, processes and are therefore not optimized for a particular process. The databases developed from a data oriented approach are known as ‘subject databases’ [67] or ‘data classes’ [50]. Subject databases are required when introducing administrative integration.

2.1.2 Administrative integration

Administrative integration is cooperation between and linking of administration activities in business functions based upon mutual interest. It implies a common use of data for different business functions, and this may lead to the introduction of subject databases and data infrastructures, see section 2.2.2. When the administrative processes of business functions within the same organization are integrated, it is called ‘internal integration’, integration of administrative processes of business functions between different organizations is called ‘external integration’.

Internal integration and the value cycle

Internal administrative integration is cooperation between administrative processes within the same organization [8]. It can be achieved by exchange of data, or common application of data between different business functions. The relationships between business functions of an organization are shown in figure 2.1 in what is known as the ‘value-cycle’ [53]. Cooperation between the business functions purchase, production and sales is possible through the exchange of data. This enables a higher quality of both the product and the service. For example: the customer can be given a more accurate estimate of date of delivery, and if there is a production delay, this can be reported to the customer. Advantages in production and marketing can be gained by closer cooperation among internal business functions.

2.1 ENVIRONMENT

43

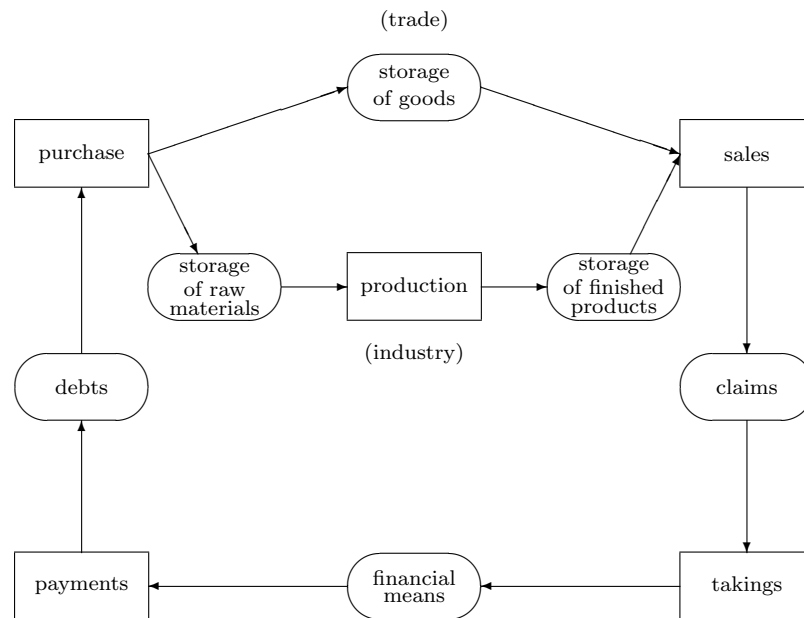


Figure 2.1: The value cycle

External integration and the value chain

External administrative integration is cooperation between administrative processes between organizations [8]. It can be achieved by exchange of data, or common application of data between business functions in different organizations. External administrative integration is discussed by Wierda in terms of ‘interorganizational information systems’ (IIS) [108]. The role of Electronic Data Interchange (EDI) in interorganizational information systems is discussed by Streng [98]. This linking of the business functions of different organizations is also known as ‘chain integration’ as it links the value cycles of individual businesses into a single value chain, see for example [98] and figure 2.2. As well as pure chain integration there is a second kind of external integration in which providers of basically the same products or services participate in an exchange of data [8].

Chain integration establishes consecutive links in the flow from raw materials to finished products for the consumer, standardization of (for example) order forms facilitates cooperation between many different participants in a particular value chain [8]. A uniform article numbering scheme (‘bar-codes’), i.e. the European Article Numbering (EAN) scheme, for consumer goods sold in supermarkets and department stores

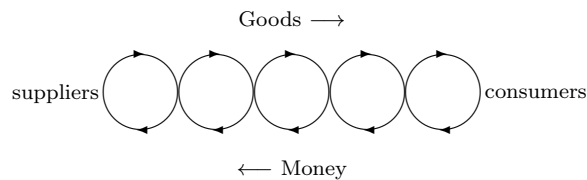


Figure 2.2: Linking value cycles into a value chain

is one example of standardization in the interest of integration in a value chain that combines manufactures, wholesalers and retailers.

The second kind of external administrative integration is concerned with exchange of data between participants who in principle, perform the same functions and thus by nature are closely related, such as banks where money transfers and cheques need to be processed [8]. In this kind of external administrative integration standardization is also an important issue, for example all cheques should have the same format. Airline tickets issued by airlines and standardized by the International Air Transport Association (IATA), are another example of the second kind of data exchange among organizations that perform the same functions.

2.1.3 Owner and custodian of data

Traditional perceptions concern the “ownership” of data and data management [103] but with increasing distribution of the data environment, caused by administrative integration, matters of ownership are less obvious. Strictly speaking, the owner of the data is the person or corporate body who has a legal right over it. Within a corporation, usually the board of directors can be called the owners of the data because they decide what means, both capital and labour, can be employed on behalf of data. Due to the complexity of the organizational structure of the corporation the board of directors are seldom actually able to control the data. This control is then delegated to a particular unit within the organization which receives a mandate from the owner to become the custodian of the data. As such the custodian becomes the effective owner of the data, however within the authority of the board of directors [10]

This organizational unit, or person, has custody [51] or stewardship [65] over data, and a data oriented approach requires a separation of ownership and custodian of data [67]. Although in the strict sense the owner and custodian do not belong to the real system and thus the data environment, they do belong to the organizational structure in which the data are placed. When a value chain is set up, using external administrative integration, attention has to be paid to where the custodian of the data is positioned. Ownership of the data can be dispersed within the chain, but the

2.1 ENVIRONMENT

45

custodian should be defined clearly. This could be a problem, because the introduction of data used in common can shift the existing balance of power within organizations and between organizations [61].

2.1.4 Distribution of the environment

The data environment consists of those interested parties that have a stake in the requirements and preconditions of data that are to be met. Distribution of the data environment is influenced by a number of factors, a few of which are mentioned below.

An environment with a process oriented approach will have data centered around a single application, and its utilization restricted to single locations and within a single organization unit. Environments with a data oriented approach tend to utilize data for more than one application, leading to utilization of data at several geographical locations within organizational units, and thus a more distributed environment. Environments that are characterized by extensive administrative integration, both internal and external, see a substantial increase in the geographical and organizational distribution of data utilization. When organizational borders are crossed, as happens with external administrative integration, matters of ownership and custodianship need to be defined clearly.

As illustrated by the above discussion, distribution of the data environment manifests itself by the way the stakeholders are placed. The stakeholders can be distributed in two ways: at various geographical locations (deconcentration); and at various organizational positions (decentralization).

Deconcentration

Deconcentration is also called geographical decentralization [53] and concerns the physical dispersion of facilities over several geographical locations [73]. The decision to concentrate or deconcentrate an organizational unit reflects the tradeoff between work-flow dependencies (namely, the interaction with the users of the services provided by the unit) and the need for specialization and economies of scale [73].

For example; the Joint Office for Social Security Administration¹ (GAK), has district offices placed at thirty locations throughout The Netherlands close to the users of the administrative services provided by GAK regarding the execution of Dutch social security legislation. Specialized functions for administration auditing and financial transactions etc. are concentrated at the head office of GAK in Amsterdam. For more on GAK see chapter 3.

Decentralization

Decentralization is also called functional decentralization [53] and concerns the dispersal of decisional powers within an organization [97]. In a centralized organization all

¹In Dutch: Gemeenschappelijk Administratie Kantoor (GAK)

the power for making decisions rests at a single point in the organization, ultimately in the hands of one person, in a decentralized organization the power is dispersed among many people [73]. Decentralization can further be distinguished into vertical decentralization and horizontal decentralization, and into selective decentralization and parallel decentralization [73].

Vertical decentralization refers to the amount of decisional power dispersed down the line of authority of managers, horizontal decentralization refers to the extent non-managers, support specialist in the support staff and the analysts in the technostructure of the organization, control decisional processes. Selective decentralization refers to dispersal of power over different kinds of decisions that rest in different places in the organization. Parallel decentralization refers to the dispersal of powers over different kinds of decisions to the same place. From these (two-by-two) forms of decentralization five structural configurations for an organization emerge, with increasing levels of decentralization: the simple structure, the machine bureaucracy, the divisionalized form, the adhocracy and the professional bureaucracy [73].

The simple structure is characterized by horizontal and vertical centralization. In this type of organization decisional power is concentrated at the top of the organization, strategic apex, typically in the hands of a single person. This chief executive retains both formal and informal power, making all important decisions and coordinating their execution using direct supervision. The machine bureaucracy is characterized by selective horizontal decentralization. This type of organization relies on coordinating unskilled tasks using standardization of work processes. Formal power remains at the top of the organization, but because of their importance to standardizing work processes, the analysts in the technostructure are able to gain some informal power. The divisionalized form is characterized by parallel vertical decentralization. This type of organization is usually divided into autonomous units, or divisions, and a great deal of formal power concerning their own unit is delegated (in parallel) to the managers. An adhocracy is characterized by selective horizontal and vertical decentralization. Power for different types of decisions is delegated to work constellations of specialized workers that make selective use of staff experts. Coordination within as well as between the work constellations is effected primarily through mutual adjustment. The professional bureaucracy is characterized by horizontal and vertical decentralization. Decisions are taken by highly skilled professionals in the operating core, they work relatively autonomously without direct influence from managers or analysts. What coordination there is, is achieved through standardization of skills (e.g. previous education). This taxonomy of decentralization as defined by Mintzberg [73] is summed up in table 2.1, see also figure 1.7 for placement of the five basic parts of an organization.

Organizational structure	Prime coordinating mechanism	Key part of organization	Type of decentralization
Simple structure	Direct supervision	Strategic apex	Vertical and horizontal centralization
Machine bureaucracy	Standardization of work processes	Technostructure	Selective horizontal decentralization
Divisionalized form	Standardization of outputs	Middle line	Parallel vertical decentralization
Adhocracy	Mutual adjustment	Support staff	Selective vertical and horizontal decentralization
Professional bureaucracy	Standardization of skills	Operating core	Vertical and horizontal decentralization

Table 2.1: Five types of decentralization

2.2 Data

As mentioned in section 2.1, data are one of the five components of an information system. A more detailed discussion of what data are, and what their role in information systems is, can be found in section 2.2.1. The process oriented approach and data oriented approach to defining the relationships between data and the real system, lead to process databases and subject databases. Administrative integration requires process and subject data to be combined leading to an infrastructural approach to data. The nature of data infrastructures is discussed in section 2.2.2. The basis upon which data management is performed is the operation, control and maintenance of data in accordance with user requirements and preconditions (see page 22), an overview of requirements of data that can be found in current literature is given in section 2.2.3. Distribution of data is discussed in section 2.2.4, in this section the reasons why data might (or should) be distributed, the architectures that make data distribution possible are considered. The section ends with specific (management) aspects that have to be taken into consideration when implementing data distribution.

2.2.1 Data viewed as an information system component

Each set of coherent, dynamic phenomena (each system) can be abstracted into a real system and an information system that determines the behaviour of the real system. To this end the information system contains and creates representations of the real system, to control the real system. An information system is a structured composition of hardware, software, data, procedures and people (see page 21) and as such data are a component of the information system [11]. The function of data is derived from the function of the information system, i.e. to make representations of entities, states

and processes in the real system. The functional relationships between data and the real system are discussed in detail in section 2.1.

From an information systems management viewpoint, information system components include documentation to support the installations, utilization, and maintenance of hardware, software and data. They also include the technical facilities that provide electricity, cooling, fire prevention, safety and security. Hardware and software are ordered into several classes and subclasses of automation facilities. Examples of subclasses in relation to data are database management systems, data dictionary systems, cache memories, magnetic disks, and tape cartridges [60] (see also figure 4.1).

The functional relationship between data and the real system can be considered at the pragmatic level of the effect and utilization of data in the real system, and at a semantic level where the meaning of data is defined in the relationship between data and the real system. The technical relationship between data and automation facilities can be considered at the syntactic level of data structures, and at the empiric level of data representations of physical media [96].

2.2.2 Data infrastructures

The common use of information systems and data via the introduction of administrative integration will lead to functional relationships for data involving more than one real system. This requires an infrastructural approach to data where technical relationships with automation facilities enable shared use of data and integration in multiple information systems. Information systems also incorporate specific data that is not shared with other information systems and that can be considered to be built on top of the infrastructural data, these two types of databases are referred to as process databases and subject databases.

Process databases

Process databases are related to the process oriented approach of implementing the information provision function of an organizational unit. Process databases are specific to a single information system which controls a single business process. A process database contains all the data needed for controlling this single business process. This means that a description of a product can be stored in several databases, for example in a sales database, a storage database and a production database [61]. Each database has its own definition of the product, with its own attributes for the product. To determine a particular attribute of the product, you will have to know which database it is stored in.

Process databases are also called application databases because systems analysts tend to create a separate database for each new application program [66]. Database management systems are used to obtain some degree of data independence, i.e. data and software can be maintained independently to a certain degree. As more systems are implemented the number of databases grows, creating a high degree of duplication

2.2 DATA

49

of data. Maintenance of this high number of databases takes up a large portion of the work of data management personnel, causing high costs for database maintenance. Duplicated storage of data is also costly. The popularity of process databases can be attributed to the fact that they are easier and faster to implement than subject databases [66].

Subject databases

Subject databases are related to the data oriented approach to implementing the information provision function of an organizational unit. Instead of storing data concerning a business entity in different databases, all facts concerning the business entity are stored in a single subject databases. Examples of entities to be stored in a subject databases are: products, product specifications, customers, orders, suppliers, and parts. Subject databases are also called data classes [50, 65].

Initially subject databases are slower and more expensive to implement due to the detailed data analysis and modelling required, once implemented operating and maintenance cost are lower. The data to be used are already available, this means savings in data definitions, data structures and testing the database when new information systems are developed. All that is needed, is testing of the subject databases together with the new application software of the information system. Thus overall, subject databases provide savings in development and maintenance of information systems. The number of subject databases does not grow linearly with the number of information systems. The data needed is already available in one of the subject databases, and at a certain point all the data is already available somewhere. As the number of databases is reduced, integrity checking and other data management tasks can be done using a single management process. When process databases are used, multiple integrity checks have to be performed by several management processes [66].

Subject databases interconnect the information systems that are used to control the processes that manipulate the entity. As such the subject databases constitute the data infrastructure upon which the information systems are built. The data infrastructure itself is built on top of the technical infrastructure which contains all automation facilities employed in common use within the organization [60].

Relationship between process and subject databases

The above sections describe the distinction between process databases and subject databases. Process databases contain data which are specific to controlling a single business process. Subject databases contain data which are used in controlling a multitude of business processes. Some authors consider process database to be outdated and propose keeping all data in subject databases [66, 82], although they can be implemented only under strict conditions [61].

Combining process databases and subject databases within an infrastructural approach enables integration of information systems. This means that the information system has access to both its own process databases and the subject databases. A

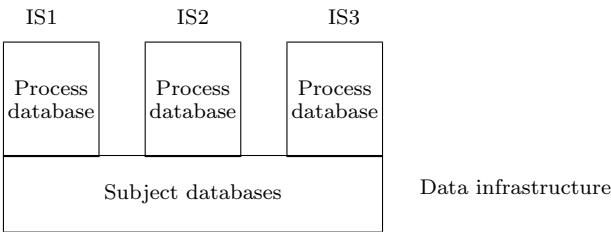


Figure 2.3: Relationship between process databases and subject databases

process database that is part of a particular information system, contains references (e.g. foreign keys) to entities in the subject database. Such dependencies between databases require careful planning and management, and can only be successful when a substantial amount of personnel and financial means are committed over a considerable period of time. Without such an commitment the databases will be incomplete and the information systems fragmented, which makes subsequent integration difficult [31].

2.2.3 Requirements on data

The current literature on requirements on data can be divided into two groups: the first group views data from a business management perspective where optimal utilization of data to improve business performance is the central issue. The second group views data from a technical perspective where minimizing costs of data exploitation is the central issue. The requirements and preconditions are formulated from the management perspective and the technical perspective and are used as a guideline for the functionaries involved with data management. The two perspectives are also found in the partition into functional and technical data management (see page 78), where functional data management operates, controls, and maintains data in accordance with requirements made from a business management viewpoint, and where technical data management operates, controls, and maintains data in accordance with requirements made from a technical viewpoint. The requirements specified in the literature are not defined broadly, but have a rather narrow focus and are specialized, without an overview of data.

Business management perspective

Authors that write from a business management perspective have an EDP auditing background, a chartered accountant or an information systems development background. Requirements and preconditions on data are consequently specified for the

2.2 DATA

51

purpose of maintaining proper accounts using financial data or for the purpose of satisfying the customer of the information system development project.

Veldhuizen [102] states that the influence of accountancy on data management is justifiably large as the accountant's certification of the annual account of a corporation also implies a judgement of the information systems which provide the data to draw up the annual accounts. Audit and internal control is however, restricted to financial and economic data and a wider context for data management is required. Requirements and preconditions for data are defined from an auditing perspective of data management with criteria for the contents of data and for the form of data. The criteria for the contents of data are: completeness, correctness, timeliness, accuracy, trustworthiness, and auditability. The criteria for the form of data are: quantification, surveyability, readability, and consistency in the datamodel [102].

The control practices for databases are grouped by Fernandez et al. [33] into three categories: (1) transaction initiation and data entry, (2) database content, data processing and data access and (3) database storage. Additional database considerations used by auditors for reviewing controls are: adequacy of database administration, control functions performed by DBA, separation of data control from application development, DBMS used, the number of databases and their use, adequacy of database documentation and access control for database documentation [33].

Starreveld et al. [97] consider data criteria from the standpoint of justification of actions and internal control within organizations. They group the criteria into data content criteria and data form criteria. Data content criteria consist of relevance of data, the way in which user needs are satisfied by the data and the trustworthiness of data. The data form criteria are the presentation form of the data and efficiency of the data. Data relevance is further subdivided into ten aspects: abundance, completeness, level of detail, critical appraisal, quantification, accuracy, delay in data provision, speed of data provision, period in which data is available and finally continuity. Data trustworthiness is subdivided into the three aspects: correctness, completeness, and timeliness [97].

Jans [53] provides a number of quality criteria for data from the perspective of business administration and the administrative procedures that need to be developed. The three criteria mentioned by Jans are: effectiveness, efficiency and trustworthiness. Data effectiveness is subdivided into: continuity of the provision of data and flexibility of data processing systems. Data efficiency is subdivided into: cost, duration of action, and productivity of personnel. Data trustworthiness is subdivided into: division of labour, auditable division of functions, authorization, timeliness, correctness, and completeness of data [53].

Mollema [75] defines data quality criteria from an EDP-auditing viewpoint. The list of data criteria given by Mollema includes: accessibility, availability, semantic correctness, syntactic correctness, consistency, validity, accuracy, timeliness, completeness, sensitivity, auditability, confidentiality [75].

This EDP-auditing viewpoint is also taken by Van Biene-Hershey [5]. The list of auditing aspects for data given by Van Biene-Hershey includes: physical security, au-

thorisation and controlled access, integrity (which includes completeness, correctness and timeliness), efficiency, controllability, auditability and continuity [5].

Jackson [51] determines the productivity of data centers by measuring the response time of on-line data access with regard to: transaction volumes, average response times and percentage of transactions completed within a certain period, and by measuring batch turnaround times and the availability and reliability of computer hardware [51].

Brevoord [7] groups criteria for the functioning of management information systems into user requirements, technical requirements and management requirements. User requirements consist of: user acceptance, user aspects (responsiveness, frequency timeliness, flexibility and adaptivity) and user friendliness. Technical requirements consist of: structuring of data files, algorithms and models, reliability of data, accuracy of data, integrity, security of data, uniformity of design and documentation and uniformity of semantics. Management requirements consist of: effectiveness and efficiency of design, realization and operational functioning of the information system, maintainability of the information system, expandability of the information system, compatibility of hardware and software, and data privacy protection [7].

Van Hulzen and De Moel [74] define a number of performance characteristics for information systems that are used to support control mechanisms employed during the development of information systems. The main characteristic of an information systems is its trustworthiness, this is divided into five elements: availability, continuity, integrity, accessibility and timeliness. Each of these five elements is subdivided into a number of characteristics for the information system as a whole. In the list below the characteristics are attributed to the data component of the information system. Characteristics that contribute to the availability of data are: media robustness, implementation correctness, resilience, and recoverability. Characteristics that contribute to data continuity are: adaptability, imperturbability, protectability, and survivability. Characteristics that contribute to data integrity are: actuality, correctness, consistency, and auditability. Characteristics that contribute to data accessibility are: user friendliness, operability, communicativity, and screenability. Characteristics that contribute to data timeliness are: speed, frequency, response time, and readiness [48].

Delen and Rijsenbrij [27] take a step further in quality of information systems and information systems development, and define four quality dimensions, divided into twenty-one quality aspects and subdivided into forty-one quality attributes. This same subdivision of quality is used by Delen and Looijen [26] for the functional management of the information provision function². The four quality dimensions are: the process dimension of information system development, the static dimension of information system properties, the dynamic dimension of information system functioning and the information dimension. The (development) process dimension is not of interest in this thesis. The static dimension is divided into the quality aspects: flexibil-

²In Dutch: informatievoorziening

2.2 DATA

53

ity, maintainability, testability, portability, connectivity, reusability, and suitability for the technical infrastructure. The dynamic dimension is divided into the quality aspects: trustworthiness, continuity, efficiency, and effectivity. The information dimension is divided into the quality aspects: correctness, completeness, actuality, accuracy, auditability [26, 27].

Technological perspective

Authors that write from a (information) technology perspective have a background in computer engineering and software development. Requirements of data are consequently specified to make optimum use of automation facilities and to support efficient execution of application algorithms when implementing databases.

Date [23] defines requirements for database systems to enable recovery of failures and concurrent use of the database, and for maintaining the integrity and security of the database. The recovery of failures assures the reliability of the database, the principal mechanism for achieving this is the use of transactions with 'undo' and 'redo' operations, and frequent backups of the database. Concurrency assures the correct execution of simultaneous transactions without loss of updates or other inconsistencies in database contents. Integrity of the database concerns the accuracy, correctness and validity of the database contents. The rules governing the integrity of the database can be divided into two categories: domain integrity rules and relation integrity rules. Protection of the database against unauthorized disclosure, alteration or destruction can be achieved by the introduction of authorization rules in the database system [23].

De Jonge and Schijf [55] discuss the performance of database access structures, and in particular of B-tree indexes. They consider the access speed of the index, dependent on the height of the index-tree, and the load factor of both the B-tree index and the sequence of records. Restructuring the index, splitting and recombination of index nodes, and concurrent retrieval and update of the database are further aspects taken into consideration in the assessment of different B-tree implementations [55].

Tharp [100] uses a number of metrics to determine the efficiency and effectiveness of different data access structures and data storage structures. The first metric is simplicity; if the alternatives are equal in all other respects the simplest access structure and storage structure must be chosen. Other metrics are: reliability, programmability, maintainability, space complexity and time complexity [100].

Data representation is also discussed by Van de Goor[39], in which a differentiation is made between scalar data types, structured data types, access data types and system data types. The set of values a scalar data type can assume (the domain) is characterized by two aspects: the range, which is the number of values that data can assume, and the precision, which is the distance between successive data values. Possible scalar data types are: integers (signed and unsigned), boolean numbers, characters, fixed point and floating point numbers. Criteria for the representation of integers are: ease of implementation of operations, uniqueness of the value zero and symmetry of range. Structured data types are composed of a number of elements with

the same data type (sets, arrays and strings) or with different types (records). Sets can be represented by bit vectors, and vectors and arrays are stored in such a way that they occupy a contiguous memory area. Multi dimensional arrays are stored using row-majoring ordering; that is, the rows are places in consecutive memory locations. Access data types are used to implement dynamic data structures (e.g. linked lists and trees), pointers are used to access elements of the data type. System data types are used by, for example, the operating system. Stacks and queues are important examples of system data types [39].

Mackenzie [63] provides insights into the factors that govern the design of computer coded character sets, both technical and economic. Factors that need to be taken into consideration are the number of bits used to represent characters (e.g. ASCII seven bits, EBCDIC eight bits), efficiency related to CPU byte size, bit numbering and which bit is first. How to deal with plus and minus zero, the collating sequence of characters (important for sorting) in the character set, use of the character set in programming languages (e.g. APL, PASCAL), and character set translations (e.g. from ASCII to EBCDIC and vice versa) [63].

Williamson [110] provides guidelines for the care and handling of the 3480 type tape cartridges. The report lists temperature and humidity conditions during operation, storage and transport of tape cartridges. The 3480 tape medium records data on 18 parallel tracks, data transfer rate is 3 million bytes per second, and approximately 200 million bytes can be stored on a single tape. The 18 tracks are divided into two nine track groups, each group consists of seven data tracks and two error check tracks. The error correction mechanism can correct up to three failed tracks in any nine track group and up to four failed tracks in the 18 tracks. The tape is expected to retain data for at least ten years [110]. Further recommendations for care and handling of computer magnetic storage media in general are made by Geller [36].

Performance of disk I/O subsystems is measured and modelled by Houtekamer [46] who has developed several queueing models for I/O service times of different IBM mainframe disk storage subsystem setups. Disk service time is the sum of six individual service time components: pending time, seek time, latency, Rotational Position Sensing (RPS) miss delay, path wait time and connect time. The pending time is the average wait time until a channel path and a storage director become available to a disk, and depends on storage system work load in particular on the device and controller utilization. The seek time is the average time needed to position the disk head on the cylinder requested. The latency is the time needed to reach the rotational position specified by the RPS mechanism, on average this will be half the rotation time of the disk. The RPS miss delay results from failed RPS connection attempts, this happens when no controller path is available. When an RPS miss occurs data transfer has to wait a full disk revolution. The path wait time is the average wait time for a reconnection for non-RPS I/O requests, it lasts until both a path and storage director are available. The last component is connect time, it consists of the command decoding time and the data transfer time [46]. Further literature on simulation of I/O subsystems that is not limited to IBM type storage devices can be found in [14, 87].

2.2 DATA

55

Additional discussion of IBM direct access storage devices (DASD) is provided by Johnson and Johnson [54] who describe characteristics of various generations up to the 3390 type DASD and RAID (redundant array of inexpensive disks) configurations. DASD management concerns, tuning considerations, systems managed storage and the use of various IBM software tools to assist DASD management are discussed [54].

Hoagland and Monson [44] consider two main objectives for magnetic storage technology: increasing storage density and reducing access time. Bit resolution determines linear density, and is determined by the distance between the magnetic-head and the medium. Actual physical contact (such as with tape recording) offers the highest potential recording resolution, but causes wear on the recording-head defined in terms of minimum number of passes guaranteed for the tape and hours of life for the magnetic head. The other factor that determines storage density is track density, which is determined by the capability of the servo system to position the magnetic head accurately over the track and by the signal-to-noise ratio. There are four major noise sources: the medium, the magnetic head, preamplifier electronics, and adjacent track interference. Bit error rate (BER) is an important characteristic of magnetic storage media, it can be divided into a hard error rate and a soft error rate. A hard error is caused by a medium defect while a soft error is caused by noise or interference (e.g. track mispositioning) and will not be repeated regularly on reread operations [44].

Mee and Daniel [68] provide detailed characterization of the main computer storage media: rigid disks, flexible disks, magnetic tape, helical scan tape, and magneto optical disks and they discuss performance and error characteristics of the various media in great detail. They also discuss the different signal modulation codes in detail, as well as the various error detection and error correction codes [68].

Semiconductor memories are discussed by Van de Goor [40]. DRAM chips with a storage capacity of 64 Mbit and an access time of 50 ns are available, however they need to be 'refreshed' about every 4 ms or else the data is lost. Faults in semiconductor memory can be divided into two kinds: permanent faults and non-permanent faults. Examples of permanent faults are: incorrect connections between ICs, boards etc., broken components, incorrect IC masks and functional design errors. Non-permanent faults can be divided into transient and intermittent faults. Transient faults are often called soft-errors and are caused by environmental conditions such as: cosmic rays, α -particles, pollution, humidity, temperature, pressure, vibrations, power supply fluctuations, electromagnetic interference, static electrical discharges, and ground loops. Intermittent faults are caused by non-environmental conditions such as: loose connections, deteriorating or aging components, critical timing, resistance and capacitance variations, physical irregularities, and noise. Failure mechanisms can be classified into: electrical stress, intrinsic failure mechanisms (crystal defects caused during wafer fabrication) and extrinsic failure mechanisms. The failure rate depends exponentially on temperature which is why temperature is the most important stress condition. Due to small cell capacitance, high capacity DRAM chips suffer from soft-errors caused by α -particles, special coatings have reduced this problem. For high capacity DRAM chips, noise is a serious problem which can be reduced by twisting

and shielding the bit lines. Power consumption can be reduced and reliability can be increased by lowering the supply voltage [40].

2.2.4 Distribution of data

Much attention is given in the literature to the implementation of distributed database systems and the related algorithms to be used in software, such issues are outside the scope of this thesis. Distribution of data is discussed in the literature in terms of the databases and the database management systems that are used in the deconcentrated storage of data at several physical locations as illustrated by this definition given by Date [23]:

A distributed database is a database that is not stored in its entirety at a single physical location, but rather is spread across a network of locations that are geographically dispersed and connected via communication links.

As a consequence the sections which follow are concerned mainly with the deconcentration of data. In this thesis deconcentration of data and decentralization of data are considered to be two types of data distribution (see page 19) and this why they are referred to together as data distribution in the following sections.

In the next section the reasons why data might (or should) be distributed are discussed and the various forms of data distribution are discussed in the second section. The objectives for data distribution are discussed in the third section, and the alternatives for distributed databases are discussed in the fourth section. In the fifth section the architecture for distributed databases is discussed.

Reasons for data distribution

Both concentrated storage of data and deconcentrated storage of data have their advantages and disadvantages, which have to be taken into consideration when deciding to implement data distribution. The advantages and disadvantages of concentration and deconcentration for each of the arguments found in the literature are given below.

Local autonomy Concentration of data is advantageous when the data are used by central corporation applications such as payroll, purchasing or general accounting, or when users in all parts of the corporation need access to the same data and need the most recent data values [67]. Deconcentrated data permits greater user control of data [31]. Deconcentration of the database allows individual groups within the enterprise to exercise local control over their own data, with local accountability, and more generally makes them less dependent on some remote data processing center that cannot be deeply involved in purely local issues [23]. Distributed databases fit the structure of decentralized organizations more naturally [17], as this permits the setting and enforcement of local policies regarding the use of the data [25, 83].

2.2 DATA

57

Local data management Data management activities are better coordinated and controlled when performed at a single location than activities performed at a number of remote locations. It is easier to recruit qualified data management personnel, skilled management personnel is more likely to work at a large computing centre than at a small remote computing centre. The opportunities for promotions are greater at a large computing centre and this factor attracts more motivated employees. Forms, programming languages and operating procedures within the data management organization can be standardized. User utilization of data can expand with lower increases in overall use of automation facilities and personnel. It is easier to train new data management personnel since a larger staff can absorb trainees more quickly. More sophisticated automation facilities can be employed because higher qualified personnel is available.

Data management staff at the computer centre are unfamiliar with the local users' problems and they find it difficult to relate to the local managers, as a result local users may not accept the database systems designed at the computer centre. As local users cannot control priority over data processing jobs at the computer centre, they may not receive timely reports to support decisions. If the central computer crashes, the entire information system is down and cannot be used [18].

Autonomous local computer centres, with their own database, require their own data management function. While this might not be a full time task it is a skilled job requiring trained personnel [25]. Local computer centres are more sensitive to user needs resulting in a rapid response to local data needs. Information systems developed locally are more likely to be accepted by the users. The local sites do not have to compete with other locations for data processing and high priority data processing can be accommodated providing managers with timely reports. It is easier to identify costs and proposed benefits with specific local information systems, and local computer centres are easier to manage from a budget standpoint. It is more difficult to recruit data management personnel to a small computer centre, as there are fewer opportunities for promotion at the local computer centre. Distribution of data over multiple computer centres may result in more complex control and coordination problems and standardization problems [18].

Interconnection of existing databases Distributed databases are the natural solution when several databases already exist in an organization and the necessity for performing general data processing exists. In this case, the distributed database is created bottom-up from the preexisting local databases. This process may require some local restructuring; however, the effort required for restructuring is less than that needed for the creation of a completely new centralized database [17]. Distributed databases enable sharing of data among geographically dispersed organization operations [83], and enable a corporate view of data [25]. It is often desirable to integrate databases that are implemented using database management systems supplied by different vendors and that support different data models. The distributed database can hide the different data models supported by these heterogeneous databases [57].

Capacity and incremental growth A common reason for implementing deconcentrated databases is the fact that no single computer system has the capacity to store all the data, and growth can proceed in smaller increments than with a single concentrated database. If it becomes necessary to expand the the database because the volume of data or the amount of use of the data has increased, then it may be easier to add a new site to an existing deconcentrated database system than to replace a concentrated database system by a larger one [23]. When an organization grows by adding new, relatively autonomous organizational units, perhaps by acquisition, growth can be achieved with minimal impact on the existing units by adding new sites to the network [17, 25]. Expansion can be handled without the need for major system overhauls, however, distribution overheads also mean that not all installed computing power is available for data processing [83].

Performance When data are updated frequently, data may be concentrated to avoid the problems of real-time synchronization of multiple copies with a high update frequency [67]. Communication between sites involves a number of overheads, which when added together, often mean poor response times. Such overheads may be acceptable when communication interactions are few, but in a distributed database setting in general, a considerable number of messages are involved with fetching and storing data from each site [25]. Distributed databases have the advantage that the decomposition of data reflects utilization dependent criteria which maximize data utilization locality; in this way mutual interference between data processing by information systems is minimized. The workload is shared between the different computer systems, and critical bottlenecks, such as the communication network or common services are avoided [17, 25]. Data that is used at one location only, and rarely or never used at other locations can best be stored at or near the location. When the update frequency is too high to be processed by a single concentrated storage system distribution over several computer systems can enable updates to be processed in time. Complex data operations may cause loss of performance of the single database system, and may be better performed in a local database system with local responsibility for use and costs [67]. Data retrieved by a transaction may be stored at a number of sites, making it possible to execute the transaction in parallel [83].

Reliability and availability Duplicated systems components can provide improved reliability for concentrated data; however, disasters such as fire or floods can render the data inaccessible [57]. A deconcentrated database offers greater reliability than a concentrated database in that it can function (at a reduced level) in the event of failure of a site or of a communication link. If data is replicated, availability is improved because the replicated data remains available as long as at least one replica remains accessible. Each replica can be used as a free backup of data that happens to be destroyed by a failure [23]. A deconcentrated database does not by itself guarantee higher overall reliability, but it ensures a graceful degradation property. Graceful degradation means that failures in a deconcentrated database can be more frequent than in a single concentrated database because of the greater number of components,

2.2 DATA

59

but that the effect of each failure is confined to those information systems that use the data of the failed site, and a complete systems crash is rare [17, 25, 83]. Redundant data storage and redundant data processing capabilities can be used to obtain higher reliability and availability; however, obtaining this goal is not straightforward and requires the use of techniques which are still not completely understood [57]. Deconcentration unfortunately creates problems of synchronization and coordination, and distributed control of data transactions can therefore easily become a liability if care is not taken to adopt adequate policies to deal with these issues [17].

Efficiency and flexibility Concentrated data has a greater economy of storage, with good backup facilities at a central site [31]. It is cheaper to concentrate data when users of data travel among many separate locations, this eliminates overlapping and duplication of data and thus lower data storage costs [67]. A single large computer provides economies of scale with lower overall information systems development costs and information systems management costs [18]. Deconcentrated data permits fast access to data, and less database contention problems [31]. Response times and communication costs can be reduced when data is stored close to the point where they are used most [25]. If the pattern of use changes, data can be moved or replicated, or existing replicas can be removed when no longer needed [23]. Increased redundancy in data storage and the fact that a number of small databases may cost more than a single large database, means that economies of scale may not be realized. Increasing developments in computer systems technology means that the economy-of-scale argument no longer applies, and the reasoning behind one large computer system becomes [17]. It costs less to put together a system of smaller computers with the equivalent computing power of a single large computer [83]. While computer costs have fallen dramatically, communication costs have not. In particular, where high performance is required, expensive high bandwidth communications is necessary [25]. Perhaps the most important cost component is the replication of effort in data management. When automation facilities are set up at different sites, it becomes necessary to employ more people to maintain these facilities. This results in an increase in the number of personnel in data management operations. Therefore, the trade-off between increased profitability due to more efficient and timely use of data and the increased personnel costs has to be analyzed carefully [17].

Security Concentrated data has the advantage that a large amount of control can be provided over access to data and better data consistency. In a central location security can be controlled more easily, with ease of internal control and audit [31, 83]. When a high level of data security is necessary concentration is advantageous. Protection procedures may be expensive, possibly involving well-guarded, secure, vaults, and tight control of authorized access. Data are more easily guarded if they are in one location, with external backup copies, than if scattered. Audit trails, that keep details of what transactions update certain data, are more securely stored at a single location. When accuracy, privacy and security of data is a local responsibility deconcentrated data are better suited to the structure of decentralized organizations [67]. With

deconcentrated data a network is involved and this is a medium that has its own security requirements. Maintaining adequate security over computer networks is a serious problem, and thus the security problems of deconcentrated data are by nature, more complicated than those of concentrated data [83].

State of technology Relatively few available commercial products have an established track record in distributed database operation, and many do not support essential distributed database capabilities such as integrated recovery and the performance characteristics of others remain poor. No method available offers a comprehensive and proven approach to distributed data [25]. Distributed data has problems which are inherently more complex than concentrated (and centralized) data, as they include not only the problems associated with concentrated data, but also a new set of (as yet unresolved) problems. General purpose distributed database technology is not yet commonly used; what we have are either prototype systems or systems that are tailored to one application (e.g. airline reservations). This means that the solutions proposed for various problems have not been tested in an actual operating environment [17].

Forms of data distribution

Distribution of data can be divided into seven forms. The first three forms of data distribution concern the coherence between data stored in different databases, these forms are, in increasing order of data coherence: incompatible data, separate schema data and reorganized data. The remaining four forms of data distribution concern the deconcentration of data storage at different locations, these forms are, in increasing order of data decomposition: replicated data, subset data, partitioned data and fragmented data.

Incompatible data Independent information systems are set up by autonomous departments to provide for their specific needs. In this process oriented approach (see page 41) each department follows its own course and defines and implements its own data without coordination [66]. Someone who wants to create a summary of all invoices that still have to be paid by the customers has to access multiple, separately developed, databases. They have to learn where the invoices are kept, e.g. one department keeps the data required in a customers database and a different department keeps the data in a sales database to learn the various ways to gain access to the data and how to interpret and relate the data retrieved from the different databases.

Separate schema data Separate schema data are part of different information systems and have different data definitions. Although separate database schemas are used, the data originates from a data oriented approach where common use of data is emphasized (see page 41). In this manner inconsistency and unnecessary redundancy of data is avoided and the separate databases appear as an integrated whole for someone who wants to access the data. An example of separate schema data can be

2.2 DATA

61

found in the personal data register case (see section 6.3.3) where regional copies of personal data have a simplified database schema compared with personal data kept in the main database.

Reorganized data Reorganized data have the same data model, and consequently have the same data definitions stored in the data dictionaries; however, different kinds of use of the data in information systems require the data to be stored and accessed in different manners [66]. Reorganized data improve performance of specialized on-line transactions or batch processing by employing a variety of data structures such as inverted lists, secondary indexes, different sort criteria, or other storage methods.

Replicated data Replicated data refers to the keeping of identical copies of the same data at different locations [66]. This means that any logical data object can have several distinct stored representatives, stored at several distinct sites [23]. The main reason for replicating data is the reduction in transmission of data between different sites, this is becoming more attractive as the cost of storage units is dropping faster than the cost of telecommunication. Replication makes sense only when the frequency of data retrieval is much higher than the frequency of data updates [66].

Subset data Subsets of data are stored locally for ease of data entry, retrieval or local processing. In a data-entry operation, data are often keyed into a local computer and are checked there, accuracy controls and auditing controls are applied to a batch of local data, and this batch of local data is then transmitted to a large main database [66].

Partitioned data Partitioned data implies that the same database schema is used in two or more databases, but each database stores different data [66]. This means, that corresponding to any logical data object, there is precisely one physically stored representative of that object, at precisely one site in the network [23]. There may be separate databases in each branch office of an organization, each retail store, or other such location. They will all use the same information systems and transactions originate and are processed in the same location as their local data [66].

Fragmented data A stored database relation may be partitioned into separate fragments, and copies of those fragments may be stored (replicated) at multiple locations. A fragment is obtained by applying an arbitrary combination of projection and restriction to the stored relation [23]. Fragmentation of a relation along its tuples (or rows) by applying restriction is called horizontal fragmentation, fragmentation of a relation along its attributes (or columns) by projection is called vertical fragmentation [83]. Primary horizontal fragments are defined using selections on the relations; the correctness of primary fragmentation requires that each tuple of the general relation be selected in one and only one fragment. Derived horizontal fragmentation of a relation is not based on properties of its own attributes, but is derived from the horizontal fragmentation of another relation. Derived fragments are used to facilitate joins between fragments [17].

Objectives for distributed data

Implementation of data distribution requires some kind of distributed database management system. The design considerations and implementation of distributed database management systems are outside the scope of this thesis, but the underlying objectives and architecture of distributed databases are important for assessment of proposed solutions to distributed data by data management staff.

The fundamental principle for distributed database management systems is defined by Date as follows [21]:

To the user, a distributed system should look exactly like a non-distributed system.

Date derives twelve objectives for distributed databases from this fundamental principle, further discussion and explanation can be found in [21]:

1. Local autonomy
2. No reliance on a central site
3. Continuous operation
4. Location independence
5. Fragmentation independence
6. Replication independence
7. Distributed query processing
8. Distributed transaction processing
9. Hardware independence
10. Operating system independence
11. Network independence
12. Database management system independence

Relational database technology is considered to be relevant specifically to the objectives 'fragmentation independence', 'distributed query processing' and 'database management systems independence' and together with the flexibility of relational systems they give rise for the claim that "...for a distributed system to be successful, it must be relational" [21]. Notwithstanding this claim, the largest and most successful distributed database found during this research was based on CODASYL database technology (see chapter 6).

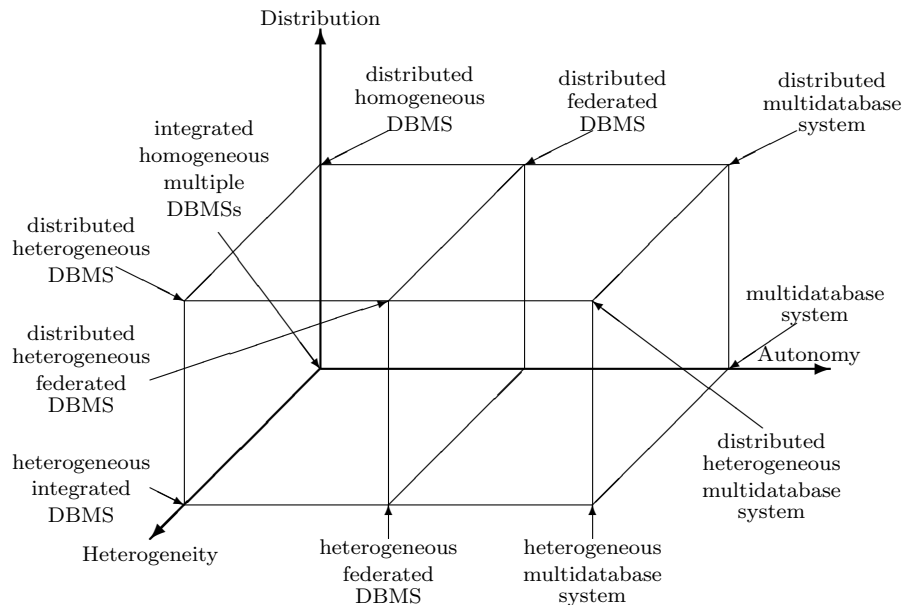


Figure 2.4: Distributed database management systems implementation alternatives

Alternatives for distributed databases

There are a number of ways in which multiple database can be put together into a distributed database management system that appears to behave to the user as a single database. A classification of distributed database management systems can be defined by characterization based on: (1) autonomy of local databases, the degree to which individual database management systems can operate independently, (2) their distribution, the different physical locations, and (3) heterogeneity, whether local database management systems consist of different kinds of database technology. Based upon this classification a diagram can be drawn positioning the different implementation alternatives, see figure 2.4 [83]. Other implementation alternatives may exist that are not shown in this diagram. The underlying fundamentals of distributed data are however, no different for the various implementations. An underlying general architecture for distributed databases is important for data management to to compare and asses the possibilities and limitations of the different distributed database management systems.

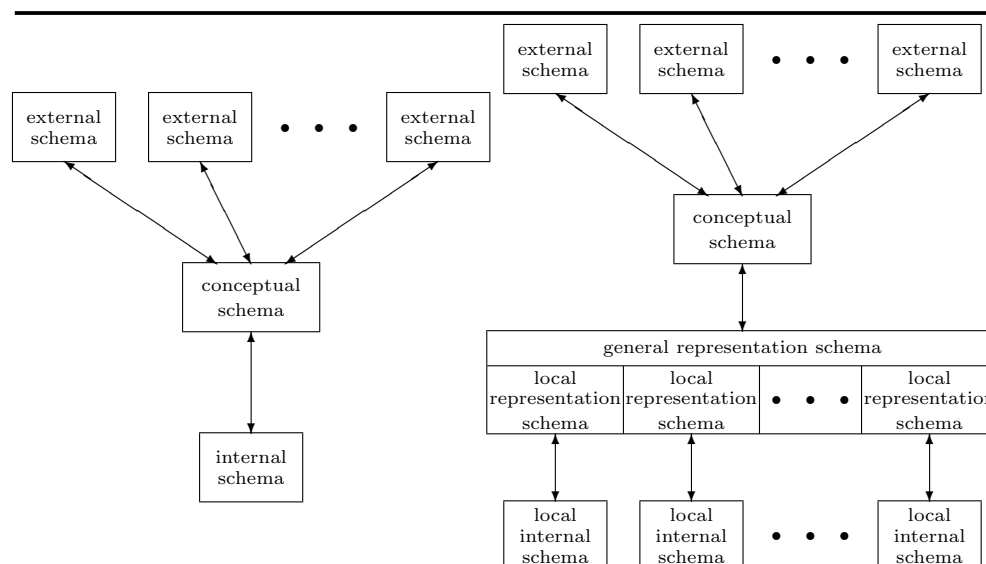


Figure 2.5: Three schema and five schema architecture for databases

Architecture for distributed databases

The ANSI/X3/SPARC study proposed an architecture for non-distributed databases that included three schema levels: internal, conceptual and external, motivated by the need for program data independence and physical data independence. The internal schema describes the data storage structures and data access structures. The conceptual schema describes the relationships between data including integrity constraints or restrictions. A mapping, or translation, relates the internal schema and the conceptual schema, and together the two schemas and the mapping achieve physical data independence. The internal schema can be modified to optimize performance without altering the conceptual schema. There are numerous external schemas that describe the different user views of the data, and external schemas can be added at any time to support new information systems. Several mappings relate the conceptual schema to any of the external schemas, and the conceptual schema and the external schemas together with their respective mappings achieve program data independence. The conceptual schema can be extended with new data to support new information systems without effects on existing information systems. Major changes in the conceptual schema however, may have an effect on the external schemas and must be executed carefully by data management[28]. The three-schema architecture is shown on the left-hand side of figure 2.5.

The five-schema architecture is an extension of the three-schema architecture

2.3 MANAGEMENT

65

needed to facilitate the twelve objectives for distributed data, listed above. The five-schema architecture adds a general representation schema and an local representation schema to the three-schema architecture to facilitate distributed transaction processing and heterogeneous local database management systems. The general representation schema contains descriptions of the partitioning and replication of data among the local databases, this allows a transaction to select the required local data copies and to recognize opportunities for parallel execution. The general representation schema is also used to coordinate the execution of data manipulation commands by the local database management systems, ensuring that the transaction either terminates successfully or aborts and restores each of the local data copies. The local representation schema is used to translate the data manipulation commands for execution by the local database management systems. These translations take advantage of the special capabilities of the local database management systems and of the knowledge of the local data storage structures and data access structures [28]. The five-schema architecture is shown on the right-hand side of figure 2.5.

Other extensions proposed to the three-schema architecture achieve basically the same effect as the general and local representation schemas. One proposal replaces the conceptual schema with a general conceptual schema and local conceptual schema [83]. The another proposal adds a fragmentation schema and allocation schema that perform the same function as the general representation schema, and adds local mapping schemas to each local database management system that perform the same function as the local representation schema [17].

2.3 Management

This section describes data management at the strategic, tactical and operational level. The strategic level identifies preconditions from outside the data management organization, it formulates the goals it must achieve, and translates this into policy guidelines. At the tactical level data management acquires the material, personnel and financial means needed to manage the data. At the operational level data management performs the tasks needed for the direct operation, control and maintenance of data. Many authors fail to make the distinction between tasks at the strategic, tactical and operational levels, and the tasks they describe are not limited to the states utilization (U) and exploitation (E) of the state model, but include policy and planning (IPP), development (D), acceptance and implementation (AI) and maintenance (M); see also the discussion on page 22.

Tasks at the strategic level of data management are discussed in section 2.3.1, tasks at the tactical level of data management are discussed in section 2.3.2, the tasks at the operational level of data management are discussed in section 2.3.3. The distribution of the data management organization that performs the data management tasks is discussed in section 2.3.4.

2.3.1 Strategic level

It is important to make a distinction between the data strategies of a bank or airline company, and the strategic level of data management. The first type of strategic importance is related to business opportunities for the bank or airline company by making proper use of the data it has available. The strategic level of data management is related to the policies and objectives the data management department has in relation to such external factors as technological improvements, suppliers of data management tools, and the business economic constraints and organizational constraints of data management. This last strategic aspect is of importance for this thesis. Few authors actually define a strategic level of data management, but instead describe strategic tasks together with tactical and operational tasks.

According to Grover [42], tasks of a strategic nature have to do with the preparation of strategic data plans and technical data plans, and the identification of opportunities for data sharing and potential database applications. Providing consultancy and training to IS personnel and users is also considered to be strategic, as well as the development and enforcement of policies governing data ownership and access control. With regard to corporate data for common use, tasks are defined that deal with organizational data resource requirement planning, and the development of an organizational data model and data architecture.

The strategic level of data management is according to Mercken [69] concerned with the selection of DBMS architecture (hierarchical, network, relational, object oriented) and with the selection of software components such as: DBMS, data dictionary/directory system, repositories and case-tools. Decisions with respect to distribution of the database, client-server architecture and strategies regarding new releases and developments are all taken at the strategic level of data management.

2.3.2 Tactical level

The tactical level of data management is concerned with acquiring the personnel and material and financial means needed to perform the operational tasks. Many authors do not make the distinction into strategic, tactical and operational and consequently many of the tasks that have to do with the acquisition of personnel, material and financial means end up on the operational level. When tactical tasks are defined by authors, they are only concerned with the material aspects of data management, not with the personnel and financial aspects.

Grover [42] attributes tasks for the evaluation and selection of hardware and software to data management at the tactical level. Further, tactical data management must set and enforce operational procedures and standards and procedures for data retention. The model of information systems management processes provided by Van Schaik [89], identifies data planning as the single tactical process for data. An historical overview of information management (IM) is given by Ortner [82], in which four phases in the history of development of data management in organizations are

2.3 MANAGEMENT

67

identified: file management, database management, data management and 'information management'. Tactical tasks in the 'information management' phase of data management are grouped by Ortner under the names: management of 'information processing', management of the 'information resource' and development of the 'information resources' dictionary system.

2.3.3 Operational level

The operational level of data management is concerned with maintaining the data in the states Utilization and Exploitation of the state model. Tasks to be performed at the operational level of data management are mentioned by a number of authors. Attention however is focussed on the definition of functionaries and organizational positions instead of on the more general applicable tasks and consequently, there are a large number of tasks attributed by authors to a 'database administrator' (DBA) or 'data administrator' (DA). Not all tasks attributed to the DBA or DA are strictly at an operational level, because the distinctions between operational, tactical and strategic levels of management are often not made.

Two approaches exist to structuring data management tasks at the operational level: the first approach groups related tasks and processes into task clusters or task fields. This approach is independent of organizational aspects such as implementation of data management in units or departments, and matters of centralization or decentralization. Task clusters are discussed in the first section. Many authors take a second approach and attribute data management tasks to functionaries. In this sense the tasks become part of a job description and fix a particular organizational implementation of data management. The most often cited job description is that of a database administrator and this described in the second section, all other job descriptions of data management functionaries are discussed in the third section.

Task clusters

At the operational level of data management tasks are performed that are involved directly in the operation, control and maintenance of data. According to Grover [42] these tasks encompass the control of data redundancy and maintaining accuracy, integrity and security of data. Conducting logical database design and establishing and maintaining a data dictionary, and physical database design are also tasks at the operational level of data management. In the model of information systems management processes provided by Van Schaik [89], two operational processes are described with regard to the control of data. These two processes are inventory and performance control of resources and data.

With regard to distributed database management systems, Walker [106] makes a distinction into management of the distributed system as a whole and management of individual (local) database components. The components can have widely differing characteristics, and management of each component may require mastering

a specialized set of procedures and utilities to perform the data management tasks. Performing data management tasks for the distributed system as a whole requires coordination among components. Walker [106] groups the data management tasks involved with administering the distributed database management system into nine areas. These tasks need to be performed for the distributed system as a whole and for each individual component.

The first task area defined by Walker concerns communication with user, management, operations, and maintenance groups. It contains tasks for maintaining a channel of communication to each of the parties concerned, and the establishment of a formal trouble reporting system. The third task in this area concerns analysis of the trouble report and routing it to the appropriate support personnel for resolution. Monitoring hardware and software is a task area in which tasks are positioned that concern monitoring status and report failures of hardware and software, planning and coordinating replacement and upgrading of hardware and software, and hardware maintenance (including processor and peripherals). Testing is a task area defined by four tasks: perform acceptance testing, schedule database structural and consistency checking, coordinate accuracy checking, and maintain internal software tables. Security is a task area that considers the data management responsibilities: consulting management about security considerations and monitoring security, review operations procedures in the light of protection of databases, and oversee physical protection of database disks, tapes and other computer media. Three tasks are defined for the task area backup and recovery: schedule backup procedures, test effectiveness of backup procedures, and develop and test contingency plans. Performance is a task area which has tasks to assign priorities to activities in the database such as transactions, report generation, to set installation and generation parameters, to monitor system performance tuning the system as required and, to monitor database size and growth, executing database reorganizations when appropriate. Training is the task area where all the training activities of data management personnel are scheduled and coordinated. Documentation is the task area where a complete list of the release numbers of all components and software is compiled and maintained. Scheduling is a task area containing tasks for scheduling regular activities such as database integrity checking (both structural and consistency), backup procedures and preventive maintenance. A separate task is defined for scheduling of special activities such as a switch over to new a software release, installation of new hardware or modifications to existing hardware, and special user requests.

At present a great deal of attention is being paid, within The Netherlands, to the "IT Infrastructure Library" (ITIL) [15]. The IT infrastructure library has quickly become the guideline for structuring processes in computer centre operations in The Netherlands. ITIL is a commercial product which consists of a series of (at present) about forty booklets, each booklet comprises a module in which points of attention and rough outlines of processes are given for specific topics of computer centre operations. Module examples include: managing facilities management, capacity planning, help desk, unattended operations, cable infrastructure strategy, and fire precautions in IT

2.3 MANAGEMENT

69

installations. An IT infrastructure is defined by ITIL as: the hardware, software, computer related communications, documentation and skills required to support the provision of IT services [16]. From this discussion it is clear that data management is not one of the topics of interest for ITIL and is consequently not discussed in any booklet from the series.

Data management tasks are clustered into task fields by Looijen [59] based upon their nature and coherence. A comparable approach is taken by the Dutch Society for Informatics (NGI) which defines 86 different task clusters for information systems management [77]. Task clusters concerning data management can be positioned within the task fields model defined by Looijen, and they will be discussed together below. Activities within task clusters are, however, not all confined to Utilization (U) and Exploitation (E) of data, see the introduction to this section, and some task clusters contain activities that are attributed to different task fields, causing overlap between tasks fields and task clusters.

Looijen [59] distinguishes five task fields within information systems management that concern data: functional data management, data administration, database administration, physical data management, and data processing.

Functional data management is a task field aimed at authorization of data utilization, management of databases with regard to integrity and removal of data and drafting reports on the nature and frequency of utilization [59]. NGI task clusters positioned within functional data management are: drawing up the object model, drawing up the data model, management of meta-data, management of data and management of software parameters [60, 77].

Data administration is a task field responsible for monitoring the integrity and proper application of data, further it provides advise on available data as well as their meaning to promote access to data and unambiguous utilization of data [59]. Data administration manages the data descriptions [84] and its task further comprise the establishment of a data architecture, a naming policy, information modelling, defining reference data elements, configuration management, and the security, integrity, reliability, and archiving of data [86]. Data administration is also involved in the clean-up of data definitions and the control of shared data definitions, management of distributed data and supervision of security functions [95]. The NGI task clusters positioned within data administration are: drawing up the object model, drawing up the data model, management of meta-data and performing data conversions [60, 77].

Database administration is a task field that performs maintenance of database structures, being the technical translation of business data and relations between them. Maintenance is aimed at error recovery of databases, interfaces between databases and application software and maintenance of software used for removing data, copying of data and conversion of data [59]. Database administration performs tasks for the implementation of the database and monitors the performance of the database [84], and involves the control of internal names, performance improvement, security, backup and recovery, concurrency, and data redundancy [86]. The NGI task clusters positioned within database management are: drawing up the object

model, management of meta-data, drawing up the implementation data model, performing data conversions, quantify information system capacity, management of data storage components, determining systems maintenance and information system evaluation [60, 77].

Physical data management is a task field responsible for the physical management of databases, the copying of databases and the recovery of databases after calamities [59]. No NGI task clusters can be positioned within physical data management [60, 77].

Data processing is a task field that takes care of processing data by the information system with respect to performing: input of data, processing and storage of data, transportation of data, and output of information [59]. No NGI task clusters can be positioned within data processing [60, 77].

Database administrator

The database administrator (DBA) is probably the most cited functionary within data management. Many more authors provide job descriptions for the DBA than can be presented here, and therefore a selection has been made to illustrate the wide range of tasks attributed to the database administrator. Such a large number and large diversity of tasks place heavy demands on the person that has to perform them, some authors have recognized this and have drawn up job descriptions for other functionaries than the DBA, these will be discussed in the next section.

Mercken [69] recognizes at the operational level of data management the function database administrator where tasks are performed concerning: physical design and maintenance of the physical database; loading, reorganization and recovery of the database; and evaluation of database exploitation and suggestions for improvement.

Date [24] considers the database administrator to be the person or group of persons, responsible for the overall control of the database system. It is the DBA's job to decide what data is to be held in the database about the entities of interest to the enterprise; having done this the DBA must then define the contents of the database by writing the conceptual schema. The DBA must also decide how the data is to be represented in the database and data is accessed, the DBA must specify this by writing the storage definitions, in addition the DBA has to specify the mappings between storage structure definitions and the conceptual schema. By communicating with users the DBA can ensure that the data required is available, this is achieved by writing the necessary external schemas. Additional mappings between external schemas and the conceptual schema are also given by the DBA. Authorization checks and validation procedures can be considered to be logical extensions to the conceptual schema, and are therefore specified by the DBA. The DBA must define and implement an appropriate strategy for backup and recovery of the database to be able to repair damaged data with a minimum of delay and with as few of unwanted side effects as possible. Monitoring performance and responding to changes in requirements complete the tasks of a DBA.

2.3 MANAGEMENT

71

Digital's VAX DBMS design guide [29] gives some indications to database administrators of what they should do to manage their database systems. Tasks listed include design and implementation of the database, tuning the database and measures that may be taken to ensure the security of the data. A DBA also has a coordinating role within the information systems management department supervising the design of applications to ensure proper use of the database, and supervising database maintenance operations. Modifications to the database can also be applied by the database administrator as well as actions taken to maintain database integrity. The final tasks mentioned in the VAX DBMS design guide are monitoring database performance and operating the database.

Völlmar [104] attributes seventeen different tasks to the database administrator functionary. These tasks are cited here in full because the detailed descriptions show the wide range of duties of a database administrator. The database administrator monitors developments in the field of database exploitation, determines their consequences for the organization, and communicates these findings to other functionaries for who this may be of interest. The DBA develops general rules for describing data and for procedures of recording actual data values, rules for inspection of data, and rules for data security and retaining data, adapting these rules when circumstances require this. The DBA determines, together with the automation department executives, what consultation bodies and procedures are desired. The DBA gives information and instruction courses in databases. After preliminary consultation with users and systems designers, the DBA draws up a planning for transition to databases or extension of database applications, including conversion of software. Based on that planning the DBA then makes an estimate of the hardware capabilities required for the computer centre executives to plan overall hardware requirements. The DBA assists owners of data in determining which users are authorized, and in determining procedures for recording actual data values and related responsibilities, further the DBA assists the owners of data in determining data retention periods and data security requirements. The DBA assists systems designers and the data administrator in determining physical data structures and the logical data structures, both in the initial stages and in later modifications and additions. The DBA determines, after deliberation with systems designers and the database management system specialist, the technical composition of the databases; the authority to make such decisions rests with the DBA because the general efficiency of database use are the DBA's responsibility. The DBA determines, together with the systems designers, how every database will be fed with new data, how reorganizations will be carried through and how historical data will be recorded and grouped. The DBA determines which security measures will be implemented to satisfy as much as possible all security requirements, and issues security instructions and regulations. These instruction and regulations apply to both protection against unauthorized use and reconstruction of corrupted database contents. The DBA audits adherence to all security measures by the persons involved. The DBA administers the central collection of meta-data in the data dictionary, issues the meta-data and makes sure that the provided descriptions

of data that are supplied are complete and satisfy the requirements regarding quality and completeness. Together with the database management specialist, the DBA monitors the development in processing times, the response times of on-line use, and when necessary performs reorganizations or modifications of the technical composition of the databases. The DBA assists in tracing malfunctions and analyses their nature and the frequency with which they occur, to be able to remove possible causes of repeated malfunctions. Together with the systems designers, the DBA puts together testing facilities to inspect the internal structures of the databases, and initiates tests when necessary. The DBA further authorizes providing of copies of databases or parts of databases for the purpose of testing application programs.

Clark [20] approaches database administration from an auditing viewpoint. A database administrator is according to Clark a technical person who has responsibility for the maintenance of the DBMS and the physical databases within the organization, and may also have responsibilities for the logical structures of the database. Tasks include database design and development and being responsible for organizing and defining the necessary “views” of data, especially in relation to data stored in DBMSs. A database administrator provides education on database technology and support to users in operational database related activities. Optimization of physical data storage and the development of data recovery procedures for the database are also part of the job description. As final tasks, the database administrator performs control of data security mechanisms and data privacy mechanisms [20].

Other functionaries

As well as the considerable attention paid to the database administrator, some authors go to the trouble of drawing up job descriptions for other functionaries because they recognize that all the tasks mentioned above cannot be performed by the same person, or same kind of person. Other functionaries, or ‘executives’ [67], mentioned are: data strategist, data administrator, data dictionary/directory administrator, database designer, data operations supervisor and security officer.

A data strategist performs tasks pertaining to enterprise-wide planning of data to achieve uniformity in data definitions throughout the enterprise, where possible. Educating top management about the need for strategic planning of data and the enforcement of common data definitions and models is another task of a data strategist, and paying attention to the handling of human and political problems with regard to the establishment of common data throughout the enterprise. Establishing common data is achieved by assessing future requirements that may lead to adaptations or additions to data models, and developing an entity-relationship overview chart. The data strategist can control the use of common data by clustering entities into subject database and coordinating of the work of separate data administrators [67].

The function of data administrator is recognized by many authors, and who sometimes devote a complete book to the data administrator, for example Durell [30], and consequently many tasks are defined. Durell’s job description of a data administrator

2.3 MANAGEMENT

73

contains tasks for naming, definition and relating of data entities. A data administrator further defines data classes from the data entities and defines subject databases to implement these data classes. A data administrator makes the logical database design that is implemented according to plan, the data structure design needs to be approved by the data administrator. Additional advising and assisting in the design of other data structures are also tasks of the data administrator [30]. Clark [20] approaches data administration from an auditing bias, and according to Clark a data administrator is responsible for the management of the corporate data including ownership, access rights and definitional matters. According to Mercken, a data administrator designs conceptual schemas, defines directives and procedures for data management, and determines the logical design of systems for security, backup, recovery provisions and integrity inspection of data [69]. Martin [67] gives an extensive list of tasks to be performed by the data administrator. A data administrator is involved in setting up end-user committees to be involved in data analysis, modelling an definition, and collects end-user views and systems analysts' views of what should be in the database. A data administrator further performs data analysis to determine what should be in the database, defines the data-items, and performs data model consistency analysis and stability analysis, to determine which application programs need to be rewritten when data-item definitions change. The feedback of the model and dictionary output to the end-user committees ensures that these meet the present and future needs as far as possible. The data administrator identifies data-item synonyms and their representations in the data dictionary and tries to eliminate them where practical. The data administrator gives advice to programmers, systems analysts, and user departments concerning data and their logical structure, and attempts to maintain a view of the future needs and applications, so that the database can be made to evolve in an appropriate manner. A data administrator defines rules and functions to ensure validity, consistency, and accuracy of data, and may determine which users are responsible for the accuracy of data. Special security protections for certain records or data-item can be defined, and the data administrator may determine which users are authorized to see, create, update or delete data. The data administrator further plans for compatibility with existing data structures or for conversion of these structures. The data administrator develops appropriate reports from the data dictionary, including cross-reference lists stating which programs use what data-items. When changes are proposed, the cross-referencing will show what has to be modified. When users, programmers, or systems analysts require data that are not in the database, the data administrator determines how (or whether) they can be incorporated. The data administrator may determine what policies govern the retention of data, for example, to comply with government regulations [67].

A data dictionary/directory administrator is a functionary defined by Mercken [69]. The job description contains tasks related to the data dictionary/directory system (DD/DS). These tasks comprise the design of DD/DS schema to give concrete form to standards for naming of meta data, quality standards and their compliance; the design and implementation of a system to enforce meta data consistency;

the verification of authorized modifications of DD/DS software or DD/DS databases; documenting meta data and maintaining a audit trail; fulfilling information requests of authorized users; and finally tasks relate to development of technical functions such as backup and recovery of a DD/DS and to commission software maintenance [69].

A database designer is a functionary who is responsible for the design of conceptual schemas and coordination of subschema development [69]. Martin [67] gives an extensive job description for database designers in which twenty tasks are identified. The database designer is responsible for physically structuring the data and for planning access methods. The database designer employs design tools such as mathematical models of the physical storage to help determine which methods of physical organization best meets the performance criteria. The database designer assists in the selection of which database hardware and software should be used, and specifies any additions to the database software that might be needed. The data designer assists the programmers when using the database, and generates their data descriptions from the data dictionary, and checks their programs to ensure that they use the database efficiently. A database designer is responsible for the design of the means for restart and recovery after system outage, the means of backup of data, and the means of reconstructing data in the event of loss of records or of catastrophic destruction of entire files. Using the logical model of the database, the database designer designs the software schemas, alone or together with systems analysts or programmers in the user department that may be competent file designers. Techniques for monitoring database performance are specified by the database designer to correlate the actual database performance to the predictions of the design models, in this way the database designer can react to future performance problems before they arise. Measurements of how full the file spaces are should be checked periodically so that appropriate restructuring can be carried out when necessary, based upon these measurements the database designer determines when and how the physical database needs tuning to improve its performance. Viewing the database as a whole, the database designer determines what categories of security techniques should be used, and designs the detailed structures and techniques for maintaining privacy. The database designer designs any database searching strategies, or plans any use of inverted files or inverted lists. The rules relating to access constraints, including rules to prevent concurrent updates or interlocks, and rules to prevent excessive time-consuming search operations, are also defined by the database designer. Policies for deleting or dumping of old data, or data migration are defined by the database designer, as well as the data compaction techniques to be used. The database designer may determine that the data model needs splitting into separate disjoint structures, which may be in the same computer or distributed over several computers. If the data are to be distributed, the database designer participates in determining how this should be done and ensures that data distribution is done and controlled way that avoids deadlocks and integrity problems.

The data operations supervisor is a functionary who deals with problems that arise on a day-to-day basis [67]. The data operations supervisor investigates all errors that are found in the data, and marks all data that found to be in error. The data

2.3 MANAGEMENT

75

operations supervises restarts and recoveries after failure, and the reorganization of databases or indices. The data operations supervisor cleans up any data conflicts that are found, which usually happens when the contents of existing files are merged into the database. The data operations supervisor ensures that the volume library is controlled and maintained correctly, and supervises the transfer of files to alternative media when necessary. The data operations supervisor exerts some control over computer scheduling when necessary and initiates and controls all periodic dumps of data, audit trails, vital record procedures, and so on [67].

The security officer is a functionary who may not report to the database executives, the functions vary widely, depending on how seriously security is taken by an enterprise [67]. The security officer receives a list every morning of all violations of security procedures which could reflect attempts to compromise security, and investigates all known security breaches. The security officer determines who is authorized to use each locked facility or each locked data-item, record, area or file, and is responsible for the authorization tables which control locked facilities or records. The security officer is the only one who can make changes to the security authorization tables, and modifies the data locks and keys whenever necessary. The security officer ensures that the security procedures for the machine room and volume library are complied with, and performs periodic security audits to establish this [67].

2.3.4 Distribution of management

A large number of management tasks at the strategic, tactical and operational levels of data management are given in previous sections. These tasks are performed by personnel that are part of several organizational units. The personnel can be dispersed geographically and can have various levels of decisional powers with regards to performing their management tasks [45]. Distribution of data management is therefore divided into deconcentration and decentralization, this is discussed below. The separation of functions into distinct units for functional data management and technical data management is a specific kind a decentralization of data management. Functional and technical data management are discussed in the third section.

Deconcentration

Geographic placement of data management personnel and units is closely related to the degree of distribution of computer hardware and software [19], however due to improving technology, deconcentration of computer hardware and software is less an issue for decisions regarding geographical placement of data management personnel [45].

In a concentrated approach one main processing centre contains hardware, software and personnel. Advantages associated with a concentrated approach are economies of scale in computer resource use, division of labour, and standardization in equipment, products and common use of data. The disadvantages however, are

inflexibility and a low service level from the point of view of the users.

In a deconcentrated approach each functional organization unit has its own computer resources and its own data management personnel. Advantages associated with a deconcentrated approach are flexible implementation of specific requirements on data, higher service levels due to directly available data management personnel and more accurate costing and cost price calculation for data management services. The disadvantages however, are the inability of functional units to make joint use of data, loss of economy of scale in computer resource use and less career opportunities for data management personnel [45].

Increases in the deconcentration of data management means that decisional powers can no longer remain at a single location, leading to decentralization of the data management organization.

Decentralization

Decentralization entails the allocation of responsibilities and authorization to organizational units and departments for making decisions with regard to data management tasks that are performed by them. Large autonomous units (such as divisions) can have significant variations in management approaches and differ in the degree of responsibility and autonomy of divisional or local data management. The local data manager can control, within central data management imposed limits, cost, priorities in processing, security etc. of data. Services can be tailored to local needs because the local data manager is closer to the activities, and is in a position to exert greater management control [51].

The danger exists with the loss of central coordination, that the autonomous units will grow further apart in their definitions of data, thus leading to incompatible and uncomparable data. A central 'corporate' datamodel managed by a central data administration department might provide a solution. The central data administration department can define standard procedures to coordinate the definition and use of common data. It is not possible however, to define data unambiguously, furthermore the same concepts are likely to be used with different meanings by different organizational units. Decentralization of management of data definitions requires additional coordination between the central administration of definitions of common data, and local data management of data definitions specific for local units [81].

The position of data management within an organization is another point of interest, three likely positions for data management are: the finance and accounting department, a steering committee with representatives from (all) functional units, and a corporate executive officer [19]. A chief information officer (CIO) is a senior vice president with a staff position without direct responsibility for line activity. The CIO acts as an representative of the information systems management department and technology advisor to the executive committee [49], and the director of data management is commonly located directly below the CIO. This common approach to data management is shown in figure 2.6 [19]. When data management is centralized as a

2.3 MANAGEMENT

77

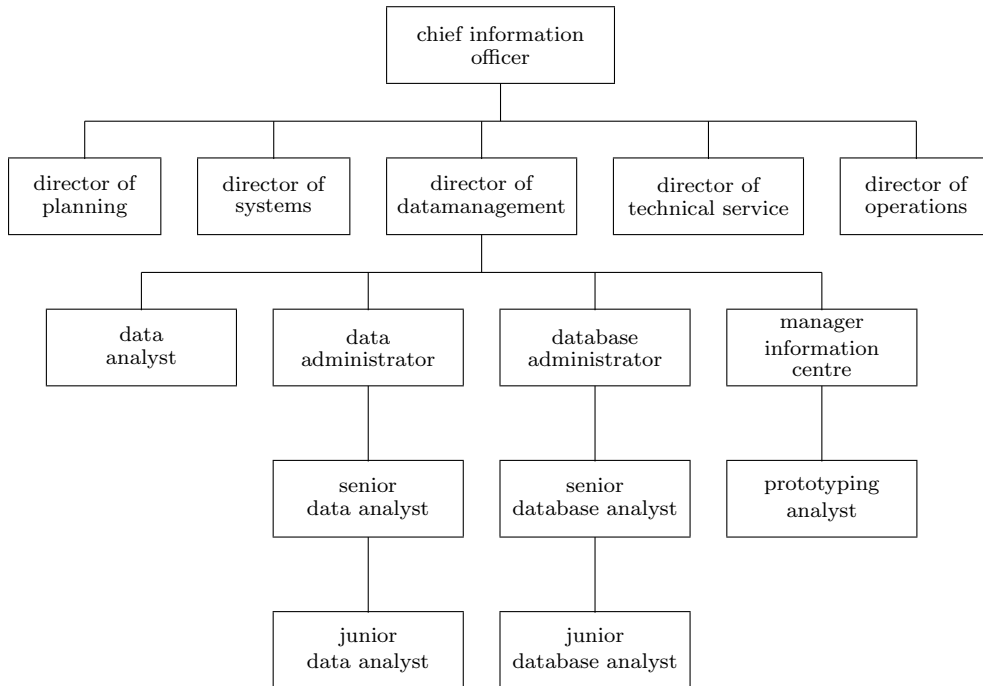


Figure 2.6: Typical organization chart of a data management department

staff bureau it is separated from the actual users of the data both geographically and functionally. This distance between users and staff-bureau enables line management to exert more power and the data management staff-bureau will be in a poor position. Common use of data will become virtually impossible to establish [61].

Decentralization of the environment was described in section 2.1.4 using the organizational theory of Mintzberg [73], giving five types of environment: the simple structure, the machine bureaucracy, the divisionalized form, the adhocracy and the professional bureaucracy. The placement of data depends on the part of the organization where the data is used, and this relationship is used by Van den Bogaart [6] to distinguish four types of ‘decentralization data’³.

In both the simple structure and the machine bureaucracy, data is directed towards the strategic apex of the organization where the decisional powers rest and which used data for issuing instructions to workers and monitoring their actions (apex data). In

³He distinguishes ‘ad hoc data’ for use by professionals, here replaced by ‘personal data’ to avoid confusion with the organizational structure adhocracy.

the divisionalized form, data is directed towards the various middle-line managers that use the data to exert control over their respective divisions (divisional data). In an adhocracy, data is tailored for the needs of the specific work constellations; e.g. financial data, marketing data and manufacturing data (constellation data). In the professional bureaucracy, data is used to satisfy the personal needs of the professionals and is not used by others (personal data). These four types of 'decentralization data' require placement of the data management department at different organizational positions.

Management of apex data requires a centralized data management organization, most likely at a main processing centre where all data management tasks are performed. Management of divisional data requires a parallel vertical decentralized data management organization in which each divisional data management unit is largely autonomous. As mentioned in the paragraphs above, the danger exists that autonomous units grow further apart in their definitions of data. A central data administration department may define standard procedures to coordinate the definitions and the use of common data. The divisional data management units may maintain local definitions and further perform all data management tasks in the division. Management of constellation data requires a selectively vertical and horizontal decentralized data management organization. Within each constellation all data management tasks are performed, just as in the case of apex data, but due to administrative integration (see section 2.1.2) the necessity of work constellations to exchange data requires the cooperation of the constellation data management units. Management of personal data requires a vertically and horizontally decentralized data management organization, where data management tasks are performed by the (professional) users at the operating core. The users determine the definitions of their own data because they do not have to be shared with someone else. A small support group, usually an information centre, remains to provide assistance and consultancy at the request of the users.

Functional management and Technical management

The separation of functions of data management into an organizations for functional data management and an organization for technical data management is a specific case of decentralization. In day-to-day data management, a conflict of interest might arise between providing the most effective support of data utilization and the most efficient exploitation of data automation facilities. One possible solution is the partitioning of data management after the nature of its tasks into [97]: 1) functional, user oriented tasks, and 2) technical, automation facilities oriented tasks.

Functional data management is a function directed at maintaining the functionalities of data. This includes support of utilization as well as evaluating utilization and modifying functionalities of data to meet new user requirements. The strong relationship with users requires functional data management to be positioned close to the users in the organization [60, 76].

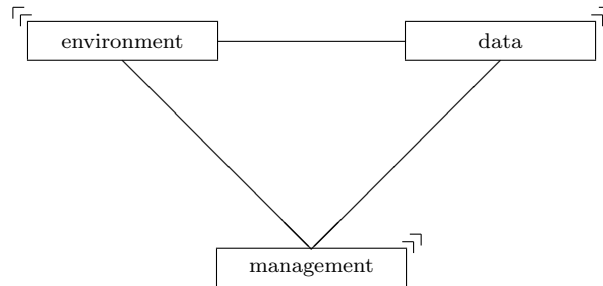


Figure 2.7: Triangular relationship between environment, data and management

Technical data management is a function directed at maintaining the technical operations of data. This includes support of exploitation of automation facilities as well as responding to deviations in agreed performance levels, and modifying technical realization of data. [60]. The strong relationship with the automation facilities requires technical data management to be positioned in the main processing centres [76].

2.4 Conclusions

The literature applicable to management of distributed data in distributed environments is discussed in this chapter. A discussion of the literature is required to obtain an understanding of current theoretical approaches and views to management of distributed data in distributed environments, and to identify the adequacy and shortcomings of these approaches. This understanding was vital for focussing the research objectives and provided a basis for the development of a framework and additional methods and techniques for managing distributed data in distributed environments.

The analysis of the theories and views described was based upon the triangular relationship between environment, data and management shown in figure 2.7. The angles shown at the corners denote the fact that each of the vertices can have some form of distribution. Distribution of the data environment can be geographical or organizational and is described in section 2.1.4. Distribution of data is described mainly in terms of geographical locations in section 2.2.4. Distribution of data management can be caused by the different geographical locations of organizational units that perform management tasks or by separation of functions and is described in section 2.3.4.

Conclusions regarding the data environment are given in section 2.4.1, conclusions regarding data are given in section 2.4.2, conclusions regarding data management are given in section 2.4.3, conclusions regarding present literature on management of distributed data in distributed environments are given in section 2.4.4.

2.4.1 Conclusions environment

The relationship between data and the real system differs for technical production processes and administrative process, this leads to different requirements and preconditions being made on data. A process oriented approach leads to requirements and preconditions on data that are optimized for a specific process. A data oriented approach leads to requirements and preconditions that are balanced for use by all processes, and not optimized towards a single process. Administrative integration is the cooperation between, and linking of, administrative activities in business functions based upon mutual interest. Administrative integration leads to common use of data and distribution of the environment of data. Environment deconcentration and decentralization means that matters of ownership are becoming less clear. The custodian of data should be clearly defined, if not, the introduction of common data can shift the existing balance of power within organizations and between organizations.

2.4.2 Conclusions data

Data infrastructures combine the process oriented approach and the data oriented approach, and enable both kinds of requirements to be satisfied. Requirements on data are made from a business management viewpoint and a technological viewpoint. The views of different authors differ greatly as do the requirements on data defined by them. The overview of requirements presented in section 2.2.3 does not pretend to be complete but serves to illustrate the large diversity and variation in opinions. These requirements should not be used in data management, rather the requirements defined in chapter 4 should be used. Introducing distribution of data is a complicated matter and is not always an obvious advantage. Data management arguments for distributed data should play a larger role than technological arguments in decisions whether or not to implement distributed data. Data fragmentation is defined in terms of the relational model using projection and restriction. Distribution of data is described primarily from a technological perspective, without relationship with the environment of the data or with data management.

2.4.3 Conclusions management

Data management is defined in the literature based primarily upon functionalities and presumes an established organizational structure. The strategic, tactical and operational levels of data management are in most instances not defined, and activities of data management are not restricted to the states Utilization (U) and Exploitation (E) of the state model for information systems management, but include the states information policy and planning (IPP), development (D), acceptance and implementation (AI) and maintenance (M). The many task clusters and functionalities encountered in section 2.3 are specified in conflicting terms and should not be used to implement data management, rather the management tasks defined in chapter 4 should be used.

2.4 CONCLUSIONS

81

When describing the distribution of data management, emphasis is placed on decentralization and the organizational position of the data management department. The relationship between the distribution of the environment, the distribution of data and distribution of data management is explored by Van den Bogaart [6] using the organizational theory of Mintzberg [73].

2.4.4 Current literature on management of distributed data in distributed environments

Current publications and theories show a fragmented image of management of distributed data in distributed environments. There appears to be a wide gap between the theories of the technical sciences and of the management sciences, and an overall view linking all the aspects of the triangular relationship as show in figure 2.7 has yet to be developed. Until now, research into data management has been highly experience oriented. The ‘expert opinion’ and ‘guru’ approach is found frequently in publications however, a scientific basis for the proposed data management approaches is often absent. A large amount of tasks are defined and these tasks are attributed to varying functions of data management. There is, further, little agreement about which tasks are performed at operational, tactical or strategic level of data management, and which tasks are involved in planning, development or data management. Current texts in managing distributed data in distributed environments are highly descriptive in suggesting what a management function should do and how it should be organized, but fail to provide underlying principles for data management.

2.4.5 Directions for this thesis

The conclusions that can be drawn from current literature emphasize the fragmented view of management of distributed data in distributed environments. Current literature does not provide a clear direction for the operational level of data management and does not provide an instrumentation for support of the operational level. This strengthens the aim of the research project to provide a management instrumentation that relates relevant aspects of managing distributed data in distributed environments and supports the operational level of data management. The instrumentation for management of distributed data in distributed environments consists of two parts: the FATO framework described in chapter 4, and the methods and techniques described in chapter 5.

Chapter 4 describes a framework of concepts designed to introduce relationships between the individual aspects of managing distributed data in distributed environments. It is called the ‘FATO framework’ after its four components: Functionalities, Automation facilities, Task fields and Organization.

Chapter 5 describes the methods and techniques that support an organization for management of data. These methods and techniques are positioned within the

82

ENVIRONMENT, DATA AND MANAGEMENT: LITERATURE

framework to define their application in managing distributed data in distributed environments.

Chapter 3

Environment, data and management: practice

The definition of management of information systems and the management paradigm, described in section 1.2, form the basis for the research programme on the management of information systems, and are used to define management of data in section 1.3. Within this management definition the key element is the utilization of data as an information system (IS) component. Utilization of data is placed within the environment of data, where data are required to perform systems functions. As a consequence, the definition of management of data presupposes a cohesive, triangular, relationship between environment of data, data, and management of data.

In chapter two it was shown that theory does not provide modelling techniques and methods to describe the structures of and relationships between environment of data, data and management of data. This makes additional insight into these relationships necessary before an instrumentation can be constructed that can be applied to the management of distributed data in distributed environments. Insight can be achieved by performing surveys on a number of situations in which distributed data are managed in a distributed environment. These surveys result in an understanding of the relationships described above and the impediments current practice is faced with. This understanding will subsequently be used as the basis for the development of a framework and additional methods and techniques for use in managing distributed data in distributed environments.

To obtain an overview of current practice of management of distributed data in distributed environments, two surveys were performed at the beginning of the research approach. These two surveys were carried out within the Central Organization (CO) of the Dutch Ministry of Defence and the Dutch Joint Office for Social Security Administration (GAK)¹.

¹In Dutch: Gemeenschappelijk Administratie Kantoor (GAK)

The survey at the Central Organization (CO) of the Ministry of Defence concerned the effect of a distributed environment on management of data. The survey showed the effect of local autonomy on definitions of data and how this complicates functional aspects of management of data. The Financial Administration (FINA) information system was used as an example of concentrated data stored at a single mainframe location, while being used and defined at many different locations.

The survey at GAK concerned the effect of distributed data on management of data. The survey showed the effects of physical distribution of data and distribution of utilization of data and how this complicates functional and technical aspects of management of data. The insured persons administration² (VZA) was used as an example of a situation where deconcentrated data is stored at many locations, and is also used at many locations.

This chapter is structured in much the same way as chapter two. First, the environments of data at CO and GAK are described. Secondly, the data present at CO and GAK are described. Thirdly, the management of data at CO and GAK is described. Finally, an evaluation of the current practice and impediments to management of distributed data in distributed environments is given in the conclusions. This chapter ends with an outline of directions for the thesis derived from the conclusions concerning current theories and practice of management of distributed data and distributed environments.

3.1 Environment

3.1.1 Ministry of Defence (CO)

The Dutch Ministry of Defence consists of the three joint services, the Royal Marechaussee and the Central Organization. The joint services are: Royal Netherlands Navy (RNLN), the Royal Netherlands Army (RNLA) and the Royal Netherlands Air Force (RNLAf). The Royal Marechaussee is responsible for national tasks which include border patrol, and protection of the Royal Family, and acts as the military police. The Central Organization (CO) is responsible for supporting tasks that cross the boundaries of the individual armed forces and concern the ministry as a whole, and for coordination between the (highly) autonomous forces. The political leadership of the ministry comes from the Minister of Defence and the State Secretary for Defence, who are both accountable to parliament. The Mintzberg diagram in figure 3.1 shows that the CO is placed at multiple positions within the ministry. The CO has some activities that concern standardization, activities that support the joint services and activities that coordinate and communicate between the joint services and the political leadership of the ministry.

The joint services and the CO are organized uniformly into four directorates-general concerned with: personnel, materiel, economic and financial affairs, and op-

²In Dutch: Verzekeren Administratie (VZA)

3.1 ENVIRONMENT

85

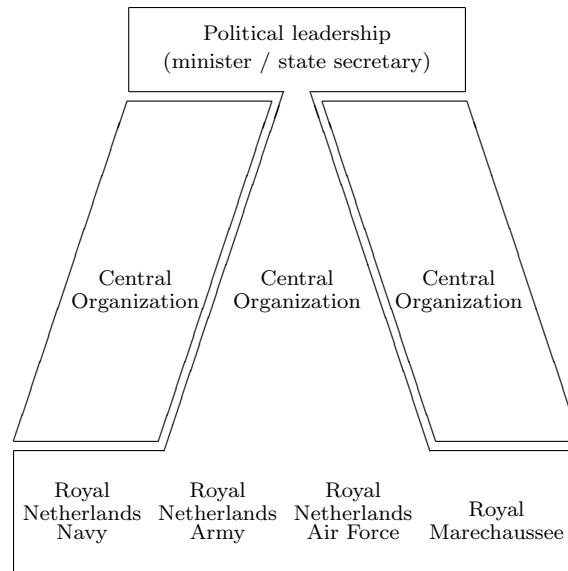


Figure 3.1: Organizational diagram of Ministry of Defence

erations. Operations within the CO is assigned to the Chief of Defence Staff who is responsible for coordination and communication between the joint services and the political leadership of the ministry. Within each directorate-general of the CO there are further directorates, of interest to the survey were the directorate for personnel (Medical Services) and the directorate materiel (Infrastructure) [112]. There are also a number of support directorates within the CO: Legal Services General Policy Affairs, Information, Internal Affairs and the Audit Board. Standardization within CO is achieved by the office of management consultancy and automation. The Central Organization is under the direction of the Secretary-General for Defence. The Mintzberg diagram in figure 3.2 shows where the organizational units of CO are positioned. The survey within CO was concerned with management of distributed data definitions. The financial administration information system (FINA) was chosen as an example of how management of data definitions is structured within CO [112].

Distribution of the environment

Strong local autonomy is a main characteristic of the ministry of Defence, especially at the level of the joint services and the central organization (CO), and also at the level of directorates-general and directorates within the CO. This means local autonomy of the joint-services, the CO and its directorates, to establish their own data

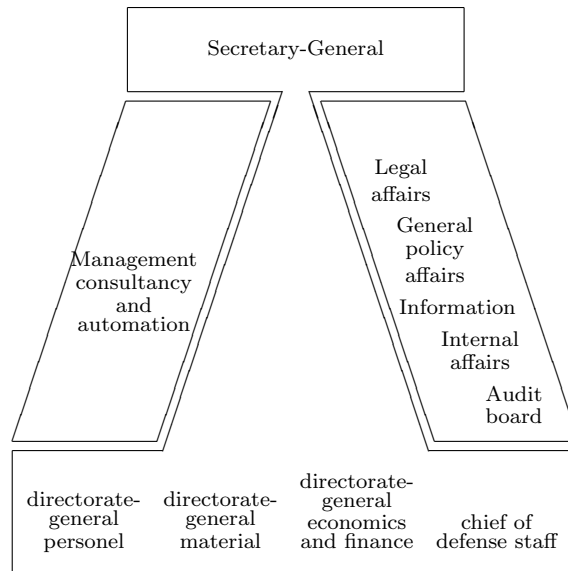


Figure 3.2: Organizational diagram of CO

definitions. Within the CO there are 9 different owners of the 29 information systems, and there are 16 different custodians of these 29 information systems. The 29 different information systems have been developed within autonomous projects, each information system for a particular organizational aspect of the CO. This has lead to little information at the corporate level that gives an overview of the ministry as a whole [112].

FINA

The financial administration (FINA) information system is the central information system of the central organization (CO). Most other information systems of the CO are connected with it, either automated or not automated. It consists of four major modules: the budget administration system (BAS), the accounts payable system (CAS), the accounts receivable system (DAS), and the monthly return system (MOS). The owner of the system is the directorate-general for economics and finance (DGEF), and the custodian of the system is the department for Financial Information, Analysis and Regulations. The FINA information system receives data from 26 organizations and information systems outside and inside the Ministry of Defence and sends data to 32 organizations and information systems, inside and outside the ministry. These organizations and information systems constituted the stakeholders that have a vested

3.1 ENVIRONMENT

87

interest in FINA, they in can be grouped into five categories:

1. Joint Services and the Central Organization:

- Partial budget-holder
- Budget-holder (about 600)
- Controller (about 15)
- Departments of economic management (about 10)

2. Financial management organizations

- Budget managers (about 65) and verifiers (about 150)
- Creditors agenda (about 25)
- Staff bureau internal audit
- Staff bureau financial management

3. Department of planning and control

- Bureau for policy affairs
- Bureau for budgetary affairs

4. Bureau financial information systems

- Section systems management
- Section information processing and provision; towards external organizations (Treasury, etc.)
- Section financial accounts; management of the bank-accounts of the ministry

5. Inspecting and consulting bodies:

- Audit Board
- Department for planing and control
- Public accounts committee
- Central accountants service of the Treasury

Plus a further twenty debt administrators and twenty auditors (inspectors).

	GAK 1953 [37]	GAK 1991 [38]
number of people working for GAK	6386	17.489
working in district offices	—	13.727
membership ^a	28	95
affiliated employers	100.000	281.410
number of insured people	1.250.000	3.498.000
total amount granted benefits	Dfl. 250.000.000	Dfl. 24.288.670.000
premium revenue	Dfl. 471.000.000	Dfl. 25.216.530.000
personnel costs GAK	Dfl. 15.384.000	Dfl. 1.145.531.000
other costs GAK	Dfl. 7.428.000	Dfl. 386.235.000

^asee below for a description

Table 3.1: Overview of GAK

3.1.2 GAK

The Netherlands has a number of social security laws that regulate a (large) number of social security insurances. These laws are divided into general laws applicable to anyone (national insurances) and laws that are only applicable to those who are in paid employment (employee insurances). The general occupational disability act (AAW) is a national insurance for anyone who is unfit for work. The health insurance act (ZW), the workers occupational disability act (WAO) and the unemployment act (WW), are examples of employee insurances. Implementation of these social insurances is entrusted to 19 occupational associations (BVs), each servicing a different sector of industry. The occupational associations are governed by a board consisting of both employers organizations and employees organizations, referred to as the social partners.

The Joint Office for Social Security Administration³ (GAK) is an association founded in 1952 by employers organizations and employees organizations (trade unions) to execute social insurance legislature for the benefit of the occupational associations. The advantages of economy of scale allow GAK to administer social insurances at a lower cost. The board of GAK consists of representatives of the employers organizations and the employees organizations: the social partners. Thirteen occupational organizations are members of GAK and have their administration carried out by GAK. Other members of GAK who have their administration done by GAK are: 23 pension funds, the joint medical services (GMD), 38 early retirement organizations (VUT) and the federation of occupational associations (FBV).

In the last forty years the amount of benefits paid has risen from 40 million guilders to over 24 thousand million guilders. With over 30 thousand million guilders in investments GAK was ranked in the top 100 investment companies of 1991. The growth of GAK over the past forty years is shown in table 3.1.

The organization of GAK is divided into three directorates. One directorate col-

³In Dutch: Gemeenschappelijk Administratie Kantoor (GAK)

3.1 ENVIRONMENT

89

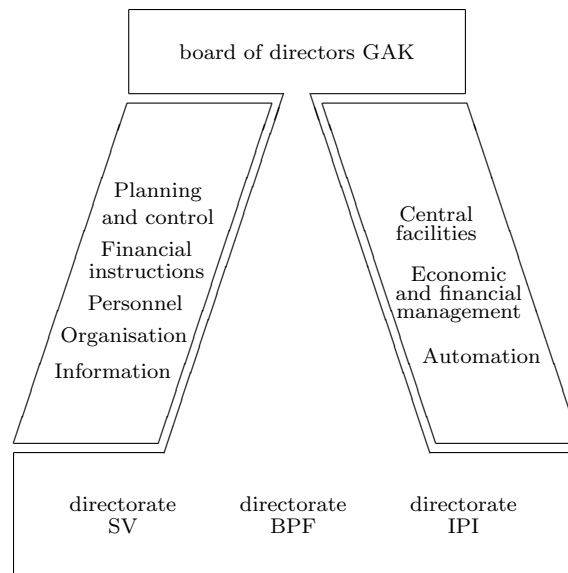


Figure 3.3: Organizational diagram of GAK

lects pension premiums and social insurance premiums (IPI). The second directorate performs administrations on behalf of pension funds (BPF). The third directorate performs administrations on behalf of social insurances (SV), this involves determining entitlements to benefits and amounts of benefit to be paid, and issuing payment orders. Each directorate is accountable for its financial results to the central board but is further autonomous. Special departments exist for personnel, finance, information and automation. Other activities are performed by central units to improve efficiency. Administration of social security insurances within the directorate SV is highly distributed, there are 30 district offices and 50 regional offices. In total GAK has 350 larger and smaller offices throughout The Netherlands.

Social Security directorate (SV)

The social security directorate executes the social insurance legislation commissioned by the occupational associations (BVs) and joint medical services (GMD). There are three units within the directorate. One unit is the Social Medical Service (SMD), it covers the medical aspects of the social insurances, especially health insurance (ZW) and occupational disability (AAW/WAO). SMD is responsible for medical examinations and activities for the prevention of sickness absenteeism and occupational disability. The medical tasks are carried out at the 30 district offices of GAK, while

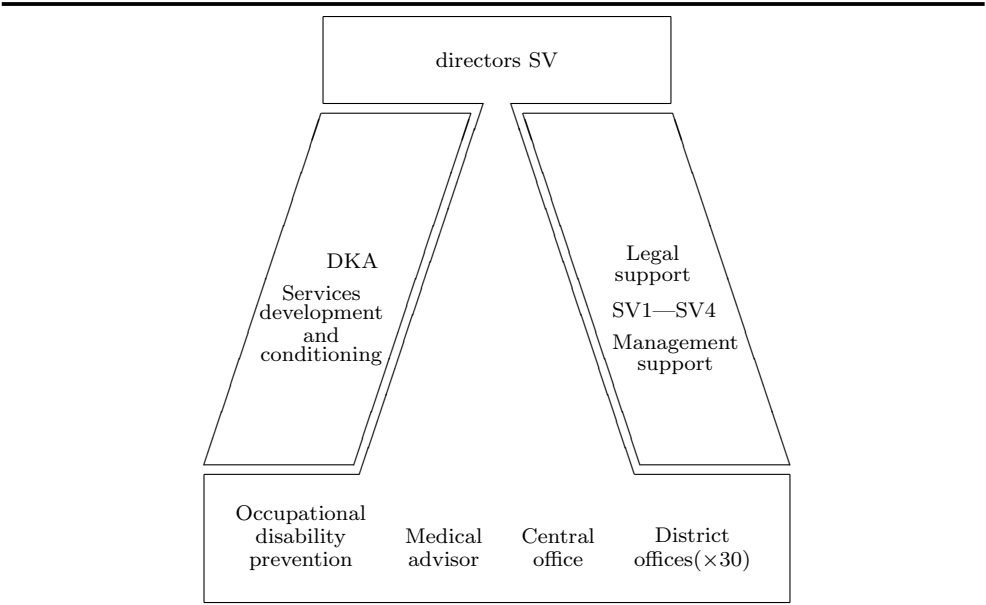


Figure 3.4: Organizational diagram of the Social Insurance directorate

medical examinations are performed by doctors (GPs) at local offices provided by GAK.

The second unit, Enforcement, actually performs the administrative tasks on behalf of the social insurances. Its activities are mainly based at the 30 district offices, and include determination of entitlement to benefits and issuing payment orders for benefits. Some tasks, such as interpreters and recovery of payments, are centralized for efficiency.

The third unit, Management Contacts, maintains contacts with the occupational associations, the legislature, the social security council and others. Based upon these contacts it modifies administrative tasks to meet new requirements. The information systems used by the directorate to implement social insurance legislation are developed by the department “district office automation” (DKA). The department DKA also manages three regional computing centres (RCCs) where the information systems are run. The social medical service (SMD) has its own bureau that develops information systems on behalf of the SMD.

Distribution of the environment

The environment of data managed at GAK has a wide geographic and organizational distribution. The stakeholders are both inside and outside the GAK organization.

3.2 DATA

91

Inside GAK there are thirty different district offices of the social insurance (SV) directorate, and the head office containing the pension funds (BPF) directorate and the premium collecting (IPI) directorate. The head office itself is housed at three locations in Amsterdam. The wide distribution of the environment however is caused by the external relations of GAK. GAK is involved in intensive exchange of data, concerning employers and employees, between its 95 members and 280.000 affiliated employers. Personal data is verified with the Ministry of Finance and close to 750 municipalities, addresses are verified with the postal service. People living abroad but who have social insurance or pension rights in the Netherlands also have to be checked.

3.2 Data

3.2.1 Ministry of Defence (CO)

The survey at the central organization of the Ministry of Defence (CO) was concerned primarily with data definitions [112]. Therefore this section only concerns the distribution of data definitions. There are about 12,000 data definitions that have to be managed by CO [112]. All these definitions have to be kept separate from each other, if this not done there is a danger of synonyms and homonyms and wrong interpretation of the data. The definitions are difficult to access, as they are “on the shelf” with the system developer, and although the data definitions are well documented, fast retrieval is not possible.

The data definitions are well documented due to the amount of attention and standardization dedicated to this matter in the ‘Handbook Project realization and Systems management’ (HPS). HPS is the standard for development of information systems at the Ministry of Defence, and is closely related to ‘System Development Method’ (SDM). The documentation of data definitions is kept up to date, however this is done manually. This reduces the accessibility of the data definitions. The quality of data definitions within CO varies. Data definitions range from a minimal description for one information system to an extensive description for another information system. This range in quality can be attributed to the project approach to creating definitions of data. One project will pay more attention to it than another.

Distribution of data

One way to achieve integration of data definitions but to still maintain local autonomy is to create separate contexts of interpretation of the data [81], and manage the definitions in each context separately. This means the definition of data is ‘context dependent’ and that the same piece of data is allowed to have different meanings in different contexts. Homonyms and synonyms of data are allowed between contexts, but not within contexts. Connecting the contexts requires building interfaces between these different definitions. Within the Central Organization twenty different contexts

were distinguished [112], however no context at the corporate level could be defined. This means in effect, that corporate data does not exist within the Central Organization. Data that is of interest in several contexts and several information systems, is not stored at a separate system, and does not have a separate definition. Such data is stored in all the information systems within the context where it is used, and is therefore defined implicitly at different locations without definitions for the context as a whole. This is a direct consequence of the project approach followed when developing information systems according to the HPS standard. Little attention is paid to coordinating data definitions within other contexts, and integrating the information systems with interfaces is considered problematic by CO.

FINA

The data component of FINA is concentrated at the Duyverman Computer Centre (DCC) total data storage available is about 8500 MB. The large IBM 3090 mainframe at DCC performs about 3,5 million on-line transactions per month on behalf of FINA. There are two large IDMS databases giving 400 MB in total. There are about 350 reports, 170 intermediate files, 10 input-files, 40 output-files and 5 'history' files of 4000 MB of data in total. Each database has 200 subschema's, 175 schema-records and there are about 1400 other kinds of records. These numbers show FINA to be a large and complex system, together with its large and diverse environment it requires a well defined data management organization. The management organization of FINA is addressed in section 3.3.1. The automation facilities involved in FINA are shown in figure 3.5.

3.2.2 GAK

In 1953 administration was mechanised using punch card and punch tape machines. In 1961 GAK was one of the first Dutch organizations, and the first social security administrator, to buy a computer: the Univac Solid State. This was a punch card computer that soon became too slow for the amount of operations it was required to perform, GAK therefore acquired a third generation magnetic tape computer in 1966: the ICL 1904. Spending by GAK on automation at present amounts to about thirty million guilders in leasing and maintenance of automation and another 29 million in purchasing automation facilities [38]. Starting from a single Univac, the IT-infrastructure of GAK has grown to the following conglomerate:

- 6 ICL mainframes (VME) concentrated at the Amsterdam head office
- 15 ICL mini-computers (UNIX) concentrated at the Amsterdam head office
- 70 DEC VAX computers highly distributed throughout the GAK organization
- 15 DEC micro-VAX computers distributed throughout the GAK organization

3.2 DATA

93

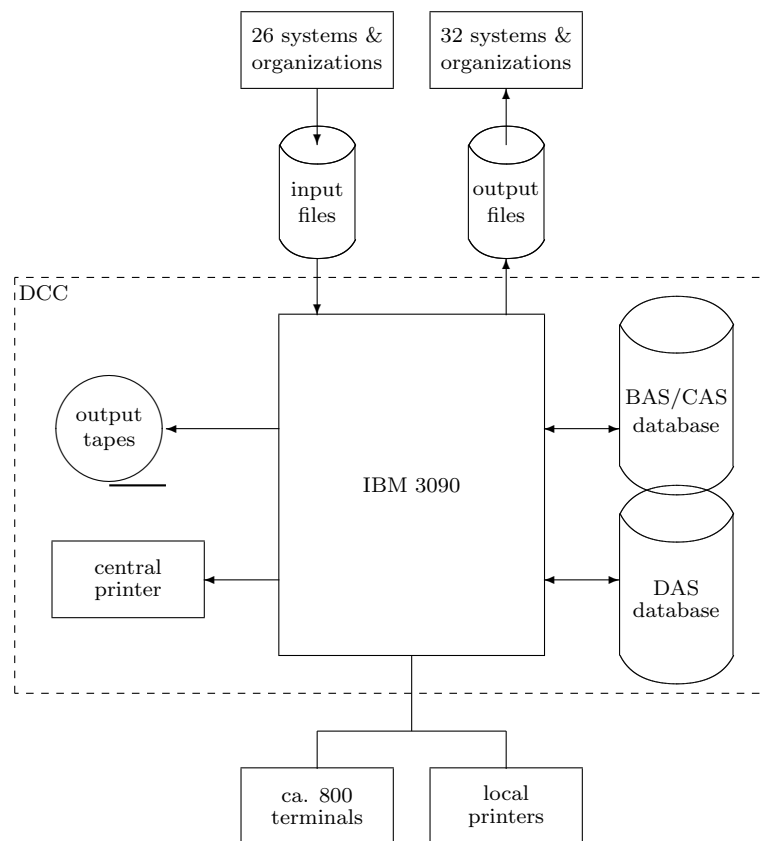


Figure 3.5: The implementation of FINA

- 3000 personal computers distributed throughout the GAK organization
- 12000 terminals distributed throughout the GAK organization
- 2300 printers distributed throughout the GAK organization

GAK's IT-infrastructure for processing and storage of data is distributed over five locations:

- two computer centres (CCCs) in Amsterdam. One computer centre contains ICL equipment, the other contains DEC equipment.
- three regional computer centres (RCCs) at Hengelo, Den Haag and Eindhoven which contain DEC equipment.

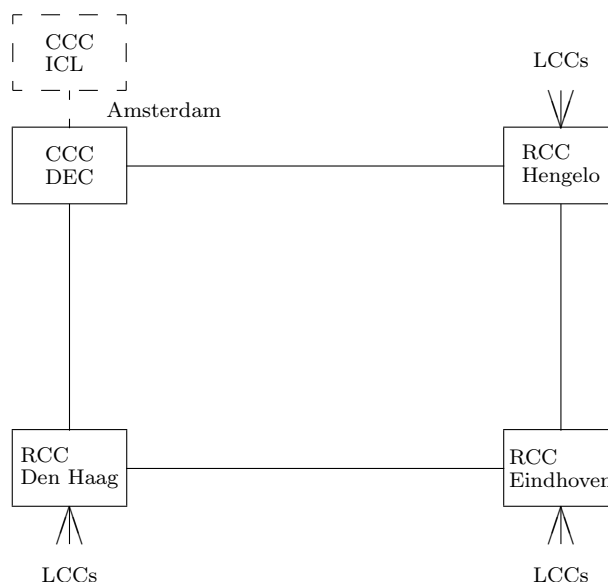


Figure 3.6: Technical-infrastructure of the Social Insurance directorate

- thirty local computer centres (LCCs) at thirty district offices which contain DEC equipment.

The CCCs are connected by a 100Mb/s optical network, the RCCs are connected to the CCCs by a 2Mb/s token ring network, and the LCCs are connected to the RCCs by 2Mb/s links. The IT-infrastructure of GAK is divided into two parts; one consists of DEC equipment used by the SV directorate, the second consist of ICL equipment used by the directorates IPI and BPF.

The insured persons administration (VZA)

In 1985 the Dutch state-secretary for social security gave orders to the social insurances council to instruct the occupational associations to build an insured persons administration⁴ (VZA) and to regulate its construction. The purpose was to reduce the amount of fraud in social security. During this period GAK implemented a major change in the way GAK worked shifting from case oriented administration to person oriented administration. This means that someone is now registered as an insured person the moment they enter into a contract with an employer, where before they would be registered each time they applied for an insurance benefit.

⁴In Dutch: Verzekerden Administratie (VZA)

3.2 DATA

95

GAK opted for a layered information architecture in which three subject registers and several process administrations were linked with each other. This architecture was slightly modified to implement the insured persons administration within GAK. The three subject registers are:

- The personal data register⁵ (BRP), contains data of all insured persons with whom GAK maintains administrative relations on behalf of the occupational associations and other members of GAK. By 1992 over 6.2 million people were registered by GAK.
- The employers register⁶ (BRWG), contains data of all employers with whom GAK maintains or has maintained administrative relations on behalf of the occupational associations and other members of GAK.
- The labour contracts register⁷ (BRDV), contains data of all current and historic labour contracts between insured persons and employers that are, and were, affiliated with one of the occupational associations that are a member of GAK.

The process administrations (PA) contain data that are specific for a particular social security legislation, e.g. unemployment or health insurance. The legislative data can be redundant with respect to other process administrations or subject registers. The subject registers together form a data infrastructure upon which the process administrations, and other information systems in use within GAK can be connected. The connections between process administrations (PAs) and subject registers (BRs) is established by a persons Social-Fiscal (SoFi) number. This number is used as an access key in every database, so retrieving data from the databases is relatively easy. The process administrations are connected with the data infrastructure through a mechanism that is known as the 'Path'. This mechanism is explained below.

Distribution of data

The distribution of data within GAK is shown in figure 3.8. The personal data register (BRP) and employers register (BRWG) are replicated over three regional computer centres (RCCs) and one central computer centre (CCC). The labour contracts register (BRDV) is not replicated and is placed at the CCC. The BRP has two central copies, the ICCP and COPY-BRP, which are used for parallel batch processing during the evening. The 'integral regional copies' (IRCPs) are placed on a separate computer at the regional computer centre. This is to prevent performance deterioration in one application from causing other applications to be unable to use the regional copies.

The process administrations (PAs) are split into thirty fragments, each fragment contains data needed by a particular LCC. An insured person is administered by one district office, and so it is possible to fragment a process administration based upon the

⁵In Dutch: Basisregistratie Personen (BRP)

⁶In Dutch: Basisregistratie Werkgevers (BRWG)

⁷In Dutch: Basisregistratie Dienstverbanden (BRDV)

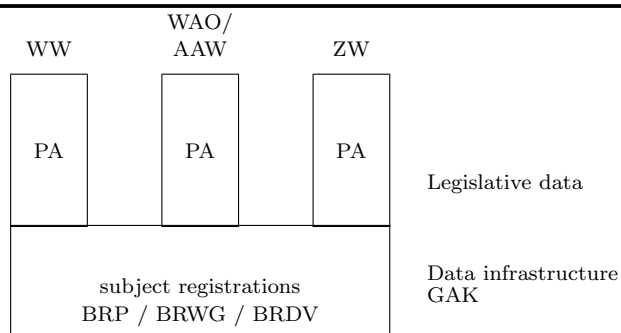


Figure 3.7: Data architecture of Social Insurances directorate

district offices where the LCCs are located. The fragments of process administrations are stored at the RCCs, and as ten LCCs are connected to one RCC, ten fragments are stored by the RCC. Every fragment is a different database managed independently from the other fragments.

The process administrations (PAs) were built using a client-server architecture application. The LCCs, located at the district offices of GAK, have front-end computers which perform the menu and screen handling part of an application. The RCCs have back-end computers where the actual computer applications are run and data is accessed. The back-end of a process administration is connected to subject registers using a special 'Path' mechanism. The process administration uses the 'Path'-mechanism to access data from the subject registers without having to know where the data is actually stored. The 'Path' is also a client/server application that is split into a front-end and a back-end. The front-end at the RCC accesses the regional copies placed there and the back-end of the 'Path' accesses the central registers. The 'Path' consist of three sub applications: the ADB application accesses the BRP, the ADW application accesses the BRWG, and the ADD application accesses the BRDV. These sub applications are used to determine if the data is to be read from the regional copy or from data stored centrally. The 'Path' is used to maintain a mutation indicator for every record of data in the regional copy. When the mutation indicator shows that the value of a particular data has changed during the day, the data isn't read from the regional copy but from the central register. Mutations are always performed on data stored centrally, and these mutations are distributed to the regional copies at night. When the regional copies are re-synchronised with the central data, the mutation-indicators are reset and data can once again be read from the regional copy.

3.2 DATA

97

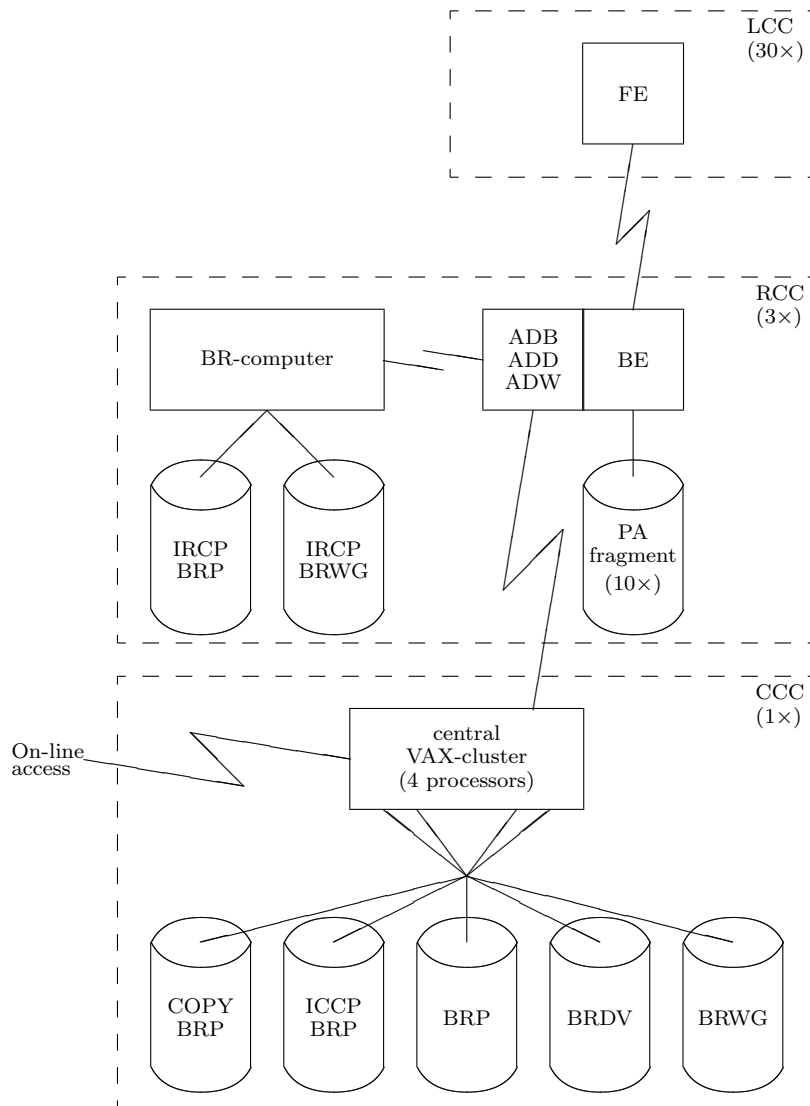


Figure 3.8: Implement of the data architecture

Avg. number of transactions each day		
	actual	SLA-BRP
on-line	22.500	4.800
via ADB	500.000	190.200
via ADW	140.000	–
via ADD	20.000	–
Mutations BRP each month		
	actual	SLA-BRP
Batch	2.444.749	600.000
via ADB	367.989	265.000
on-line	310.316	17.800

Table 3.2: Use factor subject register BRP

Implementation subject registers

The deconcentration of automation facilities within GAK is in sharp contrast to the single mainframe for the FINA information systems at the CO. Within the central computer centre there is a VAX-cluster that consists of four processors used for subject registers. Thirty six magnetic disks with a total of 24 Gigabytes of storage for the BRP are connected to the central VAX-cluster. The most important disks of the VAX-cluster are doubled as ‘shadow-disks’. This means that all I/O operations are copied to both disks keeping the disks identical. If one disk fails, the second will continue to operate. The three RCCs in total have 21 magnetic disks with a total of 17 Gigabytes of storage for the BRP, these are not shadowed. The BRP has 57 magnetic disks in total with 41 Gigabytes storage capacity in total at the three RCCs and the CCC. All the subject registers together have 67.1 Gigabytes of disk storage available at the RCCs and the CCC. This deconcentration of automation facilities is shown in figure 3.8

To give an impression of the use factor of the subject registers, some numbers together with the service levels agreed upon in the service level agreement (SLA) are shown in table 3.2. These numbers are an indication that the actual performance of the BRP is much higher than is agreed upon.

3.3 Management

3.3.1 Ministry of Defence (CO)

The ‘Handbook Project realization and Systems management’ (HPS) pays strong attention to the management of developed and operational information systems. In the case of management of data, the HPS distinguishes activities that manage the following objects:

3.3 MANAGEMENT

99

- Meta data and data structures (the object model, the conceptual model and the implementation model)
- Actual data (contents of files and databases)
- Physical storage of files and databases (disks, tapes etc.)
- Security of stored data and providing recovery or reconstruction when needed
- Authorization for using the data
- Supervision of the correct application of the security procedures

Distribution of management

There are 19 units within CO involved with functional aspects of management of data. There is a strong one-on-one relationship between information system and management organization, for every information system a separate unit is involved in management of data definitions. This is a consequence of the project approach to developing information systems prescribed by the HPS standard. As part of a development project, a separate management organization was installed. A more integrated approach to data definition management is found at the Directorate of Infrastructure and Environment and at the Directorate of Military Medical Policy. These problems have long been acknowledged within CO, and some coordination mechanisms have been created at the level of the Ministry of Defence and of the Central Organization.

Coordination of management of data definitions for the Ministry of Defence is achieved by standards issued from the directorate organization and information (DOI). Mutual adjustment is achieved by the Sub Committee Organization and Information (SuCo O&I), a second mutual adjustment mechanism is the Steering Committee Of Military Medical Information Systems.

Coordination of management of data definitions for the CO is achieved by the standards issued from the Office of Management Consultancy and Automation. Mutual adjustment is achieved by meetings of the Coordinators Organization and Information (COI-meetings). Management of FINA is formally split between technical management at the 'Duyverman Computer Centre' (DCC) and functional management at the directorate general for economics and finance (DGEF).

3.3.2 GAK

In contrast to the Ministry of Defence, management of data within GAK is distributed over three organizational units: district office automation (DKA) within the directorate SV, the unit information management GAK (IM-GAK) and the automation department (AUT). In the following paragraphs several sections within these units that are involved in managing data are described. Together they form the data

management organization of GAK that is involved with management of the insured persons administration (VZA).

Data management sections

The subject registers department (BRs). This is a sub unit of the Organization & Information unit. The department BRs maintains the data contents of the subjects registers on behalf the central administrative processes of GAK. Central register of changes of address and corrections of the contents of the data are performed by this group and are checked against authentic sources such as the Local Government Population Registration Network⁸ (GBA) or the Exchequer.

The district offices. The decentralized administrative processes at the district offices of GAK also performs corrections with regard to the contents of the data. A special functionary at each district office is authorized to modify directly the central subject registers.

The central employers register department (CWR). This is a unit within the directorate IPI which collects pension fund and social insurance premiums from employers. The department CWR maintains the data contents of the employers register (BRWG) on behalf the directorate IPI.

The internal auditing department (IAD). The IAD performs audits on behalf of the board of directors on the way data management is implemented within GAK. It sets quality requirements for the data that are aimed at the reliability of accounts rendered by the various business units.

The Organization department. The Organization department defines a framework and directives at a general level for performing data management tasks. It also maintains contacts with standardisation organizations outside GAK and develops EDI messages that are used in electronic exchange of data between GAK and (the major) employers. The group 'GAK Wide Coordination' within the Organization department acts as a mediator between the different parts of the GAK organization involved in data management. The groups also participates in talks between all organizations in the Netherlands that are involved in the execution of social security legislation. Based on these talks it sets up GAK wide data definitions.

The group EXSO. The group external contacts and special assignments (EXSO) functions as the secretariat of the Committee Coordination Administrative Automation (CCAA). This committee is a cooperation between several social security organizations to standardise social security data so that data interchange among administrations of the same security insurance is possible.

⁸In Dutch: Gemeentelijke Bevolkingsadministratie (GBA)

3.4 CONCLUSIONS

101

The department operational information systems (OIS). This department consist of several sub departments a few of which are involved in management of data for the process administrations (PAs) as well as the subject registers of GAK (see figure 3.7).

- sector support (OIS-O), within this sector management of data is implemented by means of ‘reverse engineering’ of existing data definitions.
- sector district office (DK) exploitation, this sector performs application management of process administrations such as unemployment insurance and health insurance.
- sector subject register (BR) exploitation, this sector performs application management of the subject registers. It gives directions to the technical management of both central and regional subject registers.
- the group service management, this group supports the three sectors mention above in defining and monitoring service level agreements (SLAs). There are two such agreements. One is between the department BRs as custodian and the department OIS as maintainer, for putting the personal subject register (BRP) at the disposal and keeping it at the disposal of the department BRs. The second agreement is between OIS as maintainer of the users and ATI as supplier of the automation facilities needed by the users of the BRP.

The department systems management and production control. This department is a sub unit of the department DKA, and manages the computer systems at the LCCs and the RCCs, and it manages the regional copies of the subject register and the fragments of process administrations.

The department for management of the IT-infrastructure (ATI). This is responsible for the communication network that connects the central computer centres (CCCs) with the regional computer centres (RCCs). ATI also manages computer systems at the CCCs.

The group Quality and Development (K&O). This is a support group within the development department of AUT that develops standards for new data definitions made during the development of new information systems.

3.4 Conclusions

This chapter examines two practical situations of management of distributed data in distributed environments. It does so to obtain an understanding of current practice, the relationships between the environment of data, data and management of data and the impediments that management of data is faced with. This understanding is used to provide directions for the research described in this thesis, focussing on the

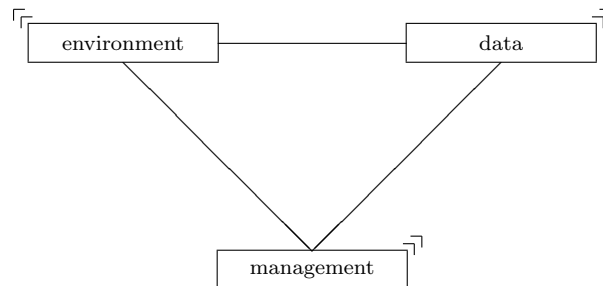


Figure 3.9: Triangular relationship between environment, data and management

objectives and to provide a basis for the development of a framework and additional methods and techniques for managing distributed data in distributed environments.

The analysis of the results of the surveys at the Ministry of Defence and GAK was based upon the triangular relationship between environment, data and management as shown in figure 3.9. The angles at the corners denote the fact that each of the vertices can have some form of distribution. The environment of data consists of the stakeholders that have a vested interest in the functionalities performed by data, and, as was shown in sections 3.1.1 and 3.1.2, distribution of the stakeholders will cause distribution of the environment of data. Distribution of data is both functional and technical, as shown in sections 3.2.1 and 3.2.2. Distribution of management of data is caused by the different locations of organizational units that perform management tasks, as shown in sections 3.3.1 and 3.3.2.

3.4.1 Conclusions Ministry of Defence (CO)

In the survey carried out at Central Organization (CO) of the Ministry of Defence it was concluded [112] that CO has a clear understanding of the effects of distribution of the environment on the management organization. Relationships between the distribution of the environment and the different definitions of the same data are difficult to establish in CO, and it was found that it was not possible to define in what way management of data maintains these relationships. It was also found that it is difficult to assign management of the contexts of data to organizational units. Allocation of authorisation, responsibilities, work and expertise of tasks to organizational units and establishing coordinating among these organizational units is another problem. The 29 different information systems have been developed within autonomous projects in accordance with the HPS standard, each information system is designed for a particular organizational aspect of the CO. This has lead to little information at a corporate level that gives a view of the ministry as a whole [112]. Local autonomy within CO, together with the large number of information systems

3.4 CONCLUSIONS

103

(29) and data definitions (12,000) leads to a number of bottle necks for management of data:

- There is no integrated overview of all the data that is in use within the CO
- As a consequence it is not possible to obtain high level management information quickly. How reliable the data is and where to obtain it is not known.
- There is no umbrella organization for management of data definition within the CO, and management organizations determine their own definitions. The lack of a central CO data dictionary makes exchange of data between information systems unfeasible.
- Some data is stored redundantly. The costs of data storage are falling, but increased storage leads to increased demands on management of this storage. Redundancy will lead to inconsistency to between data stored redundantly and will further increase the demands made on management of data storage.

Solutions to these bottle necks need to address the local autonomy of the stakeholders within CO, but should still enable integration of data definitions so data exchange between information systems is possible.

The introduction of contexts, allows different data dictionaries with data definitions to be used along side each other [81]. This also allows for flexibility with external definitions, e.g. the personnel information systems of the joint services, by creating special contexts in these cases. The larger contexts will have a 'local' data definition manager, the smaller contexts can be managed by the CO 'corporate' data definition manager. The joint services will each have a corporate data definition manager. There will be seven local data definition managers at CO, one for each of the four directorates-general, one specifically for the directorate of military medical policy, one specifically at the directorate of infrastructure and environment, and one at the strategic level of CO. An enterprise administrator is placed above the corporate data definition managers of the joint services and CO. The enterprise administrator will coordinate the corporate data definition managers and issue standards concerning management of data definitions.

3.4.2 Conclusions GAK

In the report on the survey at GAK it was concluded [70] that GAK has a clear understanding of the effects of distribution of the data on the data management organization. GAK, however, is unable to relate the distribution of the environment to the data management organization and the distributed data. GAK has difficulties in establishing what requirements and preconditions of data are made by the environment of data and what management tasks are performed on behalf these requirements and preconditions. It can be seen from table 3.2 that the actual performance of the BRP is much higher than agreed upon. Management of the BRP should take action

to agree the required and delivered service levels of the BRP. There is almost no coordination of management of data through standardization by a 'technostructure' department within GAK. The most important coordination mechanism in data management within GAK is mutual adjustment. Coordination through direct supervision is not possible due to the fact that the management organization is split over three autonomous organizational units.

3.4.3 Current practice of management of distributed data of distributed environments

The conclusions and recommendations made in both surveys [70, 112] suggest that the triangular relationship between environment of data, data, and management of data, is not present at GAK or the central organization (CO). There has been no equal trend in distribution of data in distributed environments and its management within GAK and CO. Management has not kept up with the distribution, and only now is it beginning to appreciate the advantages and disadvantages of distribution; for this reason GAK and CO are looking for instruments which can improve on this historical, uncontrolled, situation to be better prepared for the future.

These conclusions are drawn from two of the many organizations where management of distributed data in distributed environments can be found. The results of these two surveys and their conclusions do not rely on any specific attribute of these organizations, and it is believed that they give a fair description of the current state of management of distributed data in distributed environments within The Netherlands.

3.4.4 Directions for this thesis

The conclusions that can be drawn from current practice emphasize the lack of insight and coherent view of management of distributed data in distributed environments. Current practice does not have clear guidelines for data management and does not use clear and predetermined methods and techniques at the operational level. This strengthens the aim of the research project to provide a management instrumentation that relates individual aspects of managing distributed data in distributed environments and supports the operational level of data management. The management instrumentation for management of distributed data in distributed environments consists of two parts: the FATO framework described in chapter 4, and the methods and techniques described in chapter 5.

A framework of concepts designed to introduce relationships between the individual aspects of managing distributed data in distributed environments is described in chapter 4. It is called the 'FATO framework' after its four components: Functionalities, Automation facilities, Task fields and Organization.

The methods and techniques that support an organization for management of data are described in chapter 5. These methods and techniques are positioned within

3.4 CONCLUSIONS

105

the framework to define their application in managing distributed data in distributed environments.

Chapter 4

The FATO framework

This chapter is devoted to the framework for the management of distributed data in distributed environments. The object of this framework is to define what management of distributed data in distributed environments is by placing all elements in relation to each other. In chapters 2 and 3 it was shown that many definitions and views exist concerning environments, data and management. The FATO framework was devised to give a practical sub-division of all the aspects of management of distributed data in distributed environments.

The FATO framework is named after the four components of the framework: functionalities (F), automation facilities (A), task fields (T) and organization (O). The concept of FATO was introduced by Looijen [58] to model the management and organization of the automation facilities of information systems. In this thesis the concept of FATO is taken one step further, and now acts as a framework to position all the aspects of management of distributed data in distributed environments. Functionalities of data are determined by the underlying corporate processes and are the starting-point of management of data. Automation facilities are the technical means used to implement the data functionalities. Task fields are groups of tasks that have to be performed for managing the functionalities and automation facilities. An organization that is adequately staffed and structured to perform the tasks completes the FATO framework. The framework is further sub-divided into the following elements:

- four levels of data: pragmatics, semantics, syntactics and empirics [96]
- four task fields: actual data management, data administration, database administration and storage management
- two organizational units: functional management and technical management

Table 4.1 gives an overview of the framework and shows the relationships between each of the elements of the framework.

F	A	T	O
pragmatics		actual data management	Functional management
semantics		data administration	
	syntactics	database administration	Technical management
	empirics	storage management	

Table 4.1: The FATO framework

Pragmatics	Semantics	Syntactics	Empirics
correctness	convertibility	access speed	reliability
completeness	relatability	restructuring speed	repairability
actuality	domain integrity	data capacity	recoverability
timeliness	static integrity	load factor	access time
durability	insert integrity	use factor	transfer rate
confidentiality	modification integrity	transformation	storage capacity
auditability	securability	accuracy	safety
		concurrency	retention time
		error detection	storage costs
		error correction	access pattern
			fill factor

Table 4.2: Characteristics of data

Management of distributed data in distributed environments works in accordance with user requirements and preconditions and characteristics of data (see definition on page 25). Section 2.2 describes the many characteristics of data that literature defines from a number of viewpoints: e.g. engineering, EDP-audit, IS-development and DBMS-programming. Table 4.2 shows the characteristics that are defined, within this thesis, to be important for management of data. The characteristics are positioned within the FATO framework by assigning them to one of the four data levels.

It may be possible that not all the characteristics from this list are important to the management of data. The first research question (page 31) therefore asks what characteristics of data are important for performing tasks in management of distributed data in distributed environments. The three case studies described in chapters 6, 7 and 8 help to determine what characteristics are of practical use in actually managing distributed data in distributed environments. The case studies can be expected to yield characteristics not on the list but that are actually used when performing tasks for management of distributed data in distributed environments.

4.1 Functionalities

Functionalities of data are determined by the role data play in corporate processes. These functionalities are defined by the owner or custodian of the data to fulfill the needs of the stakeholders of the data, and are the starting point for management of data. By definition data are one of five components of an information system together with hardware, software, procedures and people. Data make representations of entities of the real system to enable the information system to obtain knowledge of, to control and to support the real system (see the definition on page 21). Representations of the real system are characterized on two data levels: the pragmatic level and the semantic level [96].

4.1.1 Pragmatics

Pragmatics of data covers the relationships between actual data and behaviour of the real system [96]. Actual data values are the contents of files and databases used by the information system to control the real system. Data acquires meaning within this context of utilization, where data values are determined by the stakeholders within the real system. Stakeholders set requirements and preconditions concerning data values, and maintaining the actual data values in accordance to these requirements is a task for management of data.

Several characteristics can be attributed to data concerning their actual values. These are: correctness, completeness, actuality, timeliness, durability, confidentiality and auditability.

Correctness

Correctness of data requires that data should represent all the relevant corporate entities within the real system [53]. Good control of corporate processes on the basis of data requires data to be a correct representation of corporate entities. This means that validity inspection is necessary to monitor relationships between the data values stored in the database and the reality represented. If, for example, the database at the Dutch joint office for social security administration¹ (GAK) contains data about a person with the surname “Mersel”, this person should appear in the real system of GAK as someone for whom a social insurance administration relationship exists.

Completeness

Completeness of data requires that all relevant corporate entities within the real system should be represented by data [53]. All the properties of the person “Mersel” relevant to awarding a social security benefit must be included. Completeness should hold both for meta-data and for actual data values: all classes of corporate entities

¹in Dutch: Gemeenschappelijk Administratie Kantoor (GAK)

should be described by a data type, and all instances of corporate entities should be described by a data instance.

Actuality

Actuality of data requires that data must describe the most recent state, and therefore the current state, of the real system [48], changes in the real system should be reflected in the data as quickly as possible. Although it is not practical to keep data completely up to date with the real system, effective control of the real system requires the most recent data. Aged data can be a complete and correct representation of a former state of the real system, and as such can be quite valuable, but for effective control of the real system by the information system the most recent data are necessary.

Timeliness

Timeliness of data requires that data has to be available in time to be effective in the control of the real system by the information system [53]. Timeliness of data depends on three types of control of the real system: real-time, interactive (on-line) and non-interactive (batch)

Real-time control is used in process-control where high demands are set for the speed of intervention by the information system. The information system must intervene in the real system within a pre-set period or things could go dramatically wrong. Real-time information systems are highly, or completely automated because the reactive power of a human is not high enough.

Interactive control places less high demands on timeliness of data because man is the limiting factor. The data should be available in time to allow the human operator to intervene in the real system. The period in which the information system must intervene in is not guaranteed, due to the changing load of the information system data can be available earlier or later. Usually there is a upper limit to allow the human operator to work as efficiently as possible.

Non-interactive control determines beforehand what data should available at what time. This is an important difference with the previous two means of control where the information system has to react to changes in the real system. This high level of predictability allows for the planning of the processing of data for the most efficient exploitation of automation facilities. This kind of control is therefore used primarily with expensive mainframes.

Durability

Durability of data requires that the duration of usefulness of data, and therefore the length of storage, in the control of the real system is limited [76]. Data have historic value, for example for making monthly or annual reports. Data are used to describe

4.1 FUNCTIONALITIES

111

the real system during this period. Although such data are no longer correct, complete, and up-to-date, they do have meaning for controlling the real system. Historical data can be used to recreate past situations in the real system, on behalf of analysis or to settle liability claims, because historic data have a low use factor they are often archived to improve efficiency.

Confidentiality

Confidentiality of data requires that access to data should be limited to those people, and applications, that are authorized, [75], because data are accurate representations of corporate entities they need to be as well secured as corporate entities. Trade secrets and privacy-aspects play an important part. To prevent data falling into the wrong, that means unauthorized, hands access to data is restricted to those people and applications that are entitled to access. What is more, people and applications are only allowed to perform those actions they are entitled to perform. These measures taken beforehand are called preventive measures because they prevent illegal actions being taken with the data.

Auditability

Auditability of data requires that it must be possible to check whether confidentiality of data has been breached [53]. If a breach of confidentiality has occurred steps must be taken to limit the damage, to restore the confidentiality of the data and to prevent confidentiality being breached in the same manner in the future. These measures are called repressive measures and corrective measures because they respectively limit the breach in confidentiality and restore the confidentiality of data.

4.1.2 Semantics

Semantics of data is the function and meaning of data defined by the relationships between the data and their meaning in the real world [96]. These relationships are laid down in data definitions that are drawn up by stakeholders within the real system that have a vested interest in functions performed by data on their behalf. Together with the stakeholders, the organization for the management of data draws up detailed descriptions of the data and monitors their application in software and utilization by people.

Several characteristics can be attributed to data concerning their definitions: convertibility, relatibility, domain integrity, static integrity, insert integrity, modification integrity and securability. These characteristics of data are defined by Ter Bekke [4], for an explanation of semantic modelling see also section 5.3.1. Convertibility and relatibility are also referred to as inherent constraints [4], while insert integrity and modification integrity can also be referred to as dynamic constraints [4].

Convertibility

Convertibility requires that there is a one-to-one relationship between the subject and predicate of an assertion. See section 5.3.1 for definitions of predicate and assertion.

At the type level of data convertibility makes every type unique; a data type is allowed only one definition and the same definition must not be used by two different data types. For example, the definitions:

```
type home-address = person, location
type home-address = person, street, number
```

are a violation of convertibility because the data type home-address is specified by different attributes. The definitions:

```
type home-address = person, location
type residential-address = person, location
```

are also a violation of convertibility because the same combination of attributes is used for different data types.

At the instance level of data, convertibility makes every object identifier unique because the identifier may only be used once to denote a particular instance of the type. Convertibility also requires every combination of attribute values to be unique. The instances:

home-address	person	street	number
ID1	Mersel	De Haenstraat	30
ID1	Mersel	Korvezeestraat	436

are a violation of convertibility because the same identification 'ID1' is used for two different combinations of attribute values. The instances:

home-address	person	street	number
ID1	Mersel	De Haenstraat	30
ID2	Mersel	De Haenstraat	30

also violate convertibility because the same combination of attribute values is used for different identifications.

Relatability

Relatability requires every predicate to be related to exactly one subject with the same name. Every subject however can be incorporated as a predicate with a number of subjects.

At the conceptual level of data relatability requires every attribute from a type definition to be defined as a type somewhere, possibly as a base type without attributes or as a compound type. Relatability prohibits removal of a type definition from a data dictionary when other data types use that type definition as an attribute. In relational databases relatability at the type level is known as the foreign key.

4.1 FUNCTIONALITIES

113

At the instance level of data, relatibility requires every attribute value to occur as identification of exactly one instance of the related type. The converse does not need to hold. Relatability prohibits removal of an instance of a type from a data base if another instance uses its identification as attribute value. In relational terms relatibility at the instance level is known as referential integrity.

Domain integrity

Domain integrity requires data to have only certain values, and is somewhat the 'odd man out' because domain integrity relates to the realisation of the conceptual model in a database management system. Domain integrity concerns possible data values and the meaningfulness of data to the real system and therefore is a requirement at the semantic data level. Domains of data are typically specified in the data dictionary.

Static integrity

Static integrity requires attribute values not to contradict one another. The inherent characteristics convertibility and relatibility concern individual data instances, static integrity concerns the relation between aggregated data. Aggregated data is data that is deduced from previously stored data by operations like: total, maximum and average. The deduced data must be updated as soon as the data from which they are deduced alter. In this manner we can be certain that for example, the total amount of an invoice is always the sum of all the amounts specified in the invoice.

Insert integrity

Insert integrity requires a valid transition between two states of the database while inserting an new instance of a data type into the database. After insertion both static and inherent integrity rules must hold. Two types of insert integrity exist: default attribute values when they are not provided during insertion, and default relationships between instances.

Modification integrity

Modification integrity requires a valid transition between two states of the database when modifying attribute values of an instance. Modification integrity specifies for a particular data type the valid relationships between old and new attribute values of all its instances.

Securability

Securability requires limitation of the access of a user or application to the data. At the conceptual level securability limits access to data types. This can be achieved by using "views" of "sub-schema" that provide only some data types to the user, not the

complete data model. At the instance level the user can be limited to for example, read-only operations on the data. This can be achieved by using “security-schema’s” that provide only limit use of the query language.

4.2 Automation facilities

Automation facilities are those components of the information system that automate data structures and data storage. A complete list of automation facilities is described by Looijen [59]. One of the types of automation facilities described by Looijen are database management systems (DBMS) that automate the structures of data. There are two types of data structures: access structures and storage structures. Within this thesis data structures are considered to be independent of the algorithms that manipulate them. Computer media automate the storage of data and give a physical representation of data. Within this thesis the computer media are viewed as part of the data component of information systems and not part of the hardware component of information systems. For example: the magnetic surface of a direct access storage device (DASD), is considered to be part of the data component, but the actuators and the rest of the disk assembly are considered to be part of the hardware component. Automation facilities are characterized on two data levels: the syntactic level and the empiric level [96].

4.2.1 Syntactics

Syntactics of data concerns the computer storage and retrieval of data [96]. Syntactics is related to the structures that are used in the representation of data as sequences of bits within the computer system. The syntactic level does not concern itself with conceptual matters like abstractions or aggregations of data-types of semantic models, but with form and structure of data that are described in logical (physical storage independent) models. These models are however database technology dependent; logical models of CODASYL databases differ from the logical models of relational databases.

The description of formal relations and transformation of one sequence of bits into another can be both static and dynamic. The dynamic descriptions in the form of data processes in application software are maintained by application management of information systems (see section 1.2.3). Management of data concerns itself with maintaining the static descriptions in the form of access structures and storage structures. Examples of access structures are B-trees, hashing and linked lists. Storage structures can be scalar such as floating point numbers and characters or can be structured such as records and arrays.

At the syntactic level of data characteristics that can be attributed to data access structures and storage structures are: access speed, restructuring speed, data capacity, load factor, use factor, transformation, accuracy, concurrency, error detection and error correction.

4.2 AUTOMATION FACILITIES

115

The unit of access for a database management system is a page. One database page can contain several physical pages. A typical physical page size of a storage medium is 512 bytes. A database page can consist of (for example) eight physical pages making a database page 4096 bytes. Large database pages can store more records and thus reduce physical access to secondary storage. Note that access to secondary storage is influenced by the use of buffers, cache and other means described in section 4.2.2. In the characteristics described below, the unit of access is always the database page which is simply referred to a “page”.

Access speed

Access speed, or access complexity, is the number of steps through the access structure that is required to access a particular piece of data [100]. Each step typically requires the retrieval of a page from secondary storage. Two units describe the access speed: the expected number of steps for successful retrieval of data denoted by C , and the number of steps for unsuccessful retrieval denoted by C' . These are expectations of random variables so they usually have non-integer values. Access speed depends on other characteristics of the access structure such as load factor and data capacity. Examples of calculations are given in [2, 24, 100].

Restructuring speed

Restructuring speed, or restructuring complexity, is the number of steps through the access structure that is required to re-arrange it [2]. Every time a record is stored, deleted, or its key is modified, the access structure has to be adjusted. This is done by the database management system (DBMS), and poor choice of the type of access structure will lead to poor performance of the database. As utilization of the database changes over time, initial choices, made when designing the database, will have to be re-evaluated.

Data capacity

Data capacity is the amount of data that can be stored or accessed using of the storage and access structures [2]. Common storage structures within a database are the area, page and records. A database area consists of a number of pages, each page can store a number of records, the bucket-size, and each record can store a number of fields. Access structures have to be able to access a certain amount of (usually) records efficiently within the available storage space, and off course to have minimum overhead for storing the access structure itself.

Load factor

The load factor of a storage structure is the ratio between the space occupied by stored records and the space available in the storage area of the database [100]. The

most fundamental definition of load factor is:

$$\alpha = \frac{N}{Mb}.$$

where N is the number of records, M is the number of pages in a storage area and b is the bucket size of a page.

One of the complications of load factors is the fact that records normally do not have a fixed length. Due to compressed storage (in an effort to reduce space requirements) or variable length fields, records may vary in length. When modifying a variable length record, its size may increase so it no longer fits on a page. This will lead to fragmentation of the record, where the first part of the record is stored on the original page and the second part of the record on another page. This will increase the load factor, and a high load factor has a negative effect on access speed.

Use factor

The use factor of an access structure is the amount of access of that structure by (user) applications during a certain period [83]. Both successful and unsuccessful retrieval of data are counted as accessing the structure. The use factor differs from the access factor in that the access factor counts the number of physical I/O operations to a medium. A high use factor can cause locking conflicts and thus delays in the database. This is the reason why intensively used B^+ -trees should store as few keys in a node as possible. Infrequently used B^+ -trees should however, store as many keys as possible in a node to reduce the height of the tree.

Transformation

The transformation of storage structures in a database is the conversion of one representation of data into another [24, 39]. Examples of where this occurs are between the storage-schema and the external-schema of a CODASYL database, or between views with virtual attributes in a relational database with e.g. averages or totals. Transformation also plays a role with distributed heterogeneous databases where data are placed on different computer platforms [21, p.509]. Other examples are:

- encryption and decryption of data
- compression of data
- translation of ASCII text into EBCDIC and vice versa
- FORTRAN stores arrays by column-major ordering where other programming languages use row-major ordering

These are just a few possible transformations of data. Such transformations are normally performed by systems-software or by hardware, and usually lead to a higher CPU-load.

4.2 AUTOMATION FACILITIES

117

Accuracy

Accuracy of data is the range and precision with which data values can be represented by binary structures [39]. Accuracy of data is important for the precision of floating-point numbers, and for this reason it is inadvisable to use floating-point numbers as a key when retrieving a record in a database. The application program may use a different precision floating-point from the database management system (DBMS). Rounding errors in both calculation and storage of floating-point numbers may cause the DBMS to be unable to find the record. Other types of representations of data values are faced with similar problems. Integers and fixed-point numbers can have overflow and underflow just as floating-point numbers can. Data that is used by multi-media such as (moving) pictures and sound must address accuracy topics like: resolution, sampling rate and colour range of a pixel.

Concurrency

Concurrency is the ability of multiple application programmes to simultaneously initiate database transactions [33]. This has the advantage that the central processor does not have to wait for I/O-operations to complete, however, two problems can arise when performing simultaneous transactions:

- a modification can become lost
- invalid data values can be read

One solution is to place 'locks' on the data involved in the transaction, other transactions have to wait until the data are 'released'. The number of 'locking-conflicts' determines the level of concurrent use of the database.

A second solution is to use 'snapshots' of data in the database. This is primarily used with batch transactions that need the use the database exclusively for a long period of time. A snapshot 'freezes' the current state of the database for use by the transaction. Other transaction are still allowed to modify the database, but these modifications are not applied to the snapshot used by the batch transaction.

Error detection

Error detection is the ability to detect errors in transmission and storage of data [68]. Detecting errors in the data can be achieved by two methods: first through direct inspection of deliberately introduced redundancy in the data. This is what parity bits and cyclic redundancy checks do. A second method is to determine later what the correct value of the data should have been, and establish any differences. This can be achieved through special verification programmes that, for example, check the pointer references in a CODASYL database. This second method places less of a load on the DBMS but has the disadvantage that errors may be detected to late.

Error correction

Error correction, or error recovery, is the ability to restore the database into a state that is known to be correct [22]. Errors fall into one of two categories: local errors and general errors. Local errors cause a single transaction to fail, for example when an integrity rule is violated, and can be corrected using ‘before-image-journals’ that log the transitions in the database. When an transaction error occurs, a ‘rollback’ operation restores the database to a state just before the transaction started. General errors cause all database transactions to fail, for example a head crash or a power failure, these errors are corrected using backups or ‘after-image-journals’ that log former states of the database.

Errors caused by the computer medium are corrected by introducing redundancy into the coding of data. Two popular error correcting codes are: adaptive cross parity code (AXP) used with the IBM 3480 tape cartridge, and the Reed-Solomon code that is used with linear tapes, helical scan tapes, DASD units, and magneto-optical disks [68]. Error correcting codes work on the bit level of data but are also tightly linked with the storage medium, and operate at the boundary of the syntactic and empiric level of data.

4.2.2 Empirics

Empirics is the engineering level of data [96]. The bits mentioned in the syntactical level are given some physical form by modulation of a signal on a medium. The way the bits are represented depends on the storage technology selected and the desired empirical characteristics. Currently many types of media are available, from semi-conductor memories to optical disks, together they form a hierarchy of media. Figure 4.1 shows a storage hierarchy in which along the three axes the three most important characteristics of storage media are placed: access time, storage cost, and storage capacity. Together with the three characteristics mentioned above, other characteristics concerning the physical representation of data are: reliability, repairability, recoverability, transfer rate, safety, retention time, access pattern and fill factor.

Reliability

Reliability of data is the time it takes for a computer medium to fail and for it to be no longer possible to read or write data [68]. There are two modes of failure of a medium: hard errors and soft errors. Hard errors have a physical cause and are permanent. Examples include a head crash and wear of the bearing of the spindle of a direct access storage device (DASD). These errors require repair or replacement of the medium. Soft errors are of a random nature but are not permanent. Examples include an alpha particle hitting a memory cell in a chip, or positioning errors of the head of a DASD caused by temperature fluctuations. These errors can be repaired by repeating the I/O operation or by re-writing the correct data using an error correction method.

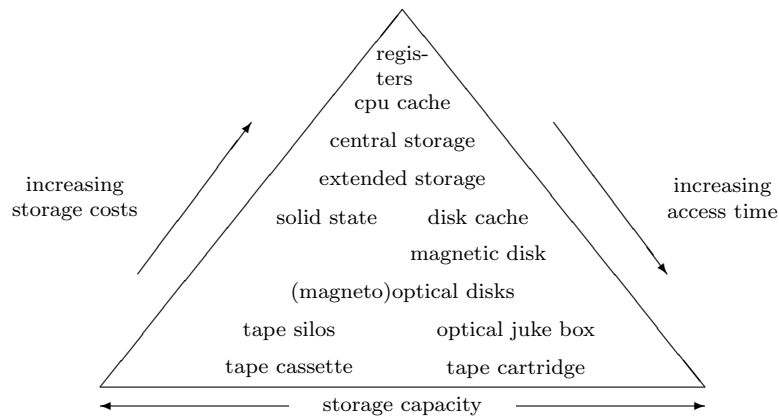


Figure 4.1: Storage hierarchy

The time between failures is usually an exponential distribution probability and is called the survival probability. One of the properties of the survival probability is that it has a constant failure rate, called the Mean Time Between Failure (MTBF). This means that during (most of) the life time of a medium, the expectation that it will fail is the same. In the beginning the medium has a short ‘burn in’ period with a sharply decreasing failure rate, during most of its life the medium has a (low) constant failure rate, then at the end of its life the medium has an increasing failure rate. Together this forms a ‘bath tub’ shaped curve for the failure of media.

The bit error rate (BER) is another way of looking at failure that is normally used for soft errors. The bit error rate is the number of bits that can be read or written without failure. A typical bit error rate for tapes is 10^8 [68].

Repairability

Repairability is the amount of time it takes to repair or replace the medium once the medium has failed [48]. After repair the medium is considered to be as good as new. The length of repair is influenced by a number of (random) factors. Examples include: how long it takes to detect the failure, how long it takes to start the repair and how long it takes to complete the repair. Consequently the repair time is modelled by an exponential distribution that has an expectation called the Mean Time To Repair (MTTR). Availability of the medium is determined by both reliability and repairability [79]:

$$\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$$

After the medium is repaired, or replaced, the data that was originally stored on it, still has to be reclaimed.

Recoverability

Recoverability of data is the amount of time it takes to put the data back on a medium after the medium has been repaired [48]. Recovery can be manual, with backup copies, or automated. Automated recovery of data on a medium is improved by the use of ‘Redundant Array of Inexpensive Disks’ or RAID [47, 54]. Recover time is modelled by an exponential distribution using an expectation called the Mean Time To Recover (MTTREC). Availability of data is therefore:

$$\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR} + \text{MTTREC}}$$

Access time

Access time is the time it takes to read or write the data to or from the medium [100]. The access time is medium dependent. For example the typical access time for ‘single in-line memory modules’ (SIMMs) used in PC’s is 70ns, for direct access storage devices (DASD) typical access time is approximately 21ms [54]. This is 300,000 times slower! Access time for DASD depends on the rotational speed of the disk and how fast the arm moves across the disk.

Transfer rate

The transfer rate, or bandwidth, is the number of bytes per second that can be read from or written to the medium [46]. For DASD this depends on the rotational speed and storage density of the disk, this determines the number of bytes that pass underneath the read/write heads. Transfer rate is increased by dividing the data blocks over several disks and accessing these disks in parallel. This is called ‘striping’ and is used with the redundant array of inexpensive disks (RAID) storage technology [54].

Storage capacity

The storage capacity of a medium is the number of bits the medium can store [100]. With almost any type of computer media the amount that can be stored increases continuously. For tapes the storage capacity depends on the density (bits per inch, or BPI) and the length of the tape. With DASD storage capacity depends on the number of bits per track, the number tracks per surface (cylinders) and the number of surfaces. Note that DASD manufacturers quote the storage capacity in million bytes (10^6) not the usual megabytes (2^{20}).

4.2 AUTOMATION FACILITIES

121

Safety

Safety of data is the restriction of physical access to data stored on the medium [12]. The areas where media are stored should be appropriately secured, and precautions against theft, fire and burglary should be installed. Destruction of a medium (e.g. burning it) when the medium no longer is required is also a safety measure. The consequences of off site storage of data, e.g. of backup tapes, should also be considered with respect to safe storage of data.

Retention time

Retention time of data is the period a bit pattern can be read back from the medium [36]. Data cannot be stored indefinitely on a medium. The period a bit pattern can be read reliably depends on internal and external factors. Examples of external factors are: temperature and humidity, exposure to alpha radiation and ultra-violet radiation. Dynamic RAM has a very short retention time and must be 'refreshed' about every $100\mu s$ [40], estimates for cartridges range from 10 years to 25 years [110].

Storage costs

Storage costs of data is the costs per unit of storage in cents/bit [88]. These costs differ for each medium, and the cheapest medium will not always be the fastest one, or the most reliable. Cost of storage is not only determined by the medium itself, but also by the cost of the floor space the medium occupies, climate control equipment (temperature and humidity are important) and power supply.

Access pattern

The access pattern of data is expressed in the number of read and write operations per seconds performed on the medium [54]. Historical data concerning the access pattern, sometimes called the access profile, is important for noticing trends in utilization of the medium and taking appropriate measures to control it. There is an upper limit to the amount of operations a medium can perform per second. When this limit is reached performance starts to deteriorate.

Fill factor

The fill factor of data is the fraction of the storage capacity that is allocated to data files. This is described in the literature as the loading factor [99]. To avoid confusion with the load factor defined at the syntactical level the term fill factor is used. The difference is that load factor is defined for data files, whereas fill factor is defined for storage devices. Historical data concerning the fill factor, is important for noticing trends in utilization of the medium. A low fill factor means the medium is not used

efficiently, a high fill factor may lead to a short fall in data storage space in the near future. Measures must be taken to control the fill factor and maintain agreed levels.

4.3 Task fields

Several tasks are performed within the management of data to control, operate and maintain data in accordance with user, or stakeholder, requirements and preconditions. These tasks serve to maintain the characteristics of data concerning both functionality and automation facilities. The tasks are grouped into four task fields based upon the four levels of data: pragmatics, semantics, syntactics and empirics. Each task field deals with data and characteristics at its own level. In this sense there is a strong relationship between the F, the A and the T of the FATO framework. The four task fields are: actual data management, data administration, database administration and storage management.

4.3.1 Actual data management

Definition

Management of the quality of the actual data values.

Explanation

This task field takes care of the correct operation of data, and of the processing of mutations in data as quickly and correctly as possible. In the case of distributed databases this implies a proper handling of mutations in every copy of the data in each of the databases. The actual value of the data is determined by the user environment, not by actual data management. This task field is concerned with the fulfillment of the requirements and preconditions of data from a pragmatic point of view.

Tasks

1. Managing authorization of data utilization
 - Granting and modifying of authorizations of utilization of data to applications or users.
 - Monitoring and reporting of (un-)authorized utilization of data to the owner or custodian of the data.
2. Managing actual data contents
 - Controlling and reporting on levels of correctness, completeness, actuality, timeliness, durability, confidentiality and auditability of the actual data according to the requirements and preconditions of the stakeholders.

4.3 TASK FIELDS

123

- Processing mutations of actual data values in each of the databases in which the data are stored.
- Monitoring and reporting on consistency of data stored in each of the databases in which they are stored.
- Removing data from each of the databases the data is stored in, based on criteria set by the owner or custodian of the data.

3. Ad hoc data provision

- Providing copies of, or selections from, data to applications or users that have the proper authorizations.

4.3.2 Data administration

Definition

Documentation of detailed descriptions of data and monitoring utilization of data descriptions by applications and users.

Explanation

Unambiguous description of data is a prerequisite for utilization of such data, in particular when more than one information system makes use of the same data, possibly stored in multiple databases, from possibly different sources. Absence of accurate descriptions can lead to different interpretations by users of the information system (or information systems) and will in most cases lead to ambiguity between stored data and the real world. This task field is concerned with the fulfillment of the requirements and preconditions of data from a semantic point of view.

Tasks

1. Establishing and documenting data names and a data naming standard
 - Establishing or coordinating the establishment of the way data are described and the way the data is to be named.
 - Documenting or coordinating the documentation of the way data are described and the way the data are to be named.
2. Managing the application of data definitions
 - Controlling and reporting on the levels of convertibility, relatability, domain integrity, static integrity, insert integrity, modification integrity and securability of data definitions in accordance with the requirements and preconditions set by the stakeholders.

124

THE FATO FRAMEWORK

- Supervision of the use of data descriptions and the data naming standards, and signaling improper or incorrect application by users.
- Signaling possible shortcomings of the data definitions or the data naming standard.

3. Consultancy

- Give recommendations for improvements in data definitions or data naming standards to the stakeholders , based upon signaled shortcomings, wishes and suggestions of users.

4.3.3 Database administration

Definition

Management of data structures that are documented at a logical level and translated to a physical level.

Explanation

Database administration is concerned with access structures and storage structures. These structures determine the ways in which, and the speed with which data can be accessed once they are stored on computer media. Various automated tools are available to assist in accessing and manipulating data. Attending to modifications of access structures and storage structures as a result of errors or desired performance improvements with respect to retrieval processed are all part of database administration. This task field is concerned with the fulfillment of the requirements and preconditions of data from a syntactic point of view.

Tasks

1. Management of access structures and storage structures

- Management and modification of documentation of access structures and storage structures.
- Attending to modifications of access structures and storage structures.
- Controlling and reporting on the levels of access speed, restructuring speed, data capacity, load factor, use factor, transformation, accuracy, concurrency, error detection and error correction of access structures and storage structures in accordance with the requirements and preconditions set by the stakeholders.

2. Availability of data

4.3 TASK FIELDS

125

- Monitoring and if necessary adjusting of the operation of software and databases.
- Applying reconstruction procedures for the recovery of a database in the event of a calamity.

3. Testing and evaluation

- Setting testing standards for newly developed or modified database software.
- Measurement of the workload and making recommendations for tuning in case of too high a workload of the database.

4.3.4 Storage management**Definition**

Management of the physical data as they are stored on computer media that is incorporated in the technical infrastructure.

Explanation

Data are stored in files and databases on a wide range of computer media, ranging from solid state devices and direct access storage devices (DASD) to disk-arrays and automated tape storage systems utilizing robots. The data forms part an information system or is shared by several information systems. Storage management must take preventive measures against the mutilation of data, either accidental or intentional. None the less, in the event of mutilation of data, storage management should have installed effective measures to restore the data to their original state. This task field is concerned with the fulfillment of the requirements and preconditions of data from an empiric point of view.

Tasks

1. Managing the storage of data

- Controlling and reporting the levels of reliability, repairability, recoverability, access time, transfer rate, storage capacity, safety, retention time, storage costs, access pattern and fill factor of data storage in accordance with requirements and preconditions set by the stakeholders.
- Allocation of space for storage of data and the type of media where data is stored, is based upon the requirements and preconditions of the stakeholder. Sometimes not enough space is allocated and has to be 'extended'. This involves the allocation of extra space, usually not contiguous, on the same medium.

- Migration moves data from one type of medium to another, typically up or down the storage hierarchy, this is known as ‘staging’ of data. Migration allows data to use fast media when active, and to use cheaper media when inactive. This allows more efficient use of the fast but expensive media.
- Scratching involves removing the data altogether when the data is no longer needed.
- Reorganization is sometimes necessary if a number of extends cause fragmentation of the space allocated for storage and slow down access to the data.

2. Safe keeping

- Storing a medium safe from damage by, for example, fire, water or other (external) influences. Preferably at a location away from where the data are normally used, but not too far away, the medium must be retrievable within a required time.

3. Reconstruction

- Reconstruction of data stored on a medium after mutilation by some failure by retrieving a back-up copy of the data from safe keeping.

4. Destruction

- When computer media are discarded, all contents should be wiped. This prevents someone from outside the organization reading the data. This includes DASD units replaced by a service mechanic.

4.4 Organization

The functionalities (F), the automation facilities (A) and the tasks fields (T) that are involved in the management of data have been described in the preceding sections. Given this F.A.T. structure there has to be some kind of organization that performs task fields, this organization is the ‘O’ that completes the FATO framework. This organization can be as small as one person or as large as an entire department, and can be centralized or decentralized, concentrated or deconcentrated.

4.4.1 External influences

Figure 4.2 shows the management paradigm specialized for management of data with six external influences along its sides. External influences enable new possibilities, or impose constraints upon an organization for data management that enhance or limit the possibilities for implementing the data management tasks fields [51].

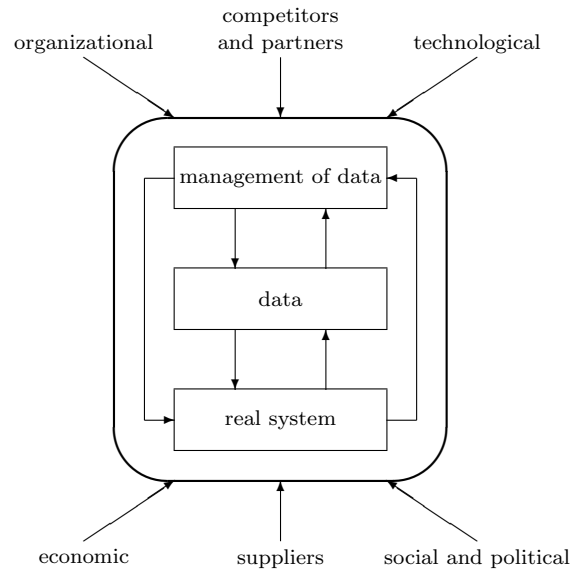


Figure 4.2: External influences

Organizational

Developments within the users organization requires other controlling mechanisms, other organizational structures and other data than it uses currently. Mergers and strategic alliances lead to new needs for data about business entities and business processes. The introduction of electronic data interchange (EDI) is good example of this, where different management organizations suddenly have to work together and have to adjust to one another, the new data that has to be managed, and the changes in the real system.

Automation facilities are redistributed within the users organization, out of the computer centre into the users organization or back into the computer centre. Other examples of recent interest are outsourcing and downsizing of the management organization.

Competitors and partners

Organizations responsible for management of data increasingly have to compete in the market for service providers. User organizations are increasingly concentrating on their "core business" and are making their automation departments independent of the rest of their organization.

Data are components of one or more information systems, and this means that the department in charge of management of data cooperates with departments that manage other components of the information system, for example, when information systems are linked within a value chain. Another kind of partnership can be seen when the organization for management of data itself has some tasks performed by others, for example specialized companies for the destruction of computer media.

Technological

Computer media are subject to continuous improvement. On the technology push side storage capacities increase, while size, cost and access times decrease. On the application pull side, information systems of the future will no longer rely on characters and numbers, but on new data formats such as MPEG² and JPEG³. Multi media techniques integrating numerical data with sound and moving images already make use of every technological innovation available.

Economic

While the cost of data storage keeps falling, the volume of data storage keeps rising. The costs of personnel are however generally outside the influence of the department for management of data, and are subject to trade union negotiations with the mother organization. One of the aspects is the choice whether management of data should function as a cost centre, as a service center or as a profit centre. The last option only being realistic when management of data is an independent organization and not a daughter of a mother organization.

Suppliers

Management of data to a large extent depends on its suppliers. A supplier may go bankrupt, or could decide to no longer support or supply a particular automation facility. In a fast evolving technological area, management of data should signal such eventualities at an early stage and anticipate a course of action.

Social and Political

Society exerts all kinds of influences on management of data. Examples include: public opinion and government regulations concerning privacy sensitivity, restrictions on data encryption and cross border data flows.

²MPEG stands for the Motion Picture Expert Group that defined this standard.

³JPEG stands for the Joint Picture Expert Group that defined this standard.

4.4 ORGANIZATION

129

4.4.2 Tactical and strategic task fields

Together with the external organizational aspects, management of data must deal with internal organizational aspects. The internal organizational aspects are the three M's: Money, Men and Means, and these organizational aspects require task fields at tactical and strategic levels of management of data. They are not as specific for management of data as the task fields at the operational level of the organization mentioned in section 4.3, but are an important part of management of information systems as a whole. Looijen describes all task fields involved in the management of information systems [59]. Some examples of task fields are described briefly below, they include: strategic management, tactical management, personnel, procedures, quality assurance, charging and obtaining automation facilities.

Strategic management

Strategic management of data has to maintain contacts with all those who are able to exert external influence on the organization. Tasks with respect to strategic management are:

1. Direct supervision [73]. Strategic management assigns means of production and issues orders, takes important decisions, structures and staffs the organization, and also motivates and rewards employees.
2. Governing the organizational boundary conditions [73]. These conditions are set by the external influences of section 4.4.1. Strategic management spends a large portion of its time informing and affecting influential parties in the environment and holding negotiations and concluding important agreements with external partners.
3. Developing an organizational strategy [73]. The strategic apex tries to make the organization respond to changes in the environment without strong disturbances within the organization itself, by interpreting the wishes of the stakeholders that constitute the environment and the internal forces and needs of the organization.

Tactical management

Tactical management coordinates activities at the operational level by direct supervision, and by managing and assigning time, money and means to the operational level. Tasks with respect to tactical management are:

1. Gathering 'feed-back information' about the performance of the unit they lead. This information is passed partially or completely to a higher ranking manager.
2. Taking decisions allowed at its management level.

3. All the task mentioned for strategic management, limited to their own unit. The strategy of the unit should of course fit within the organizational strategy. The line-manager should maintain relations with influential parties within the environment of his unit.

Personnel management

Managing the personal aspects of the organization, by standardization of skills. Tasks with respect to personnel management are:

1. Draft job descriptions
2. Recruiting and selecting new personnel
3. Training personnel
4. Career supervision

Procedures

Improving the quality of management of data by standardization of work processes. This is the first component of the quality system for management of data. Tasks with respect to procedures are:

1. Defining and documenting the work processes of management of data.
2. Performing management audits to establish if the procedures are correctly implemented.
3. Evaluating procedures and improving them where necessary.

Quality assurance

Inspection of the quality of data by standardization of the conformity of the characteristics of data. This is the second component of the quality system for management of data. Tasks with respect to quality assurance are:

1. Defining and introducing quality standards.
2. Inspecting conformity of data characteristics using the quality standards.
3. Evaluating quality standards: are the quality standards feasible, do the standards continue to meet requirements and preconditions, and do the standards need improvement?

4.4 ORGANIZATION

131

Charging

Charging the services provided by the management of data to the consumer. Tasks with respect to charging are:

1. Costing, calculating the cost price of the services provided
2. Establishing an expense calculation system
3. Establishing a budgeting system
4. Invoicing

Obtaining automation facilities

All actions that are necessary to acquire automation facilities. Tasks with respect to obtaining automation facilities are:

1. Obtaining quotations from possible suppliers or contractors
2. Evaluation of quotations
3. Placing orders

These seven task fields are just a few examples of the task fields defined by Looijen [59], because these task fields are directly taken from management of information systems, this thesis assumes they are valid for the management of data and are therefore not tested in the cases described in chapters 6, 7 and 8. Figure 4.3 shows the task fields of management of data positioned within the organizational diagram of Mintzberg [73]. This is an indication of where these task fields can be placed within an organization. Further organizational development is necessary for management of data to become a complete organization, but this is outside the scope of this thesis

4.4.3 Functional management and technical management

Situational or contingency factors [73] that lead to particular organizational structures and decentralization are also outside the scope of this thesis. One factor however is important for the management of data: this is the segregation of functions to prevent a conflict of interest. This conflict necessitates two different organizational units: functional management and technical management [59, 76]. The division into two organizational units is based upon differences in expertise, responsibilities and substance of the tasks performed.

Functional management provides optimal support of users and business processes with functional expertise, maximizing effectiveness of data, and is responsible for the functional requirements and preconditions for data set by the stakeholders. The tasks of functional management are oriented towards the utilization of data.

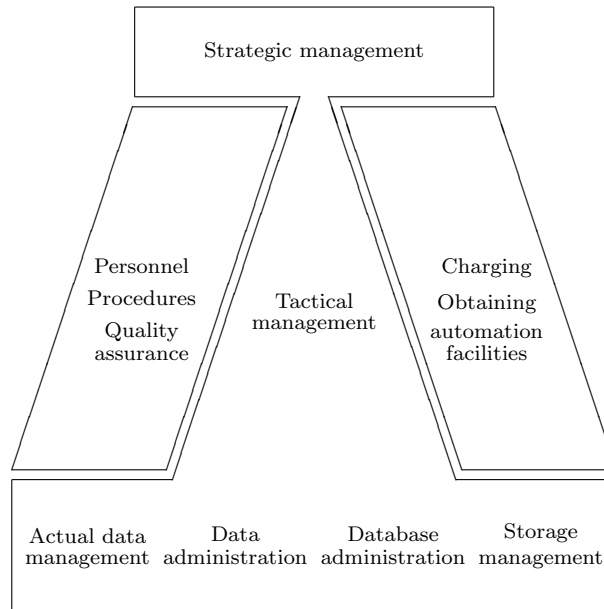


Figure 4.3: Organizational positions of task fields

Technical management provides optimal exploitation of available automation facilities with technical expertise, maximizing efficiency of data, and is responsible for the technical requirements and preconditions of data set by the stakeholders. The tasks of technical management are oriented towards the exploitation of data.

4.5 Positioning distribution within the FATO framework

Distribution can be classified, using the FATO framework, into three view points: functional distribution (F), technical distribution (A) and organizational distribution (O). Task fields (T) are not an independent factor in distribution, because they are tightly linked with the pragmatics, semantics, syntactics and empirics of data. Task fields are therefore distributed together with the functionalities and the automation facilities.

Figure 4.4 shows a graphic representation of the distribution of the environment, the data and management. On the left is a typical personal computer situation with

4.5 POSITIONING DISTRIBUTION WITHIN THE FATO FRAMEWORK

133

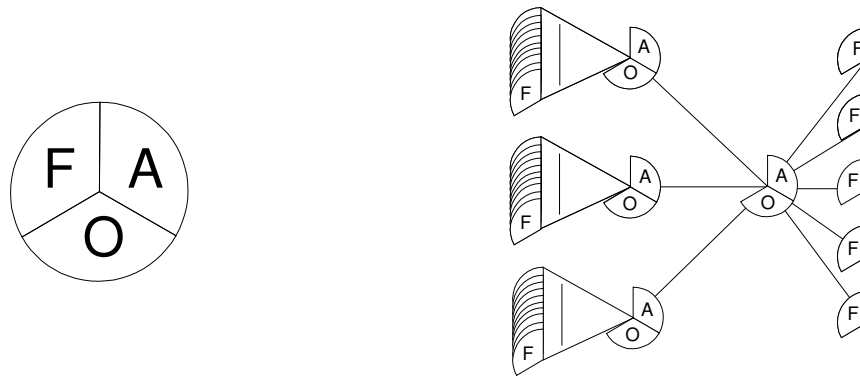


Figure 4.4: Diagram for distribution

functionalities, automation facilities, as well as the management organization and the user, all concentrated at one location. On the right is a situation where functionalities, automation facilities and the management organization are deconcentrated (dispersed) over several locations.

4.5.1 Functional distribution

Functional distribution relates to the utilization of data by the stakeholders or real system. The stakeholders or real system are the environment of data where the data perform their functionalities, and the distribution of functionalities is a measure for the distribution of the environment. Requirements and preconditions for actual data and data definitions can vary for each location of the environment. Functional management should take precautions to maintain functional coherence of data across locations.

4.5.2 Technical distribution

Technical distribution relates to the exploitation of automation facilities at different locations. The data storage, data structures and the automation facilities used to implement them can be at many location. Requirements and preconditions for structures and storage of data can vary for each of the locations of the automation facilities. Technical management should take precautions to maintain technical coherence of data across locations.

4.5.3 Organizational distribution

Organizational distribution relates to the way the organization of management of data is located at several sites and is a measure for the deconcentration of management of data. The tasks of management of data are carried out where the functionalities and the automation facilities are, but the people who perform these tasks can be at completely different locations. For example, data can be stored at many locations but can be management from a single location. Organizational distribution may be caused by local autonomy of organizational units involved in management of data. Segregation of functions into two distinct organizational units, functional management and technical management, also contributes to organizational distribution of management of data. Some mechanisms must be in place to achieve coordination between the tasks performed by each organizational unit [73]. Management of data should take precautions to maintain coherence between the organizational units at different locations, and to maintain the coherence between technical management and functional management.

Chapter 5

Methods and techniques

This chapter gives examples of methods and techniques to be used for management of distributed data in distributed environments and what kinds of adaptations are required to suit methods and techniques to the needs of data management. The management paradigm described in section 1.2.1 uses the recursion principle to show that management itself can be viewed as an information system at a higher level that serves to operate, control and to maintain information systems that are then considered at the level of real systems. This makes it possible to take well known methods and techniques that are already available for the description of information systems, and apply these to management of information systems.

The methods and techniques selected for this thesis take the systelological perspective [107] where the effect of the information system on the real system is considered. As is shown in figure 1.5, data management is considered in this thesis as the information system that controls, operates and maintains data as the real system. The informational, datalogical and technological perspectives of information systems [56, 94] are outside the scope of this thesis, this includes topics like: tools for computer assisted software engineering, distributed database management systems, systems managed storage band other solutions that can be applied to management of distributed data in distributed environments.

Business System Planning (BSP) [50] and “information systems work and analysis of change” (ISAC) [62] are well known methods that use methods and techniques described in literature as applicable to the systelological level of information systems [64, 80]. These methods and techniques from BSP and ISAC are used to describe management of distributed data in distributed environments. They are methods for use during planning (state IPP) and development (state D) of information systems (see section 1.2.2) and due to the inherent differences within management, adaptations are needed if they are to be used within the states Utilization (U) and Exploitation (E). It is important that methods and techniques used by management link up with the methods and techniques used in planning and development because

management has to work with documentation handed over from planning and development.

The next section describes what kinds of adaptations are required due to the different natures of management and planning and development. Specific adaptations of methods and techniques are described in sections concerning environments, data and management respectively. The last section of this chapter describes how methods and techniques are positioned within the FATO framework.

5.1 Adaptations required for management

Methods and techniques are used in information systems management, and data management, to assure that once a system is implemented, it keeps satisfying the requirements and preconditions. This perspective is described in the second research question of page 32 repeated below:

What methods and techniques can be made available to management to operate, control and maintain distributed data in accordance to the requirements and preconditions of a distributed environment and the characteristics of the distributed data?

One approach to how methods and techniques can be used to support development of information systems is described by De Moel and Van Hulzen [74]. In this approach methods and techniques are used to support four aspects of information systems development: (1) definition of requirements and preconditions of the information system; (2) translation of requirements and preconditions into properties of the information system; (3) measures needed to translate requirements and preconditions into properties; and (4) control mechanisms used during development of the information system. Performance characteristics of information systems by De Moel and Van Hulzen [48] are used to group previously defined requirements and preconditions with properties of information systems (see also page 52). This approach to position methods and techniques for development (state D) of information systems is adopted for management of utilization (state U) and exploitation (state E) of data. The kinds of adaptations needed to make methods and techniques applicable from a management perspective are independent of the particular method or technique involved, and are described below using the FATO framework defined in chapter 4.

Methods and techniques used to determine requirements, preconditions (1) and properties (2) of data in the 'F' and 'A' of the FATO framework need adaptation of the characteristics used to group them. The characteristics of information systems defined by De Moel and Van Hulzen, need to be replaced by the characteristics of data at the pragmatic, semantic, syntactic and empiric levels of data such as defined by the FATO framework.

Methods and techniques used to support the measures (3) taken in tasks of data management in the 'T' of the FATO framework need adaptation of the way they

5.2 ENVIRONMENT

137

are executed. Measures taken by management tasks in the states utilization (U) and exploitation (E) are closed control loops (feed-forward and feed-back) where properties of data are compared with requirements and preconditions and action is taken to correct deviations of properties from these requirements and preconditions.

Methods and techniques used to support the control mechanisms (4) used in functional data management and technical data management in the ‘O’ of the FATO framework need adaptation of the nature of the control mechanisms. The nature of the control mechanisms is determined by the nature of the states in the state model for information systems management described in section 1.2.2. The steady states utilization (U) and exploitation (E) of the state model contain tasks that are performed continuously during the whole life span of data, and therefore permanent organizational structures and coordination mechanisms need to be in place. The transient states development (D) and maintenance (M) contain tasks that are performed during short and predetermined periods of time, and project management organizations are better suited to this situation.

5.2 Environment

The environment of data consists of the stakeholders that have a vested interest in that data perform their functionalities within the requirements and preconditions set by the stakeholders. Stakeholders do not need to be people, end-users that use data as well as other automated information systems that depend on data are considered to be stakeholders of the data. From a data management perspective there is little difference at the operational level if stakeholders are within the boundaries of the real system or outside the real system, this is a general organizational aspect that is outside the scope of this thesis. The methods and techniques described in the next four sections are used by data management to support the stakeholders in utilizing the functionalities offered by data. These four methods and techniques are: the list of functionalities, the list of stakeholders, the functionality matrix and the create/use matrix.

5.2.1 List of functionalities

BSP considers objectives of information systems to “properly support the development, implementation and operation of the planned system” [50, p.3], and ISAC considers objectives of information systems to select between change alternatives. Data management uses the list of functionalities to determine if data continue to perform these functionalities on behalf of the environment. The list of functionalities such as is used within the case studies contains all functionalities that have to be performed by data on behalf of the stakeholders.

5.2.2 List of stakeholders

BSP considers business processes that use and create data and therefore make requirements of data. ISAC considers stakeholders as people who require certain problems to be solved by developing new information systems [62, p. 29]. The list of stakeholders it used by data management to identify who has an interest in the proper functioning of data. The list of stakeholders such as is used within the case studies and contains those parties of interest that require the functionalities offered by the data.

5.2.3 Functionality matrix

BSP shows distribution of business processes over locations and organizational units but not of objectives. ISAC relates objectives to problems in a table that shows the alternatives for change, but does not relate objectives to stakeholders. The functionalities are not required by all stakeholders, a functionality matrix helps data management to keep track of which stakeholders need what functionalities. A functionality matrix also enables data management to determine the level of support required by stakeholders by examining what functionalities are required by most stakeholders; the higher the number of stakeholders that require a functionality the higher the level of support required. The functionality matrix shows the distribution of functionalities of data over different stakeholders, and indicates the functional distribution of data.

5.2.4 Create/Use matrix

BSP uses create/use (C/U) matrices to show what business processes create or use which data classes, and these are used to later determine where a given database should be stored physically. The C/U-matrix is used by data management to control the authority of stakeholders to determine what actual data values should be (create) or to apply actual data values in their business processes (use). In this C/U-matrix business processes are replaced by stakeholders, and data classes are replaced by data types that are defined by the semantic model described in section 5.3.1.

5.3 Data

The different perspectives of data development and data management also require adaptations of methods and techniques. The development perspective of the methods and techniques considers data at the level of the information system. The management perspective of methods and techniques considers data at the level of the real system, where data is operated, controlled, and maintained by data management. This change of perspective is illustrated by the management paradigm shown in figure 1.1. This enables data management to consider data as a relatively independent entity to be managed in accordance with requirements and preconditions and within the external constraints on management. Due to the specific nature of data, the methods and

5.3 DATA

139

techniques used by BSP and by ISAC to describe real systems need modifications to enable them to describe the data characteristics. The methods and techniques described in the next four sections are used by data management to operate, control and maintain the requirements, preconditions and properties for data characteristics. These four methods and techniques are: the semantic model, the automation facilities matrix, the table of properties and the table of stakeholders.

5.3.1 Semantic model

BSP considers data classes without relationships between data classes at the semantic level or syntactic level, each data class may be a separate databases or all the data classes may be stored in a single database, at the level of business system planning this distinction is not important. ISAC does not consider data modelling at the systelological level during the 'analysis of change' phase of the method, but defers this to the infological level of information systems analysis. The data characteristics at the semantic level in section 4.1.2 are defined using the semantic model defined by Ter Bekke [4], the same semantic model is used to describe data as an object of management.

The semantic model is used to discover data requirements and preconditions and the data properties at the semantic level. A conceptual model of data normally exists but semantic requirements are not usually defined during development. Using the semantic model enables management to determine what the requirements and preconditions are, and using the data dictionary it is possible to determine the semantic properties of data. The semantic model is used by data management to check if distributed data that are implemented by unrelated development projects have the same definitions. If this is not the case, special interfaces must translate between different data definitions when data are transferred.

5.3.2 Automation facilities matrix

BSP considers storing data at a single location or at multiple locations but only at the level of data classes as described with the functionality matrix, automation facilities are not considered. Automation facilities concern data structures and data storage and data management needs to keep track of the different locations where computer media used for storing data are placed and where the various database management systems are placed. The automation facilities matrix shows the locations where automation facilities are kept, and thereby shows the technical distribution of data.

5.3.3 Table of properties

BSP defines a characteristics matrix where requirements and preconditions for data classes are listed. The characteristics defined by BSP are: use and audit, security,

occurrence, currency, volume/response and criticality. These correspond with data characteristics at the pragmatic level defined in chapter 4, BSP fails to defines characteristics at the semantic, syntactic and empiric level and also fails to describe how to determine properties. ISAC defines property tables that are used to quantify activities described in the activity model (see section 5.4.1). It however fails to specify what the requirements and preconditions are and ISAC does not provide predetermined characteristics to express properties. ISAC is not able to compare requirements and preconditions with properties and therefore cannot describe associated actions to correct deviations. The table of properties is used by data management to determine if data still meet requirements and preconditions made by the stakeholders. The table compares, for each data characteristic, requirements and preconditions with actual properties.

5.3.4 Table of stakeholders

The table of stakeholders is used by ISAC to relate stakeholders to problems that must be solved by developing information systems, and activities within the real system that are related to these problems. The table of stakeholders is used by data management to define the relationships between the stakeholders, what requirements and preconditions are made by a particular stakeholder, and what activities are performed by management with respect to these requirements and preconditions. To this end two adaptations are necessary: problems are replaced by requirements and preconditions, and activities are moved to the information system level where they describe data management (see section 5.4.1).

5.4 Management

In the management paradigm (see page 26) is shown that management of data is regarded as the information system for the knowledge, control and support of data (analogous of the information system definition on page 21). The methods and techniques described in the next three sections, that are employed to describe data management, take into consideration the fact that they are not only used to describe data management, but are also used by data management in performing and coordinating its tasks. These three methods and techniques are: the activity model, the table of RAW distribution and coordination mechanisms.

5.4.1 Activity model

Activity models are used within the ISAC method to model activities in the organization where an information system needs to be developed to solve the problems of this organization. The activity model uses three kinds of symbols: sets, flows and activities. Sets can consist of physical entities of the real system (people or materials), messages of the information system (documents or telephone messages) or mixed sets

5.4 MANAGEMENT

141

of both types of entities. ISAC describes activities at the level of the real system, and flows show how physical entities and messages are used by activities in the real system.

The recursion principle of the management paradigm enables adaptation of the activity model for data management, where activity models are used to describe activities performing data management tasks. Examples of using activity models to describe management tasks are given by Esseling and Van Nimwegen in [32, p.26] and by Bakker in [3, p.3]. By applying the recursion principle of the management paradigm, physical entities of the real system are replaced by data that is controlled by data management. Messages of the information system are replaced by documentation that is used by data management, and mixed sets that consist of both data and documentation to be used by data management. Activities concern the tasks that are performed by data management and flows show what data and documentation are used by management to perform its tasks.

5.4.2 Table of RAWE distribution

The table of distribution of responsibilities (R), authorization (A), work (W) and expertise (E) is used by Information Strategy Planning (ISP) [13] to describe management tasks in the complete life cycle of information systems. The RAWE table for tasks within data management is used in this thesis to show the organizational distribution of data. The RAWE table describes what organizational units are involved in tasks performed by data management defined in chapter 4. This table could for example show that an organizational units performs a task (W) but it lacks the skills (E) needed to perform this task, or that an organizational unit is responsible (R) for a task but lacks the authorization (A) to take the required decisions.

5.4.3 Coordinating mechanisms

Coordinating mechanisms describe how different organizational units of data management coordinate their tasks [73]. Section 4.4 shows that data management is an organization like any other and how tasks fields of data management are positioned within an organizational diagram. Coordination mechanisms link the individual task of the activity models with the organizational units of management in the RAWE-table. Five coordinating mechanisms are fundamental in the coordination of work within organizations [73, p. 4]:

Mutual adjustment achieves coordination by informal communication between the participants at the operating core of the organization. This is a direct communication that is initiated by the participants in an ad hoc fashion when they feel they need it.

Direct supervision achieves coordination by having one person at the strategic apex of the organization taking the responsibility for the work of others. The

strategic apex issues instructions to other people in the organization and personally monitors the execution of these orders by them.

Standardization of work achieves coordination by specifying rules and procedures regarding to performing work. These rules and procedures prescribe precisely in which order work is to take place.

Standardization of output achieves coordination by specifying the norms, these are specifications of the results of the work, and product specifications and quality inspection play an important part. This type of coordination can also be applied to the financial results of a organizational unit, where a certain amount of turnover or profit is required of it.

Standardization of skills achieves coordination by specifying the amount of skills or knowledge that is required to be able to perform the work. This can be achieved by training of people already employed, or by selecting potential employees with the appropriate education from a group of applicants.

5.5 Positioning methods and techniques within the FATO framework

To complete this chapter the position of methods and techniques within the FATO framework is described below and is shown in table 5.1. The number of methods and techniques may seem to vary between the four components of the FATO framework however, the table of properties uses every data characteristic defined in chapter 4 and the activity model uses every task defined in chapter 4.

Methods and techniques to support functionalities (F) describe the data environment and data at the pragmatic and semantic level, consequently many methods and techniques are placed within functionalities. This is in line with the fact that functional management is required to both manage data and support their utilization. Distribution of the environment is described using the functionality matrix, and the functionality matrix and the create/use matrix together describe functional distribution of data. A table of properties and a table of stakeholders are used to describe functionalities concerning data characteristics at the pragmatic and semantics levels.

Methods and techniques to support automation facilities (A) describe data at the syntactic and empiric levels. Technical distribution of data is described using the automation facilities matrix. The table of properties and table of stakeholders are used to describe automation facilities concerning data characteristics at the syntactic and empiric levels.

Methods and techniques to support task fields (T) describe the tasks that are performed by data management. The activity model is used to describe activities that perform tasks within the four task fields of data management defined in chapter 4.

Functionalities	List of stakeholders List of functionalities Functionality matrix Create/Use matrix Semantic model Table of properties Table of stakeholders
Automation facilities	Automation facilities matrix Table of properties Table of stakeholders
Task fields	Activity model
Organization	Table of RAWE distribution Coordinating mechanisms

Table 5.1: Methods and techniques positioned within FATO

Methods and techniques to support organization (O) describe the organization structure of data management. Organizational distribution of data is described using the RAWE table, that shows what tasks are performed by what organizational unit. As the physical location of each organization unit is known, the distribution of the management organization is also known. Coordinating mechanisms describe how coordination between the distributed organizational units is achieved.

5.6 Testing the management instrumentation

The results of the first and second research objectives are described in chapter 4 and in this chapter. The framework that contains all concepts that are applicable to management of distributed data in distributed environments was specified in chapter 4. In this chapter methods and techniques to manage distributed data in distributed environments are provided and it is shown how the management perspective requires adaptations of methods and techniques traditionally used for planning and development of information systems. Together, the FATO framework and the methods and techniques presented here provide an instrumentation for management of distributed data in distributed environments. The adaptations of the methods and techniques were tested in practical situations of management of distributed data in distributed environments. The third research objective entailed testing the adequacy of this instrumentation in specific situations where distributed data are managed in a distributed environment. The results of these tests are described in chapters 6, 7 and 8. The final conclusions and the research results are given in chapter 9. To draw such cross cases conclusions, a comparison of contrasts and of similarities of the three case studies is necessary. This is possible by showing the types of distribution of the

three case studies. Positioning of the distribution of a case is achieved by showing the deconcentration of data based upon the FATO framework as described in section 4.5 and by showing the decentralization of the environment, see also section 1.4.7.

Chapter 6

Personal Data Register Case

Chapters 4 and 5 provide an instrumentation for management of distributed data in distributed environments. A generic framework in which characteristics, tasks and organization involved with data management are positioned with respect to each other is described chapter 4. How generic adaptations provide methods and techniques suited for use by data management is described chapter 5. The case study described in this chapter provides a specific test for the framework and the methods and techniques used in the management of a distributed data infrastructure (see figure 3.7): the personal data register¹ (BRP).

This case study applies the management instrumentation in a specific situation that is characterised by two aspects. The first aspect of the case study is the fact that the BRP is used to control several different primary (production) processes within the Dutch Joint Office for Social Security Administration² (GAK). The BRP is part of the data infrastructure within GAK and is used to control the enforcement of various legislation concerning social security (see also section 3.2.2). Different utilization of the same data by the various stakeholders involved in corporate processes may lead to multiple, and perhaps conflicting, requirements and preconditions for the same data characteristic. The second aspect of this case study is the fact that data of the BRP is distributed over one central site and three regional sites, where the regional data is a simplification of the central data. The regional data serve a specific purpose for (social) legislative information systems used at thirty district offices of GAK³. There are six databases, three central and three regional, where BRP data is stored, each exhibiting its own properties for the data characteristics.

The main objective of this case study was to test the adequacy of the management instrumentation by its application in a practical situation for analysis of data management of the BRP and to provide suggestions for improvements. In the description

¹In Dutch: Basisregistratie Personen (BRP)

²In Dutch: Gemeenschappelijk Administratie Kantoor (GAK)

³In section 3.2.2 legislative information systems were referred to as process administrations

of the case study emphasis is placed on functional aspects of the BRP [71].

Section 6.1 positions this case study with respect to deconcentration of BRP data and decentralization of the BRP environment to enable cross case analysis in chapter 9, where the final research results are presented. Section 6.2 gives descriptions of the environment of the BRP using the methods and techniques applied in this case study. Section 6.3 describes the data characteristics and the distribution principle of the BRP. Section 6.4 describes management of the BRP with the tasks and organizational units involved in performing these tasks. Section 6.5 gives the results of application of the management instrumentation in this case study and concludes in what way the management instrumentation is adequate for management of this specific data infrastructure and what adaptations are required in this specific situation.

6.1 Positioning the case

Positioning of the case study is achieved by following the guidelines given in section 1.4.7. Within the context of this thesis, distribution of the Personal Data Register (BRP) is restricted to geographical deconcentration of functionalities, automation facilities and the organizational management units, of the BRP using the diagram for distribution of data shown in figure 4.4 at the end of chapter 4. The 'FAO'-diagram for the BRP is shown in figure 6.1. The distribution of the environment is restricted within this thesis to the decentralization of the decisional powers over the BRP within the Joint Office for Social Security Administration (GAK) using the organizational diagram of Mintzberg [73]. The Mintzberg diagram for the BRP is shown in figure 6.2.

6.1.1 Deconcentration

Deconcentration of data of the Personal Data Register is described using the FATO framework with a diagram for graphic representation where distribution of functionalities (F), distribution of automation facilities (A) and distribution of the management organization (O) are shown. See also section 4.5.

Distribution of functionalities of the BRP is shown by the placement of 'F' segments in figure 6.1. Functionalities of BRP are highly distributed. In the extreme case BRP performs its functionalities for approximately seven million people with social security insurances administered by GAK and that are registered within the BRP. For the sake of simplicity of the diagram in figure 6.1, but not for simplicity of data management, only one 'F' segment is drawn at the right hand side of the 'FAO'-diagram. All functionalities for stakeholders outside the GAK organization are provided by the central service centre (CSC). Functionalities required by the head office of GAK are also provided by the CSC, these functionalities are also represented by 'F' segments placed at the right hand side of the 'FAO'-diagram. There are thirty district offices within GAK distributed throughout The Netherlands, and at every district office there are functionaries that use data stored within the BRP to perform

6.1 POSITIONING THE CASE

147

their every day duties. Ten district offices are connected to one of three regional service centres (RSCs)⁴ where copies of the BRP are placed. The functionalities used by the GAK functionaries at the district offices are represented by three groups of ten 'F' segments placed at the left hand side of the 'FAO'-diagram. Specific information systems for control of administrative processes for different social security legislations are located at the RSCs and they are (directly) connected to the regional copy of the BRP at the respective RSCs. These information systems are represented by three 'F' segments placed at the three RSCs at the centre left hand side of the 'FAO'-diagram.

Distribution of automation facilities of the BRP is shown by the placement of 'A' segments in figure 6.1. Automation facilities that are employed on behalf of the BRP are placed at the central service centre (CSC) in Amsterdam and three regional service centres (RSCs) in The Hague, Eindhoven and Hengelo. The main database of the BRP is placed at the central service centre, to improve performance (of batch-processing) an identical copy and an simplified copy of BRP data are also placed at the central service centre. The databases and accompanying automating facilities at the central service centre (CSC) are represented by the 'A' segment placed at the CSC at the centre right hand side of the 'FAO diagram'. A database containing simplified copies of BRP data is placed at each of the regional service centres, so in total there are six database that together constitute the personal data register (BRP). The databases and accompanying automating facilities at the regional service centres (RSCs) are represented by three 'A' segments placed at the three RSCs at the centre left hand side of the 'FAO'-diagram. The mechanisms employed to maintain connections between the databases at the RSCs and the databases at the CSC that enable the BRP to function as a coherent system are described in section 6.3.3.

Distribution of the management organization of the BRP is shown by the placement of 'O' segments in figure 6.1. Organizational units for management of the BRP are located at various locations of GAK. Functional management is performed at each of the thirty district offices where a special functionary can modify directly the contents of the database, they are represented by three groups of ten 'O' segments placed at the left hand side of the 'FAO'-diagram. Functional management of BRP data is performed at the CSC by the account management unit (AMU), it is represented by an 'O' segment at the CSC at the right hand side of the 'FAO'-diagram. Technical management of the regional copies of the BRP is performed at the three RSCs, data management personnel at the RSCs performs the tasks of media management involved with producing backups of data. Technical management at the regional service centres (RSCs) is represented by three 'O' segments placed at the three RSCs at the centre left hand side of the 'FAO'-diagram. Technical management tasks concerning media management of the central BRP databases are performed at the CSC, this is represented by the 'O' segment placed at the CSC at the centre right hand side of the 'FAO diagram'. All other technical management tasks for the regional as well as the central BRP databases, are performed by the "systems management" unit at a

⁴In section 3.2.2 the central service centre and regional service centres were referred to as central computer centre (CCC) and regional computer centres (RCCs)

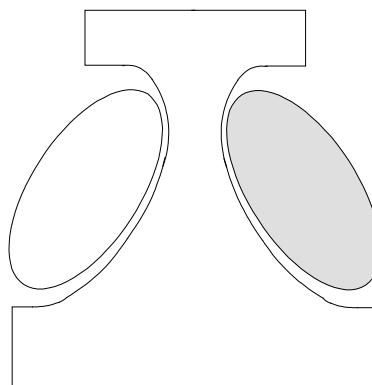
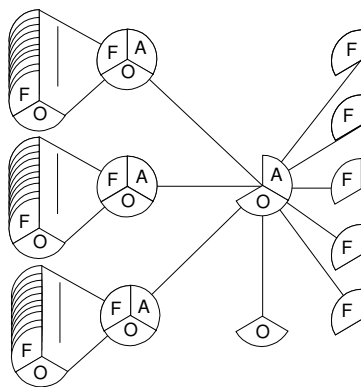


Figure 6.1: Deconcentration of BRP

Figure 6.2: Decentralization of BRP

separate location a short distance from the central service centre. At this location are also the project group BRP and the project group data management that perform tasks of functional data management. This separate location is represented by the 'O' segment placed below the CSC at the centre right hand side of the 'FAO'-diagram.

6.1.2 Decentralization

Decentralization of the environment of the Personal Data Register (BRP) is described using the distribution of decisional powers over the user organization in which data management is performed. The data environment consists of stakeholders with a vested interest in the functionalities performed by the data. For the sake of simplicity the many stakeholders that are outside the GAK organization are considered to be at the operating core of the GAK organization. The power to decide over the functionalities and automation facilities of the BRP rests with the custodian of the BRP, see section 2.1.3. The level of decentralization of the BRP is determined by the location within the GAK organization where the power rests to make decision about data in the BRP.

The fact that the BRP is part of the data-infrastructure of GAK (see figure 3.7) leads to the expectation that the central support staff of GAK has a major influence in the decision making process around the BRP, and this is in fact the case. Custodianship of the BRP rests with the department Information management and Automation (I&A) of GAK which is a support unit for the execution of social security legislation at the operating core of GAK. The department I&A decides on the functionalities to be performed by the BRP as well as the automation facilities to be employed. The

level of decentralization within the environment of the BRP is thereby characterized as a type D decentralization (see page 34), the Mintzberg diagram shown in figure 6.2 illustrates this type of decentralization.

6.2 Environment

The stakeholders and the functionalities of the MPD identified during the case study using the methods and techniques of the management instrumentation are discussed in section 6.2.1.

6.2.1 Stakeholders and functionalities

In the first section the stakeholders of the Personal Data Register (BRP) that were identified during the case study are discussed using the list of stakeholders, shown in table 6.1. In the second section the functionalities the BRP performs for the stakeholders are discussed, these functionalities were identified during the case study using the list of functionalities shown in table 6.2. In the third section the functional distribution of the BRP is discussed. The distribution the BRP functionalities over the stakeholders is identified using the functionality matrix shown in table 6.3. The distribution of the authority to determine the actual data contents or to use the actual data contents was identified using the create/use matrix shown in table 6.4.

Stakeholders

Stakeholders have an interest in the functionalities that are performed by the BRP. Stakeholders do not only comprise the organizational units or individuals within the GAK organization or outside of GAK, but they also comprise the automated information systems that depend on the BRP in some form or other. The stakeholders of BRP are shown in table 6.1.

The board of directors of GAK are the owners of the BRP and as such are responsible for the BRP conforming to all relevant legislation, in particular privacy legislation. The board of directors has laid down rules and requirements in a privacy regulation concerning the collection and distribution of personal data with regards to those persons who are registered by GAK on behalf of social security administration.

The direct users of the BRP are employees at the GAK head office and the district offices that have the ability to mutate directly data stored in the BRP to correct mistakes. All the other employees at the district offices can access BRP data only through the legislative information systems.

Legislative information systems are used for the administration of particular social security legislations (the various social insurances are discussed in section 3.1.2). The legislative information systems may contain data that is redundant with respect to the BRP or with respect to other legislative information systems. All the legislative information systems are connected with the BRP by the ADB-mechanism described

in section 6.3.3. The legislative system RESA/FASA is an information system for the administration of two social security laws: the workers occupational disability act (WAO) and the general occupational disability act (AAW). The legislative system IMF is an information system for the health insurance act (ZW). The legislative system AAW-conveniences (AVZ) is an information system for the conveniences act which compensates for extra costs of for example modifications to homes to make them suitable for the handicapped, additional information on AVZ can be found in [78]. The legislative system GMD is an information system for the joint medical services (GMD), the GMD covers all the medical aspects of the social insurances. The legislative system NWW is an information system for the unemployment act (WW), it will be replaced by the new information system WWO which is discussed in chapter 7.

Each employer is obliged to pass on to GAK all the data, and mutations in data, concerning labour contracts and the personal data of all his employees. The employers can send this data using a form, tape or diskette, to the district office where this data is entered into the information system 'Notification Occupational Association' (MBV) from where it is entered into the BRP (and BRDV see page 95) using the ADB-mechanism. Although the information systems is called Notification Occupational Association, notification is in fact the responsibility of the employer.

The 'wages and premium determination and inspection' (LPI) department at GAK determines how much social insurance premiums have to be paid by employers and employees. The department LPI also enters into the BRP the personal data supplied by the various pension funds that are administered by GAK.

The labour contracts register (BRDV) contains data about the labour contracts between employers and employees, and it uses the BRP to access the personal data hold on employees.

The access management information system (TBS) is an information system that controls and monitors access to the technical infrastructure of GAK. As such it contains data about the organizational structure of GAK and all of the employees of GAK and all of the automated information system at GAK, that are granted access to specific components of the GAK technical infrastructure. Additional information on TBS can be found in [113].

Sometimes employees at the head office and the district offices need to have access to the combined data which is stored in the three subject registers BRP, BRDV and BRWG (see page 95), without having to go through a legislative information system. A special application program 'combined access subject registers' (RBR) has been made to enable this type of retrieval.

The statistical research department supports the board of directors by providing trends and other figures relating to developments in social security and matters of policy at GAK.

The Municipal Population Administration network (GBA) supplies authentic personal data to the BRP directly from the municipal administrations. When personal data is changed, e.g. when someone changes address or marital status, this is au-

6.2 ENVIRONMENT

151

tomatically passed on to the BRP. Some municipalities are not connected to the GBA-network so these municipalities use forms or tapes to supply personal data to the BRP.

The social insurances council (SVr) determines regulations for all agencies involved in the administration and enforcement of social security insurances. The council also determines what personal data is to be stored in the BRP, in what format, to enable exchange of personal data between social security agencies.

PTT Post is the source of authentic postal addresses in The Netherlands. All the addresses in The Netherlands are stored in the BRP based upon the postal code table supplied and updated by PTT Post. This ensures that only valid addresses are possible and prevent a person from entering a false address when applying for a social security benefit. Sometimes GAK discovers mistakes in the postal codes which are reported to PTT Post and consequently corrected in the update tapes which are supplied by PTT Post each month.

The tax offices are the authentic source of social-fiscal (SoFi) numbers of persons stored in the BRP. Verification of SoFi numbers is performed on a regular basis by exchanging tapes with the government computer centre. When the treasury cannot find a valid SoFi number for a person, most commonly because of spelling ambiguities, the district office where the personal data is administered sends a verification request to the tax office. Verification of persons living abroad takes place at the tax office in Brunssum.

The ministry of Finance provides the addresses of the tax offices which are stored in the BRP, any additional tax data is also obtained from the ministry of Finance.

The registered person is the second authentic source of the personal data and social-fiscal number stored in the BRP. The registered person has the right to inspect stored personal data, to correct errors found in the personal data, and has the right to request a statement of all third parties to which GAK has provided the personal data. The registered person is however compelled to pass on to GAK any corrections to personal data.

The register external cross-references (REV) is an information system that registers which persons are also registered by occupational associations that are not a member of GAK (see page 88). Exchange of data between occupational associations enables fraud investigations and benefit concurrence investigations.

The electronic sickness notification (EZA) is an information system which uses EDI to enable occupational associations to report sickness of employees of companies that are member of the association electronically.

Social insurance registration notification (STOV) is an information system which provides the registered person with a statement of all the relevant data from the 'insured persons administration' (VZA). These statements are sent directly after a person has been registered by the VZA, of which the BRP is a component see section 3.2.2, and are sent once every year. The BRP supplies the SoFi number, name, address, and date of birth of registered persons on these statements.

Notification occupational association via Videotex (MBX) is an information sys-

B1	Board of directors GAK
B2	Direct users BRP
B3	Legislative information system RESA/FASA
B4	Legislative information system IMF
B5	Legislative information system AAW-conveniences
B6	Legislative information system GMD
B7	Legislative information system NWW
B8	Notification occupational association
B9	Wages and premium determination and inspection
B10	Labour contracts register
B11	Access management information system
B12	Combined access subject registers
B13	Statistical research department
B14	Municipal Population Administration network
B15	Municipalities
B16	Social insurances council
B17	PTT Post
B18	Tax offices
B19	Ministry of Finance
B20	Registered person
B21	Register external cross-references
B22	Electronic sickness notification
B23	Social insurance registration statement
B24	Notification occupational association via Videotex
B25	Integrated information flows
B26	Other employees GAK
B27	Third parties outside GAK

Table 6.1: List of stakeholders BRP

tem that makes it possible for employers to use the Videotex communication system to pass on personal data and labour contracts data concerning employees. The MBX system is an alternative for employers who prefer to use electronic means rather than the traditional forms that need to be send to the MBV system, which serves the same function.

The ‘integrated information flows’ (GIS) information system enables GAK employees to combine data from all the legislative information systems and subject registers that make up the VZA (see section 3.2.2). It is used for the communication between GAK and employers, examples include sending annual statements of wages and premiums determined by GAK, reports to occupational associations and wages determination on behalf of health insurance benefits.

Not all employees of GAK are involved directly in performing administration of social security legislation and related regulations. Occasionally they require personal data to enable them to perform their tasks. For example the lawyers at GAK who need personal data to take legal action against persons who have committed fraud.

6.2 ENVIRONMENT

153

Occasionally other agencies may require personal data which is registered by GAK, for example municipal social service departments. The privacy regulations determined by the GAK board of directors contains an exhaustive list of all third parties that are entitled to access personal data.

Functionalities

The functionalities of the BRP are listed in four different documents present at GAK in four different ways [71], these four lists were combined to create the list of functionalities shown in table 6.2.

BRP is the authentic source of personal data within GAK. There are a number of benefits for GAK having its own authentic source of personal data. One, this means GAK has to pay only once to obtain the data; e.g. GAK has to pay to verify personal data with the GBA network. Two, GAK is guaranteed that personal data is always available. Three, it guarantees that all legislative information systems use the same personal data and that legislative information systems can be coupled. Four, it enable GAK to comply with legislation which prescribes the construction of the insured persons administration (VZA).

The BRP registers references to legislative information systems that also store personal data. This enables GAK to check the claims made by a registered person under one or several social regulations, the legislative informations systems use it to determine multiple claims and to settle more efficiently a benefit application made by a person.

The BRP verifies personal data with municipalities and maintains data based upon mutations supplied by the municipalities. Verification request are sent to the municipalities who return the request with mutations to personal data when necessary. Verification requests can sent to the municipalities by exchanging forms or tapes, or by sending the verification request over the GBA network. GAK subscribes to the GBA network, this means after initial verification, any alteration in personal data are automatically passed on to GAK without the need to send a verification request.

BRP verifies the Social-Fiscal (SoFi) number of a person with the Tax office. The verification request is returned by the Tax office with mutations on the SoFi number. The guiding principle at GAK is that execution of social security regulations takes priority, and therefore personal data is verified later. If a person has received an undeserved social benefit, this can always be reclaimed. The SoFi number is essential to GAK because all its social security data, e.g. legislative information systems, are linked using the SoFi number as their key characteristic (e.g. as the primary-key in relational tables). The SoFi number is essential to be able to grant a social security benefit, however, sometimes a person has yet to obtain a SoFi number. The direct user BRP can create a fictitious SoFi number until the tax office provides the true SoFi number. By using SoFi numbers to link social security data, mutations in SoFi numbers have far reaching consequences for the GAK databases.

BRP processes mutations in personal data supplied by legislative information sys-

F1	BRP is the authentic source of personal data within GAK.
F2	BRP registers references to legislative information systems that also store personal data.
F3	BRP verifies personal data with municipalities and maintains data based upon mutations supplied by the municipalities.
F4	BRP verifies the Social-Fiscal (SoFi) number of a person with the Tax office. The verification request is returned by the Tax office with mutations of the SoFi number.
F5	BRP processes mutations in personal data supplied by legislative information systems.
F6	BRP processes mutations in addresses supplied by the PTT and reports unknown addresses to PTT.
F7	Each year BRP sends registered persons a summary of data stored by the BRP: SoFi number, name, address and date of birth.
F8	Whenever a person is registered within the the BRP, this person receives a notification within a month.
F9	BRP supplies personal data to employees of GAK involved with cases concerning social legislation or with management of BRP.
F10	BRP supplies on an ad hoc basis only specific personal data to other employees of GAK when this data is necessary to perform their work.
F11	BRP supplies on an ad hoc basis only specific personal data to third parties outside GAK when this data is necessary to perform their work.
F12	BRP provides the registered person, on request, with a list of all parties to whom, within the last 12 months, personal data has been supplied.
F13	BRP supplies data on behalf of policies planning and strategic management of the execution of social security legislation and related regulations.
F14	BRP destroys all data concerning a person five years after the need for registration has ceased to exist. Data held on behalf of statistical research are excluded.
F15	BRP supplies extracts of personal data of registered persons to each district office of GAK.

Table 6.2: List of functionalities BRP

tems. The mutations in personal data are distributed by the BRP to other legislative information systems during the night, more details on mutation distribution can be found in section 6.3.3.

BRP processes mutations in addresses supplied by the PTT and reports unknown addresses to the PTT. The BRP contains a table of all streets and postal codes known to the PTT. This enable addresses entered in the BRP to be verified automatically. Each month the PTT supplies a tape with mutations in addresses and postal codes. Sometimes an address or postal code is not yet known or contains errors, these are reported to the PTT.

Each year BRP sends registered persons a summary of data stored by the BRP: SoFi number, name, address and date of birth. This statement is sent to comply with Dutch privacy legislation, and to enable the registered person to correct errors found in the personal data.

Whenever a person is registered within the the BRP, this person receives a notifi-

6.2 ENVIRONMENT

155

cation within a month. This statement is sent to inform the registered person of the registration and to enable the registered person to correct errors.

The BRP supplies personal data to employees of GAK involved with cases concerning social legislation or with management of BRP. Access to BRP data is controlled by the ‘access management information system’ (TBS) which records the authorizations for each individual GAK employee.

The BRP supplies on an ad hoc basis only specific personal data to other employees of GAK when this data is necessary to perform their work. Privacy regulations limit the conditions under which personal data is made available to GAK employees not directly involved in administering social security legislations.

The BRP supplies on an ad hoc basis only specific personal data to third parties outside GAK when this data is necessary to perform their work. Privacy regulations limit the conditions under which personal data is made available to third parties.

The BRP provides a registered person, on request, with a list of all parties to whom, within the last 12 months, personal data has been supplied. The BRP supplies data on behalf of policy planning and strategic management of the execution of social security legislation and related regulations.

The BRP destroys all data concerning a person five years after the need for registration has ceased to exist, data held on behalf of statistical research are however excluded. The need for registration of a person ceases when a person has died, when a person has become 65 year of age (or more) and no longer has entitlements to social security such as a old age pension, or when this person was registered on behalf of an investigation that concluded that the person did not have to be compulsorily insured. In the past six year no personal data has been removed from the BRP. Some data is not allowed to be destroyed due to the Social Insurance council regulations that prescribe that samples of personal data need to be retained for statistical research.

BRP supplies extracts of personal data of registered persons to each district office of GAK. Each district office maintains local copies of personal data in its legislative information systems, for each of the people for whom it performs social security administration. Mutations in personal data are distributed by the BRP to each of these extracts, see section 6.3.3 for more details.

Functional distribution

The functionalities performed by the BRP for a particular stakeholders, indicates the importance of the BRP to the stakeholders and consequently the amount the support the stakeholder needs to receive from the functional data management organization. The documentation present at GAK did not indicate which functionality was built into the BRP on the request of which stakeholder. The distribution of BRP functionalities over the stakeholders is shown in the functionality matrix given in table 6.3, which is based on interviews with data management personnel at the departments IMAP and R&I (see section 6.4.2).

From the functionality matrix it can be concluded that, functionally speaking, the

legislative information systems are identical. By counting the number of functionalities for each stakeholder, it is possible to determine the relative importance of each of the stakeholders. The stakeholder 'registered person' has the most functionalities performed by the BRP, followed by the BRP's direct users, the legislative information systems and the statistical research department. The access management information system (TBS) and the Ministry of Finance are stakeholders for whom the BRP does not perform any functionalities, although they are still recognized as stakeholders in documentation.

By counting the number of stakeholders for each functionality, it is possible to determine the relative importance of each of the functionalities. The verification function of the BRP is the one most often required by the stakeholders, followed by the functions processing of mutations and acting as authentic source of personal data. Ad hoc provision of personal data and the notification sent to the person registered are required by one stakeholder only and are thus less important. From the functionality matrix it can also be concluded that there are no unnecessary functionalities as all of the functionalities are performed on behalf of at least one stakeholder.

The board of directors officially determines which GAK employees and other stakeholders are authorized to create the actual data contents of the BRP or to use the actual data contents of the BRP, in practice these authorizations are determined by the custodian of the BRP (see section 6.4.2). This distribution of authority over the BRP stakeholders is shown by the create/use matrix in table 6.4. The authority of a stakeholder to create the actual data contents of a particular data type in the BRP is shown by a "c" in the matrix, the authority of a stakeholder to use the actual data contents of a particular data type in the BRP is shown by a "u" in the matrix. The first column of the matrix contains the names of the data types of the BRP, more details of the data types can be found in section 6.3.2. The create/use matrix can be used to determine the levels of authority of the various stakeholders.

The board of directors is the owner of the BRP, and determines a large number of regulations for the use of the BRP, e.g. the privacy regulations, but they do not determine directly the actual data contents of the BRP nor do they make direct use of the actual data contents of the BRP. Most authority to create and use the actual data contents of the BRP rests with the direct users of the BRP, the legislative information systems, the combined access subject registers system (RBR) and the statistical research department. The statistical research department is given a complete copy of the BRP database to perform statistical studies. The results of these studies are used by the board of directors to determine GAK policies. The BRP direct user and PTT Post are the only stakeholders who are allowed to create the streets and postal codes stored by the BRP. When a person who is registered passes on an non-existent address, the BRP direct user enters it as an unknown address and notifies PTT Post to verify the address. The BRP direct user uses the 'control' data type to create a fictitious SoFi number when the true SoFi number of a person is not yet known.

6.2 ENVIRONMENT 157

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	B23	B24	B25	B26	B27
F1		X	X	X	X	X	X	X	X			X	X								X	X	X				
F2		X	X	X	X	X	X					X	X									X					
F3		X	X	X	X	X	X	X	X	X		X	X	X	X					X	X	X					
F4		X	X	X	X	X	X			X		X	X							X	X	X					
F5		X	X	X	X	X	X			X		X	X							X	X	X					
F6		X	X	X	X	X	X			X		X					X			X	X						
F7																				X	X						
F8																				X	X						
F9		X										X													X		
F10																									X		
F11																											X
F12													X							X							
F13	X												X							X							
F14													X			X				X							
F15			X	X	X	X	X																				

Table 6.3: Functionality matrix BRP

158

PERSONAL DATA REGISTER CASE

Data type	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	B23	B24	B25	B26	B27
administration	cu							c				u	u			c						u				u	
foreign address	cu	u	u	u	u	u	u					u	u			c			c	cu					u		u
address organization unit	cu	u	u	u	u	u	u				c	u	u			c				cu					u		u
address person	cu	u	u	u	u	u	u			u		u	u	c						cu					u		u
address group		u	u	u	u	u	u			u		u	u				c										
freepost address	cu											u	u				c										
country	cu	u	u	u	u	u	u					u	u			c									u		u
location	cu	u	u	u	u	u	u			u		u	u												u		u
mutation	cu											u	u														
mutation person	cu	u	u	u	u	u	u					u	u														
GBA entryno.								c				u	u	cu		c									c		
GAK member	cu										c	u	u			c											
organization unit	cu	u	u	u	u	u	u					u	u			c			c								
person cluster	cu	u	u	u	u	u	u					u	u			c						u		c	u	u	u
person description	cu	cu	cu	cu	cu	cu	cu	c	c	u		u	u	c						cu		u	u	c	u	u	u
person	cu	cu	cu	cu	cu	cu	cu	c	c			u	u					c		cu	cu	u					
postal code	cu	u	u	u	u	u	u					u	u														
street	cu	u	u	u	u	u	u					u	u				c										
control	cu																										
verification settlement	cu	cu	cu	cu	cu	cu	cu					u	u														
verification request	cu	cu	cu	cu	cu	cu	cu					u	u			c											
cross reference	cu	cu	cu							u		u	u												c		
task allocation	c											u	u														

Table 6.4: Create/Use matrix BRP

6.3 Data

In the sections below the data characteristics, as defined in chapter 4, are used in the description of the requirements and preconditions of BRP data. This shows how in practice the data management organization of the BRP implements the data characteristics that have been defined theoretically. In the description of this case study the emphasis is placed upon functional aspects of management of the BRP, and consequently conformity of properties of the data characteristics to the requirements and preconditions is restricted to characteristics at the pragmatic and semantic levels of data. In section 6.3.1 the data characteristics of the BRP at the pragmatic level are used to structure the requirements and preconditions. In section 6.3.2 the data characteristics of the BRP at the semantic level are used to structure the requirements and preconditions. In section 6.3.3 attention is given to the distribution of data of the BRP and the implementation of mutation distribution and retrieval distribution of the BRP. In section 6.3.4 a short description is given of the requirements and preconditions and properties at the syntactic and empiric level.

6.3.1 Pragmatic level

At the pragmatic level of data requirements and preconditions concerning the actual data values are attributed to seven data characteristics: correctness, completeness, actuality, timeliness, durability, confidentiality and auditability. See section 4.1.1 for more details.

The requirements and preconditions are defined using all of the data characteristics. A summary of the pragmatic data characteristics is shown in table 6.5, the number of requirements that are made on the pragmatic data characteristics is shown in the second column, the number of requirements that are fulfilled by properties of the BRP is shown in the third column, the number of additional properties of the BRP upon which no requirements are made is shown in the last column.

The requirements and preconditions for the pragmatic data characteristics are assigned a unique number. These numbers are used in the table of properties that were drawn up during the case study. An example of a table of properties for the data characteristic auditability is shown in table 6.6. The numbers are abbreviations of the data characteristics: correctness (C), completeness (V), actuality (A), timeliness online (To), timeliness batch (Tb), durability (D), confidentiality (Ve) and auditability (Au). The requirements and preconditions are listed in table 6.7 using these numbers.

The table of stakeholders shown in table 6.7 gives the relationships between the stakeholders, the requirements and preconditions made by those stakeholders, and what activities are performed by BRP data management related to these requirements and preconditions. The stakeholders are identified by their number as listed in table 6.1. The requirements and preconditions are identified by their number used in the tables of properties which were one the methods and techniques employed during this case study [71]. From table 6.7 it can be concluded that BRP data management

characteristic	requirements and preconditions	fulfilled by properties	additional properties
correctness	12	4	7
completeness	8	1	4
actuality	13	1	1
timeliness (on-line)	9	0	0
timeliness (batch)	6	0	0
durability	3	1	0
confidentiality	11	2	0
auditability	3	2	0
total	65	10	12

Table 6.5: Summary pragmatic characteristics

No. Data type	Requirement and precondition	Property
Au1 Complete database	The same person is not allowed to input, mutate, retrieve and verify the same data	property unknown
Au2 Complete database	Measures against misuse and unauthorized use must be auditable	Audit-logging and alarms of the DBMS are switched off
Au3 Complete database	Registration of rights of individual employees	The access management information system (TBS) registers authorized transactions of individual users.

Table 6.6: Table of properties auditability

does not know which stakeholder made what requirements, and table 6.7 demonstrates that it is not clear to BRP data management which activities are performed to fulfill the requirements and preconditions made by the stakeholders.

Correctness The BRP stores more than just personal data. The BRP contains a complete list of all Dutch municipalities, of all the nations and nationalities of the world, the addresses of the local tax offices, the occupational associations, and the pension funds. The BRP also contains the complete organizational structure of GAK to determine which organizational units are entitled to access parts of the BRP, this data is updated automatically by the access management information system (TBS).

Social-Fiscal numbers are verified with the treasury by monthly exchange of tapes with the government computer centre. The treasury uses the name, address and date of birth to find a person's SoFi number and it enters this number on the tape when it

6.3 DATA

161

Stakeholder	Requirements and preconditions	Management activities
Board of directors GAK	A2 B1 C8 D2 D3 Tb1 Ve6 Ve7 Ve8 Ve8 Ve9 Ve10 Ve11 Vo1 Vo6	unknown
Direct users BRP	A9 To7 To9	unknown
Information system RESA/FASA	A8 A13 To5 To6 To8	unknown
Information system IMF	A8 A13 To5 To6 To8	unknown
Information system AVZ	A8 A13 To5 To6 To8	unknown
Information system GMD	A8 A13 To5 To6 To8	unknown
Information system NWW	A8 A13 To5 To6 To8	unknown
Occupational associations		unknown
Wages and premiums		unknown
Labour contracts register		unknown
Access management information system		unknown
Combined access	A8 A13 To5 To6 To8	unknown
Statistical department.		unknown
Municipal population adm. network	A10 Vo8 Tb2	unknown
Municipalities		unknown
Social insurances council	C7 C9 C10 C11 C12 Vo2	unknown
PTT post		unknown
Tax offices	A11 Vo7 Tb3	unknown
Ministry of Finance		unknown
Registered person		unknown
Register external cross references.		unknown
Electronic sickness notification		unknown
Registration statement		unknown
Videotex		unknown
Integrated information flows		unknown
Other employees GAK		unknown
Third parties outside GAK		unknown
Unknown	A1 A3 A4 A5 A6 A7 A12 C1 C2 C3 C4 C5 C6 Au1 Au2 Au3 D1 To1 To2 To3 To4 Tb4 Tb5 Tb6 Ve1 Ve2 Ve3 Ve4 Ve5 Vo3 Vo4 Vo5	unknown

Table 6.7: Table of stakeholders pragmatics BRP

sends the tape back to GAK. Occasionally the treasury cannot find the SoFi number based on the personal data supplied by GAK, or the treasury returns a SoFi number already in use by another person in the BRP. In such cases the local tax offices are contacted to carry out further investigations. Each week PTT Post supplies a tape with mutations of postal codes in The Netherlands. GAK reports addresses with unknown postal codes to PTT Post.

Personal data is stored in the BRP in a shortened form. A person's first name is shortened to 12 characters, and a person's last name is shortened to 18 characters.

Completeness The GAK functionaries at the district offices or the head office are obliged to correct errors in personal data when requested by the person registered. The privacy regulations specify exactly what data is allowed to be stored in the BRP

and forbid explicitly the storage of any other data. The BRP violates the privacy regulations by storing a GBA Administration number, and adding other records to the database. Comparison of data specified in documentation of the functional management organization and of the application management organization with data actually stored in the BRP shows that there is no agreement between either specifications or BRP data.

Actuality The BRP databases are updated with the following frequency: each month postal codes are updated with PTT mutations, each month a tape is exchanged with the treasury to verify SoFi numbers, each night mutations in the central database are distributed to the regional copies, each night mutations in the central database are copied by the central copy BR. Each night mutations in the BRP are listed and sorted according to district office and legislative information system and stored in a separate mutation file. These mutations files are processed at the RSCs by the legislative information systems.

When a person moves between different municipalities the municipality where the person moves from blocks this person's address. The BRP cannot verify the address with the GBA network, and has to wait four weeks until a verification request can be sent to the municipality the person moves to.

Timeliness The requirements and preconditions on on-line timeliness and batch timeliness are defined in great detail, however, no properties of timeliness are known to BRP data management as they are not measured. The required maximum on-line response time at a terminal for 90% of retrieval transactions is five seconds. The required maximum on-line response time at a terminal for mutation transactions is nine seconds. The required maximum average on-line response time at a terminal is five seconds. Preconditions under which on-line response times are valid are: at most 10,000 mutations a day on personal data, on-line access is permitted Monday through Friday from 7,30 until 17,30.

Batch processes must be able to process 200,000 mutations per weekend and 400,000 mutations per month. Exclusive use of BRP data by batch jobs is permitted on Monday through Thursday from 18,30 until 2,00 on Friday from 0,00 until 2,00 and on Saturday and Sunday from 0,00 until 24,00. Shared use of BRP data by batch jobs is permitted on Monday through Thursday from 18,30 until 7,30 on Friday from 0,00 until 7,30 and on Saturday and Sunday from 0,00 until 24,00.

Durability Personal data concerning a person who is registered must be destroyed five years after the grounds for registration have expired. The need for registration of a person ceases when this person has died, has become 65 years of age (or more) and no longer has entitlements to social security such as a old age pension, or when this person was registered on behalf of an investigation which concluded that this person did not have to have compulsory insurance. In the past six year no personal data has been removed from the BRP. Some data is not allowed to be destroyed due to

6.3 DATA

163

the Social Insurance council regulations that prescribe that samples of personal data need to be retained for statistical research.

In reality personal data has never been removed from the BRP, as the BRP is only six years old this comes as no surprise, however, no data management activities are defined that would enable the deletion of redundant data.

Confidentiality Requirements and preconditions on the data characteristic confidentiality are defined in great detail, however, only two are fulfilled. The access management information system (TBS) limits access for every individual BRP user (both human and automated) to a list of transactions allowed to be executed. Data management personnel is allowed to access BRP data directly. The BRP is accessed using individual authorizations, TBS does not allow 'group accounts'.

Auditability Procedures and programs which perform input, mutation, retrieval or verification of personal data are secured using separations of functions. It should not be allowed that the same person is able to input, mutate, retrieve and verify the same personal data, however, as procedures and programs are not managed by data management these properties were not determined. The measures which are taken against misuse and unauthorized use of the BRP must be auditable, despite this, audit logging and the alarms of the DBMS are switched off. The BRP must maintain a record of rights of individual employees, the access management information system (TBS) registers authorized transactions of individual users. Table 6.6 gives an example of a table of properties in which the requirements and properties of the characteristic auditability are placed.

6.3.2 Semantic level

At the semantic data level, requirements and preconditions concerning data definitions are attributed to seven data characteristics: convertibility, relatability, domain integrity, static integrity, insert integrity, modification integrity and securability. See section 4.1.2 for more details.

There were two sources of semantic requirements and preconditions, one from the systems development department and the other from the account management unit, which were found to be in conflict with each other [71]. The documents from the systems development departments and the account management unit were used to specify the requirements and preconditions on the semantics characteristics of the BRP. The properties of the semantic characteristics were determined by examining the print out of the CODASYL DBMS data dictionary in which record definitions and set definitions are laid down together with the integrity specifications such as: 'insertion automatic retention fixed'.

All of the semantic data characteristics defined in section 4.1.2 are employed in either the requirements and preconditions or in the requirements on the BRP. A summary of the semantic data characteristics is shown in table 6.8, the number of requirements that are made on the semantic data characteristics is shown in the

characteristic	requirements and preconditions	fulfilled by properties	additional properties
convertibility	27	0	7
relatability	27	2	7
domain integrity	57	12	20
static integrity	19	0	1
insert integrity	3	0	13
modification integrity	0	0	1
securability	0	0	3
total	133	14	52

Table 6.8: Summary semantic characteristics

second column, the number of requirements that are fulfilled by properties of the BRP is shown in the third column, the number of additional properties of the BRP upon which no requirements are made is shown in the last column.

The data model of the central database is shown in figure 6.3, many of the semantic requirements and preconditions are defined in the semantic model. The semantic data properties of the central database and the regional copies of the BRP differ. The regional copies store the personal data in a single large record instead of using the extensive data model.

Convertibility The data model of the BRP contains a large number of data type definitions, the documentation lists different attributes than can be found in the data dictionary of the BRP database. In addition the BRP data dictionary contains a translation table of the Social-Fiscal numbers issued by the treasury and the GBA Administration numbers used by the municipalities to interchange data of their populations administration automatically. The BRP data dictionary contains six other data definitions not mentioned in the documentation. The ‘control’ record is a data type used by the DK-BR functionary to generate a fictitious SoFi number for persons who need to be registered by GAK but have not been issued a SoFi number. The SoFi number is essential to GAK because all its legislative information systems are linked using the SoFi number. There is always exactly one instance of the ‘control’ record.

Relatability The data model of the BRP defines a large number of relational dependencies of data, the documentation only agrees with two of the properties of relatability found in the data dictionary. The CODASYL definitions of relatability in a set are used to determine the properties; e.g. insertion automatic retention fixed. Relatability between data in the CODASYL database is, in some circumstances, defined in manner found in relational databases; in some data types, attributes are

6.3 DATA

165

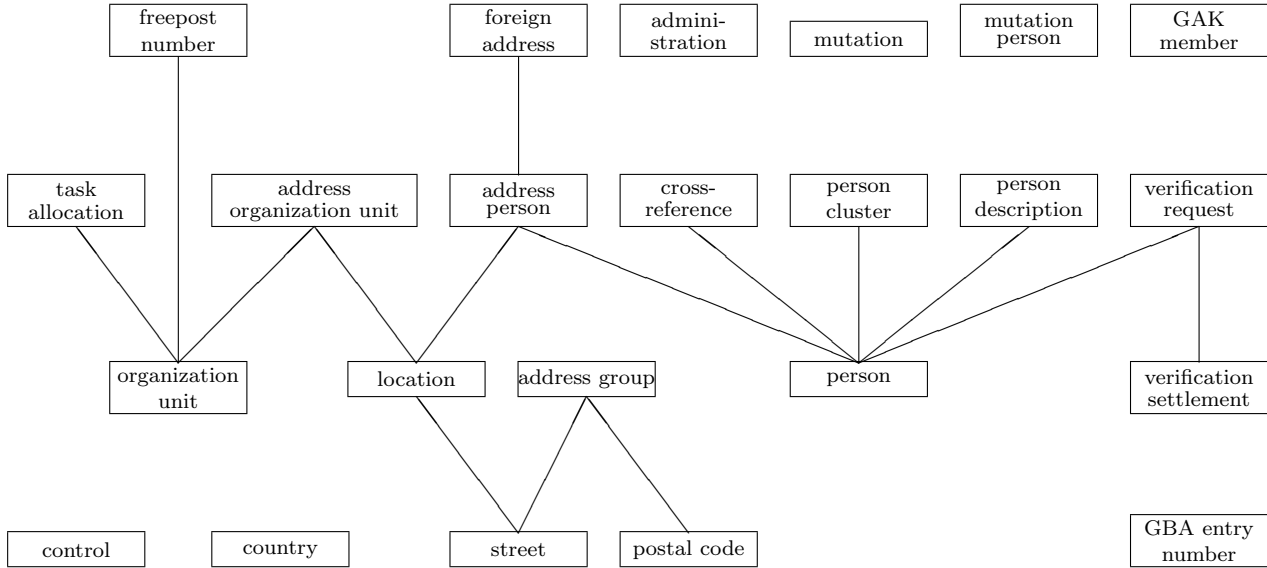


Figure 6.3: Semantic model of BRP

used as foreign keys to primary key attributes of other data types. As the CODASYL database management system uses set-connections to enforce relatability of data, it cannot enforce relatability specified using foreign keys. Additional application software needs to be written to implement the same constraints that could have been specified in the database directly. Special care needs to be taken in case of mutations of the primary key attributes.

Domain integrity The documentation of the BRP contains extensive list of valid codes that can be entered in a particular data field. There are two sets of documentation which contradict each other. The most important difference is the range of SoFi numbers allowed. One document defines a range of 1,000,007 to 299,999,993 for SoFi numbers issued by the treasury and defines the range between 800,000,000 to 899,999,999 for fictitious SoFi numbers. The second document defines a range of 1,000,007 to 199,999,995 for SoFi numbers issued by the treasury and defines the range between 800,000,006 to 899,999,992 for fictitious SoFi numbers. The BRP fulfills neither requirements and preconditions and defines a SoFi number within the range 0,000,000 to 9,999,999. Data management personnel indicated in interviews that additional SoFi number domain integrity rules are probably implemented in application software.

Static integrity An additional requirement on the SoFi number is the modulo 11 test which specifies a calculation that the digits of a valid SoFi number have to satisfy. Of all 19 requirements made on the static integrity of the BRP none are implemented in the data dictionary, but are probably implemented in the application software. The requirements and preconditions and properties of static integrity found during the case study were expressed in the semantic data language defined by Ter Bekke [4].

Insert integrity The three requirements and preconditions of insert integrity that were specified in documentation could not be identified in the data dictionary of the BRP. The data dictionary did however specify thirteen properties of insert integrity that were implemented in the BRP. The requirements and preconditions and properties of insert integrity found during the case study were expressed in the semantic data language defined by Ter Bekke [4].

Modification integrity Requirements and preconditions for the data characteristic modification integrity are not documented. The database management system does have run unit journals that enable the results of an transaction to be undone when an modification constraint is violated by executing a ROLLBACK instruction at the end of a transaction instead of the COMMIT instruction that is executed at the end of a successful transaction.

Securability Requirements and preconditions for the data characteristic securability are not documented. The data dictionary contains one sub-schema that is used by all applications programs, all application programs have access to the complete database. Securability is enforced by allowing users to execute only those transactions

6.3 DATA

167

permitted by the access management information system (TBS). The data dictionary also contains two security schemas which screen the use of data manipulation commands on the database. One security schema allows only retrieval operations on the database, the second security schema allows both retrieval and mutation operations on the database.

6.3.3 Distribution principle

The distribution of the BRP is described in more detail in this section. The need for data distribution for the BRP is discussed first, followed by a discussion of the architecture of the data distribution. The third and fourth sections contain descriptions of the implementation of mutations and retrieval of BRP data. These sections are based on an published article [72].

Necessity for data distribution

Personal data is nearly always required to perform administrative tasks within social security administration. Central registration of personal data enables GAK to guarantee consistency of personal data and the validity of social security benefits. The disadvantage of central storage is that everything depends on having personal data in time to perform administrative tasks throughout the GAK organization. This dependency makes heavy demands on the availability and performance of automation facilities employed for the BRP. Important aspects include the high intensity and national distribution of utilization of data stored in the BRP, together with the large volume of data stored.

The large volume of data is caused by the large number of people in The Netherlands for whom GAK performs social insurance administration. The database contains data on approximately seven million people, a number which is expected to increase due to closer cooperation between social security organizations, both governmental and non-governmental. The amount of personal data is shown in table 6.9, the verification and justification data that form a substantial part of the database (see figure 6.3) are not shown. The additional space required for the NEXT, PRIOR and OWNER pointers which the CODASYL database stores with each record is also not shown.

The intensity of utilization is a consequence of the high dependency of the legislative information systems on personal data stored in the BRP. About one million on-line transactions, both retrieval and mutations, are performed on data stored in the BRP every week. On-line transactions and batch transaction combined lead to three million mutations per month on data stored in the BRP. One must take into consideration that mutations are not confined to modifications in a single record, but bring into play a complex verification and justification process in which the authorization for this modification granted is after verification with municipalities and the ministry of finance and the effects on social security benefit claims have been established.

Database record	Number of records	Storage size
person	6.949.676	500 MB
address-person	10.404.359	650 MB
person-description	9.898.293	1,2 GB
person-cluster	8.493.419	325 MB

Table 6.9: Amount of data (without verification and justification data)

The nationwide distribution of utilization is caused by the strong geographical distribution of the district offices of GAK where legislative information systems are used to evaluate and grant social security benefits. Retrieval and mutations of personal data is performed by functionaries of GAK at the district offices sited at thirty locations throughout The Netherlands. Transport of data between the central service centre (CSC) in Amsterdam and the legislative information systems at these thirty locations would cause a heavy load on the internal communication network of GAK. Legislative information systems are implemented in such a way that additional data of a person concerning that particular social insurance legislation is stored locally. This is a necessity because legislative information systems are implemented using a number of different database management systems and technologies and complex queries (e.g. using joins) can not access BRP data efficiently. The BRP has to transfer all mutations in personal data to the legislative information systems to facilitate this implementation. To this end mutations are sorted each night by district office and legislative information system and placed in update files for each of the thirty locations and different legislative systems. These update files are processed by the legislative information systems to update their own copies of personal data. The head office makes use of information systems implemented on mainframes and Unix mini-computers that constitute a technical infrastructure independent of the technical infrastructure and communication network in use at the district offices. This division of the GAK automation facilities means that each night additional update files have to be made for the head office information systems.

During initial development and implementation of the BRP in 1987, relational database management systems, and distributed database management systems could not handle the large volumes of data transactions and the internal communication network could not provide the high availability required. It was decided to develop a distribution concept based upon CODASYL database management systems and client/server connections, that could provide the functionalities of a central registration of personal data, despite the large volume of the database and intensity and distribution of utilization.

6.3 DATA

169

Distribution architecture

The intensity of utilization and the large volume of the database made it impossible at the time (1987) to implement the BRP using a single computer system. The BRP implementation made use of the fact that the utilization of BRP functionalities could be divided into head office use and district office use. One large computer system was placed at the central service centre on behalf of head office utilization, and three computer systems were placed at the regional service centres on behalf of the district offices.

District office use of personal data is mainly on behalf of the legislative information systems that use the BRP as an authentic source of personal data. The emphasis on this one functionality of the BRP means that access to the database is mainly for retrieval of personal data. This enables the use of a much simplified data model in which all personal data is joined together in one large record for each person, in contrast to the main BRP database where multiple records are used to store personal data, as shown in figure 6.3 and table 6.9. In this simplified data model personal data can only be accessed by someone's Social-Fiscal number (SoFi number), and as the SoFi number is generally known to the legislative information systems this poses no problems. As the district offices are located throughout The Netherlands the extensive communications could form a bottleneck in transaction processing at GAK. To relieve the communication network, databases which use the simplified data model to store the personal data record for all persons registered (they are fully replicated databases) are held at the three regional service centres (RSCs), and for this reason these databases are called the 'integral regional copies' (IRCPs). The simplified data model and simplified utilization of the IRCPs means that it is sufficient to place a small VAX computer at each RSC.

Head office use of personal data is mainly on behalf of the verification and authentication process of personal data, as well as on-line processing and batch processing of mutations from various sources. All functionalities of the BRP are required and therefore the complete data model of the BRP (as shown in figure 6.3) is required, in which personal data is divided into several records and is related to several kinds of verification and authentication data. The complete data model enables automated checks on validity of personal data and mutation in personal data, it also enables access to personal data without use of the SoFi number, but for example, using the combination of last name and date of birth. Mutations are processed by the BRP in two ways: during working hours the mutations from legislative information systems and from GAK employees performing on-line transactions are processed, during the evenings and weekends batch transactions are performed on behalf of GAK and of external BRP stakeholders, such as the tax offices. The complex data model and the complex utilization of BRP data means that it is necessary to place a cluster of four large VAX computers at the CSC.

The large number of records in the BRP database and the complex operations performed cause difficulties in batch processing throughput which can not be solved

with the large processing capabilities of the VAX-cluster alone. To reduce the batch processing times an identical copy of the BRP database (COPY BRP) is placed at the CSC. A second copy of the database is used for batch-retrieval processing only: the Integral Central Copy (ICCP BRP), which has the same simplified data model as the IRCPs. These copies of the BRP database enable the VAX-cluster to run several batch processes at the same time. The various copies of the BRP are shown in figures 3.8 and 6.4.

Location transparency is one of the important bases of the distribution architecture of the BRP. When there is a need for legislative information systems access personal data it is not necessary to specify on which database the transactions are to be performed. The BRP is surrounded by an 'application program interface' which can be used by each of the legislative information systems to perform one of the various predefined transactions that can be performed on the BRP. This interface is called the "Attachment legislative databases on the BRP" (ADB), see also page 96. A number of parameters (such as search criteria) have to be entered when requesting a particular standard transaction, and the ADB-interface is used to return the appropriate results. The ADB-interface is implemented using a product "Application Control and Management System" (ACMS) which enables a transaction to be split into two halves: one half of the transaction is performed on a 'front-end' computer the second half of the transaction is performed on a 'back-end' computer. The ACMS uses a 'mailbox' mechanism to transport data over the network between the two halves of the transaction. In this manner a client/server division is introduced at both the ADB-interface and the BRP.

Implementation of mutations distribution

The distribution of mutations is based on the observation that mutations are relatively few compared to retrieval. Another observation is the fact that the same personal data is hardly ever used by district offices connected to two different regional service centres (RSCs).

The steps involved in mutations of the BRP are shown in figure 6.4. In the first step (1) a mutation request is sent from the legislative information system via a mailbox to the ADB server at the RSC that holds the relevant data. The ADB server in turn consists of a client part in which the mutation request and additional storage for the result are placed in a new mailbox, which is then sent in a second step (2) to the central server at the CSC. The central server reads the mailbox and performs the mutation in the central BRP database. In the third step (3) the central server sets a mutation indicator in the regional database using a remote DBMS transaction; this operation does not have to be location transparent. When both mutations are successful the ADB server returns the result of the mutation to the legislative information system in step four (4). All mutations are processed in this way during working hours within GAK. At the beginning of the evening all mutations that were performed on the central database are distributed to all regional and central copies in step five (5) and

6.3 DATA

171

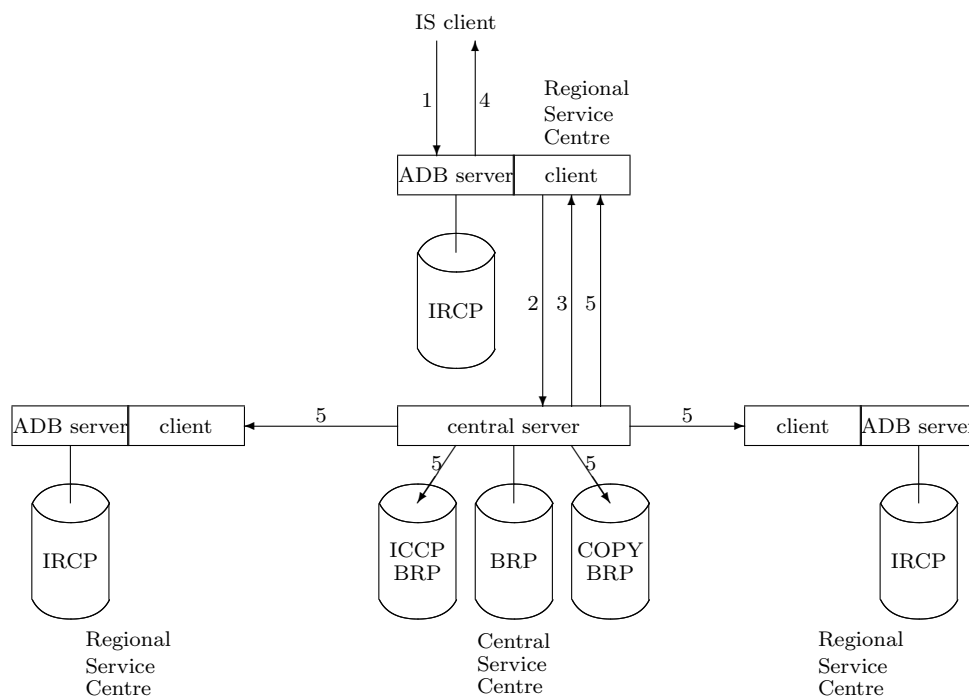


Figure 6.4: Distribution principle

processed. The mutation indicators are reset to indicate that the databases have been re-synchronized.

Distribution of mutations to the legislative information systems is established by creating separate update files for every information system at district offices and the head office of GAK. To distribute the work load, update files on behalf of the district offices are created and distributed by the IRCPs at the regional service centres, update files on behalf of information systems at the head office mainframes and Unix mini computers are created and distributed by the BRP at the central service centre.

This method of distributing mutations *restores* consistency of data stored at different locations instead of maintaining consistency continuously. This entails the risk that personal data stored in legislative systems and head office systems lose their actuality during the course of a day. There is a chance that decisions are based on false data, however, the frequency with which this happens is sufficiently low as to enable GAK to restore such errors manually.

Implementation of retrieval distribution

Retrieval is also based upon the client/server mechanism described above. A legislative information system calls a standard transaction in the ADB server at the regional service centre by sending a mail box with the appropriate data and some extra storage space to receive the result. When personal data is retrieved using a person's Social-Fiscal (SoFi) number, the ADB server checks if the mutation indicator (which is set at mutation step 3) of that particular personal data record is set. If this is not the case, personal data is read from the regional copy (IRCP), and immediately returned to the legislative information system, thus achieving a short response time. When the mutation indicator is set or the personal data is not retrieved using the SoFi number, the client part of the ADB interface places the request in a new mailbox that is sent to the central BRP server. The central server reads the mailbox, performs the standard transaction and returns the result to the client part of the ADB interface. The ADB interface in its turn reads the mailbox and the result is returned to the legislative information system by the server part. This mechanism guarantees that most retrievals can be performed at the level of the regional service centres, and that despite mutations, correct data is still being read.

6.3.4 Syntactic level and empiric level

As the emphasis in this case study was placed on the functional management of the BRP, no extensive investigation of the automation facilities employed on behalf of the BRP by using the methods and techniques defined in chapter 5 was performed. The technical documentation of the BRP and interviews with technical management personnel provided enough information however, to make the observations given in the next two sections.

Syntactic level

There are no requirements and preconditions defined for storage structures and access structures. This can be explained by the fact that no stakeholders are defined outside the systems management (SM) unit. Properties of characteristics can hardly be measured due to the size of the database, verify and analyze jobs can not be run within a single weekend time slot and have to be aborted on Monday mornings before they are completed. Analysis and verification are performed on the COPY-BRP database *not* the central BRP database itself because this is used during the weekend, and analyze and verify jobs require exclusive use of the database. Output of analysis and verification are judged by data management personnel according to their own interpretation of what properties are allowed and when properties are no longer correct. To give acceptable access speeds for the database the access structures have an average load factor of 60 percent. The simplified IRCP BRP and ICCP BRP databases lack the verification and justification data and therefore require less data capacity compared

6.3 DATA

173

Database	Location	Storage Capacity
BRP	CSC Amsterdam	17 GB
COPY BRP	CSC Amsterdam	8,5 GB
ICCP BRP	CSC Amsterdam	5 GB
IRCP BRP	RSC Den Haag	12,2 GB
IRCP BRP	RSC Eindhoven	12,2 GB
IRCP BRP	RSC Hengelo	12,2 GB

Table 6.10: Allocated DASD storage capacity

to the BRP database. The IRCP BRP databases additionally have such low load factors for their access structures that data access is very fast.

Empiric level

The six databases that store personal data together have 67 Gigabytes of DASD storage at their disposal. The total amount of DASD storage capacity allocated to BRP databases is shown in table 6.10. Some databases appear to have much more data storage allocated than others. When this happens two DASD units are connected to each other to form one ‘shadow-disk’ to obtain higher reliability and lower access times. The BRP database is equipped with such ‘shadow disks’ and consequently uses twice the storage capacity of the COPY BRP database. In addition the BRP and COPY BRP databases contain records with verification and justification data next to the database records shown in table 6.9; all the data stored in the BRP database requires approximately five Gigabytes of storage space.

The average load factor of 60 percent of the access structures of the BRP means that the BRP requires almost 8.3 Gigabytes of storage capacity. The simplified IRCP BRP and ICCP BRP databases require less storage capacity compared to the BRP database. The IRCP BRP databases are also equipped with ‘shadow disks’ and therefore have twice the storage capacity requirements compared to the ICCP BRP database.

In addition approximately 500 cartridges of 215 MB each are allocated to BRP backups. Backup tapes are stored in a vault not at the Central Service Centre (CSC) but at the head office of GAK which is also in Amsterdam. Backup tapes of the RSCs are rotated among them.

Requirements and preconditions on the safety of the BRP data are made by the board of directors of GAK. The spaces where BRP equipment is placed (e.g. the VAX computers, DASD and tapes) must not be accessible to unauthorized persons, and the measures taken to prevent loss, mutilation of extended unavailability of equipment or data must be checkable. The computer rooms at the CSC and the regional service centres (RSCs) are secured by a card key system which only permits access to autho-

rized personnel. The CSC building itself is only accessible using of a personal id-card issued to all GAK employees that work at the CSC, GAK employees from outside the information and automation department do not have access to the building.

6.4 Management

Management of the data of the Personal Data Register (BRP) is performed by personnel from the department Information and Automation (I&A) at the regional service centres and at the central service centre, additional data management is performed by a special functionary at the district offices. The tasks which are performed by BRP data management are described in section 6.4.1. The organizational units involved in performing management tasks are described in section 6.4.2. Table 6.11 gives the distribution of responsibilities, authorizations, work, and expertise in the data management organizations.

6.4.1 Management tasks

Interviews with data management personnel provided insights into the way data management tasks are performed by the organizational units involved. More detailed descriptions of the data management tasks can be found in section 4.3.

Managing authorization of data utilization At each of the district offices there is a special subject registers functionary (DK-BR) who is able to use the access management information system (TBS) to grant access authorizations for the BRP to functionaries at his own district office. The TBS maintains a list of the complete organizational structure of GAK, a DK-BR functionary cannot create authorizations for GAK employees outside his own district office.

Managing actual data contents When new data is entered, e.g. a new address, this has to be checked at the municipality involved or at the GBA network by a special verification processes initiated by functional management. The DK-BR functionary is able to modify the BRP directly to correct mistakes that have been made by employees at the district offices.

Ad hoc data provision The functional management unit (AMU-FB) provides extracts of personal data within the preconditions set by the privacy regulations for the BRP.

Establishing and documenting data names and a data naming standard There have been two attempts at implementing this task, at the level functional data management and at the level of application management; both attempts failed however. At present the data management unit is trying to establish some standards for data names, and has started documenting data names.

6.4 MANAGEMENT

175

Managing the application of data definitions This task has not been officially allocated to a specific unit, the data management unit of the development department are in the process of achieving some formal status, but in reality the data definitions are managed by the BRP project group that writes the BRP application software.

Consultancy Occasionally the data management organization is asked to advise on data definitions. The expertise to give recommendations for improvements in the application of data definitions rest with the BRP project group of the systems development department. The IT unit gives recommendations on the technical aspects of databases.

Management of access structures and storage structures An employee of the functional management unit was given the task to run programs that analyze and verify the consistency of the database and characteristics such as access speed, data capacity, load factor, and error detection. These programs require exclusive use of the database, so they have to be performed in the weekend on the COPY BRP database as the BRP database is busy performing batch processing. The functional manager starts the analysis and verification on Friday evening and has to abort the programs on Monday morning before they are completed. The functional manager searches the many megabytes of log file for errors, and as soon one is found, the systems management unit is called and is asked to solve the problem. The functional manager has the ability to start programs that perform an identical analysis and verification on the IRCP BRP databases, but does not have the authorization to read the log files that are placed on the RSC computers by the programs. As a consequence the log files are not inspected. The table for RAW distribution explains this behaviour: the functional manager is given the responsibilities and does half the work, but authorization and the expertise rests with the systems management unit.

Availability of data Backups are made by two organizational units, systems management starts batch jobs that close the database and initiate a backup, media management receives mount prompts at the systems console and mounts the cartridges requested in the designated cartridge drives, upon completion the cartridges are shipped to a vault at the head office of GAK.

Testing and evaluation New databases are tested according to criteria set by the IT unit. The technical documentation of the implementation is inspected to determine if the database satisfies the technical standards set by the department R&I which is the technical management organization at GAK.

Managing the storage of data The performance of the DASD is monitored via remote links with the RSCs and the CSC to the building where the systems management unit is housed. The fill factor of DASD is monitored continuously, when predetermined thresholds are crossed an automatic warning is issued. What actions are taken to free space on the DASD are unknown.

Safe keeping There are two types of media registration. The intake of tapes and diskettes in various formats from different suppliers of data to the BRP, e.g. employers, employment agencies. Registration can be planned in advance. Registration of backup tapes can not be planned because there is no prior communication between the unit media management (MM) and the units system management (SM) and unit account management (AMU) about the tapes required. Experience has taught the media management group when to expect certain backups processes so the tapes can be kept in readiness. The registration of backup tapes is essential for those contingencies in which data processing needs to be resumed at a different location, the emergency set of backup tapes should be readily available and easily identified. Depending on the type of backup between 20 and 40 tapes are required to make a backup of BRP data, the whole backup process takes about two hours.

Reconstruction Reconstruction of data is initiated by the systems management unit that start a restore job, the correct backup tapes are retrieved from the tape vault at the head office and placed in the tape units at the CSC.

Destruction Used tapes are taken away by a garbage incineration company, the DASD units are simply taken away by the supplier Digital (DEC) it is not known if they are erased beforehand.

6.4.2 Management units

The board of directors is the owner of all data contained in the BRP and of all the automation facilities that are deployed on behalf of the BRP, and as such has the competence to determine what functionalities are to be performed by the BRP and what automation facilities are to be deployed. This authority is given to the department Information and Automation (I&A) which acts as the custodian of the BRP. Within I&A, management of the BRP is divided into functional management performed by the department Information Management and Administrative Processes (IMAP), into application management performed by the department Systems Development (SD), and into technical management performed by the department Computer Centres and Infrastructures (R&I).

The department Information Management and Administrative Processes performs functional management of all the information systems and of the subject registers and it consists of a number of units for account management. The unit Account Management and Functional Management (AMU-FB) is responsible for supporting the functionaries at the district offices and the head office and maintains contacts with the stakeholders outside the GAK organization.

The department Systems Development (SD) performs application management of all the information systems. The department consists of a number of holders of portfolios of related applications which are responsible for the development and maintenance of the application software. Within the portfolio 'subject registers' are three project

6.5 CASE RESULTS

177

groups, one for each subject register (SD-BRP). Traditionally the project groups have the expertise to determine functional specifications and technical specifications.

The department Computer Centres and Infrastructures (R&I) performs technical management of the automation facilities and the technical infrastructures present at GAK. The department consists of service teams which perform computer operations and problem management. The service teams conclude agreements with the account managers of IMAP regarding the technical services they provide. The service team on behalf of the BRP consist of a unit for media management (MM) with personnel located at the CSC and the RSCs, and a unit for management of the VAX computer systems management (SM) at the CSC and the RSCs. Personnel of the SM unit is located in a separate building a short distance from the CSC. The department R&I also contains a support team for information technology (IT) matters. The consultants aid the personnel of the SM unit in analyzing the performance problems of the VAX cluster and of the CODASYL database systems employed on behalf of the BRP.

At each of the district offices there is a special subject registers functionary (DK-BR) who uses the access management information system (TBS) to grant access authorizations to the BRP functionaries at the district office. The DK-BR functionary is allowed to access the BRP directly to insert, modify or delete data, this enables the correction of errors made by employees at the district office that would otherwise be too difficult to correct.

The distribution of management tasks with respect to responsibilities, authorizations, work and expertise of each of the management units involved is shown in table 6.11.

A separation of functions is introduced between organizations for technical management (CSC and RSC) and functional management (AMU-FB). The organization for application management however, also performs tasks of functional management, notably in the tasks fields data administration and database administration, this thwarts the intended separation of functions. The prime coordinating mechanism for BRP data management at GAK is mutual adjustment, this is in line with the organizational structure called 'adhocracy' by Mintzberg [73] and confirms the location of the custodian of the BRP as shown in the diagram in figure 6.2.

6.5 Case results

The main objective of this case study was to test the adequacy of the management instrumentation described in chapters 4 and 5 by its application in a practical situation with an analysis of data management of the BRP. The general model of data management was tested by applying the methods and techniques in the specific case situation. Analysis of the general model and what was observed in the specific management situation of the BRP enabled the identification of possible improvements for data management of the BRP and also of possible improvements for making the

Tasks	Units	DK-BR	AMU FB	SD-BRP	IT	SM	MM
Managing authorization of data utilization		RAWE					
Managing actual data contents		RAWE	RAWE				
Ad hoc data provision			RAWE				
Establishing and documenting data names and a data naming standard				RAWE			
Managing the application of data definitions				WE			
Consultancy				RAWE			
Management of access structures and storage structures			RW	WE		AWE	
Availability of data						RAWE	
Testing and evaluation					RAWE		
Managing the storage of data						RAWE	
Safe keeping							RAWE
Reconstruction						RAWE	RAWE
Destruction							RAE

Table 6.11: Table of RAWE distribution BRP

management instrumentation more adequate for use by BRP data management personnel. The fact that the management instrumentation enabled improvements to be identified is an indicator of its applicability in this practical situation. The way of performing the case study and obtaining the case study results is shown in figure 6.5.

In general, using the management instrumentation helped the management of the BRP to identify bottle necks, and provided suggestions for improving the BRP and its management. Use of the management instrumentation helped management of the BRP to increase their understanding of what is expected from the BRP and of the BRP data management. The results of applying the management instrumentation, as published in report [71], found their way to BRP data management personnel who used them to: “finally understand what the BRP really looks like”. The results were also used in an organizational analysis of the technical management organization at GAK.

Conclusions and recommendations with regard to management of the BRP are given in section 6.5.1. Conclusions concerning the applicability of the management instrumentation are given in section 6.5.2.

6.5.1 Conclusions: Personal Data Register

The central question for this case study was whether the management instrumentation was adequate for analyzing management of distributed data in distributed

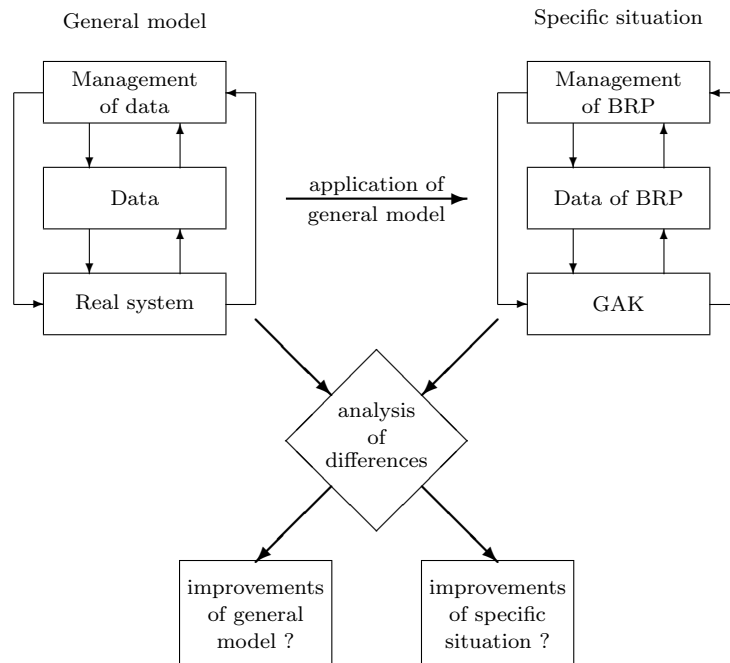


Figure 6.5: Method used to obtain the case study results for the BRP

environments and for identifying bottle necks and providing suggestions for improvement. The conclusions of the BRP case study are described below in relation to the environment, data and management of the BRP.

Environment

There are 27 different kinds of stakeholders within the environment of the BRP, and these are highly distributed. About seven million people and five million different locations are registered within the databases of BRP, plus some 200.000 foreign locations. Data is verified by 700 municipalities and 90 tax offices. Based upon table 6.3 two basic conclusions concerning the environment can be drawn:

1. The three most important stakeholders of the BRP are:
 - (a) The person registered
 - (b) The legislative systems

- (c) The department for statistical research
- 2. The four most important functionalities of the BRP are:
 - (a) Verifying personal data with municipalities
 - (b) Verifying of social fiscal (SoFi) numbers with the treasury department
 - (c) Processing mutations in personal data from legislative systems
 - (d) Source of authenticated data concerning personal data within GAK

Stakeholders are not recognized within the management of the BRP, and no formal relationships are established. The create/use table of page 158 shows that the legislative systems and the department for statistical research have the most authorizations over the actual data contents of BRP. The high importance given to the statistical research department is somewhat of a surprise because it does not perform any primary functions, it supports the GAK board of directors in policy making and strategic management.

The recommendations made in the case study with regard to the environment of the BRP emphasize the importance of close relationships between BRP data management and the stakeholders of the BRP. The stakeholders, the functionalities and the relationships between these two must be established and documented by BRP data management.

Data

Stakeholders are not recognized within the management of the BRP, as a consequence many requirements and preconditions cannot be related to the stakeholder(s) that made these requirements and preconditions. In table 6.7 is shown that at the pragmatic level about 60% of the stakeholders are not related to requirements and preconditions and about 50% of the requirements and preconditions are not related to stakeholders. At the semantic level no relationships could be found at all between stakeholders and requirements and preconditions.

Requirements and preconditions are well defined at the pragmatic level and semantic level of data, however few of the requirements and preconditions are actually met by the BRP. Few properties of the BRP are known, and those properties that are known often are unrelated to requirements and preconditions.

Distribution of data is complicated due to the existence of regional copies (IRCPs) and central copies (COPY-BRP and ICCP). The COPY-BRP database is a complete copy of the central BRP database, the IRCPs and the ICCP are however, different from central BRP database. The IRCPs and the ICCP differ in data model from the BRP and in actual data contents. The regional copies and the central copy only serve as authentic sources of personal data. Despite these differences in properties between databases, the requirements and preconditions made on all the BRP databases are exactly the same.

6.5 CASE RESULTS

181

The recommendations made in the case study with regard to the requirements and preconditions of the data of the BRP emphasize the fact that the relationship between the stakeholders and the requirements and preconditions form the basis of data management. The properties of the six BRP databases must be measured much better than is done now and the properties of all the databases must satisfy the requirements and preconditions.

Management

Tasks within the task field actual data management are well defined in handbooks, other task fields lack formal descriptions and are left to the discretion of individual data management personnel, leading to ad hoc actions in data management. Management of the BRP is not based upon requirements and preconditions of data, and is not related explicitly to tasks and activities performed. The influence of the activities of data management on properties of data within the BRP is therefore unknown.

Functional management of the regional copies (IRCPs) is not possible, due to the fact that the departments involved have no direct access to the regional copies, so they cannot perform checks on the actual data contents of these databases. Technical management of IRCPs is also barely possible, inspection of data at the syntactic level using analyze and verify software that checks the database structures is only performed on the COPY-BRP database. Inspection of the data structures of the BRP is not possible due to the continuous utilization of the BRP databases.

The recommendations made in the case study with regard to the BRP data management organization emphasize the fact that many of the data management tasks are performed on an ad hoc basis. The tasks of all the task fields must be defined in the same manner as in the task field actual data management. These definitions of the tasks should be based upon the requirements and preconditions made by the stakeholders. Correct allocation of responsibilities, authorizations, work and expertise should then be based on the task definitions to ensure that management of both the regional databases and the central databases is possible.

6.5.2 Conclusions: Management Instrumentation

The observations made during the case study show in what respect the management instrumentation is adequate in this specific situation, and what kind of improvements of the management instrumentation are needed for this particular situation. Possibilities and ways to improve the management instrumentation are indicated by answering the two research questions stated in section 1.4.4.

Characteristics of data

Research question 1 (see page 31) asks to what extent the characteristics defined in chapter 4 are of importance for performing tasks of managing distributed data in distributed environments.

The case study did not find new data characteristics that were used by the BRP data management but that were not defined in the conceptual framework. It can therefore be concluded that the data characteristics defined in chapter 4 proved to be of value to the BRP data management.

The FATO framework prescribes that data management should be based on a characterisation of data to operate, control and maintain the data in accordance with requirements and preconditions. A description of how the BRP data management implements these data characteristics in practice is given in section 6.3. The pragmatic and semantic data requirements and preconditions were defined extensively in the functional documentation of the BRP, but were not based on clearly defined data characteristics and were therefore hard to identify by BRP data management and consequently not used very often when performing data management tasks. The syntactic and empiric data requirements and preconditions on automation facilities were not defined in documentation, and were only encountered in interviews with data management personnel and during observation of the data management tasks performed. Although the emphasis in the case study was placed on the functional management of the BRP, technical management were also interviewed to obtain a complete picture of how the BRP data management handles requirements and preconditions in practice. The following paragraphs give more detailed conclusions regarding the data characteristics used by the BRP data management at the pragmatic, semantic, syntactic and empiric data levels.

Pragmatic level All the characteristics defined at the pragmatic data level were encountered in data management of the BRP. The amount and detail of requirements and preconditions is high, but they are defined in BRP management handbooks in an unaccessible manner, making them difficult to identify and use by data management personnel. The large amount requirements and preconditions is caused by the large number of stakeholders of the BRP, and the fact that many stakeholders are outside the GAK organization making clear agreements between BRP data management and stakeholders a necessity. The difficulty of identifying and using requirements and preconditions is demonstrated by the fact that 85% of the requirements and preconditions are not met, and that of 22 known properties 55% are actually properties for which no requirements have been made (see table 6.5). Using the table of stakeholders (table 6.7) it can be shown that BRP data management was unable to identify for 49% of the pragmatic requirements and preconditions which stakeholder had requested, and conversely, could not identify requirements and preconditions belonging to 59% of the stakeholders. Use of the pragmatic data characteristics demonstrated to the BRP data management that they could not manage the data stored in the Integral Regional Copies (IRCPs) of the BRP because the properties were unknown to them. They had to concentrate their attention on the central BRP database without taking into consideration that the IRCPs are completely different databases that store personal data in a different manner (see page 169). The pragmatic data characteristics provided BRP data management with a clearly outlined means to define and man-

6.5 CASE RESULTS

183

age the requirements and preconditions made on the BRP at the level of actual data content.

Semantic level All the characteristics defined at the semantic level of data were encountered in data management of the BRP. The amount and detail of requirements and preconditions is high due to the large amount of different types of data stored in the BRP database, shown in figure 6.3. Contradicting requirements and preconditions were defined in documents from two organizational units. Clear data definitions are a necessity if data exchange with stakeholders outside the GAK organization is to be successful. The requirements and preconditions contain synonyms⁵ and homonyms⁶ which endanger unambiguous utilization of data. The semantic data characteristics were of use in proving that the data management of data definitions was unable to maintain the properties of data definitions (in the data dictionary) in agreement with the requirements and preconditions, 90% of which were not met (see table 6.8). The semantic data characteristics proved in fact that many requirements and preconditions are simply ignored and implemented in completely different manner. The Integral Regional Copies (IRCPs) are implemented without any reference to the requirements and preconditions set by the BRP stakeholders and these are not met by the properties of the IRCPs. The semantic data characteristics proved to be valuable to the BRP data management demonstrating the need for more active involvement in management of the data definitions in the BRP.

Syntactic level Three out of ten characteristics defined at the syntactic level of data were encountered in data management of the BRP. These data characteristics were: access speed, data capacity and load factor. Regular technical analysis and verification of the BRP is performed, however, no requirements and preconditions have been defined that have to be met. The results of the technical analysis and verifications are judged by data management personnel using individual criteria. The main difficulty with performing analysis and verification of the BRP is the fact that the large size of the database makes it impossible to perform proper measurements of the properties of data characteristics; the analysis cannot be completed within the available time slot at the weekend and has to be aborted prematurely. Access speed was measured once by a consultant of the IT support team on an ad hoc basis, but measurements were performed on four records in the database (out of 29 records). Analysis of the database yields the data capacity of the database and the load factor of all the database areas in which the records are stored, verification of the database yields any possible errors in the database (e.g. missing pointers in sets). The reports generated by the verification utility and the transaction log files could be used to specify the error correction and error detection characteristics, however, this has not been done by data management personnel.

⁵synonym: a different name for the same data definition

⁶homonym: the same name for different data definitions

Empiric level Three out of eleven characteristics defined at the empiric level of data were encountered in data management of the BRP. These three data characteristics were: storage capacity, security and storage costs. The security requirements and preconditions are determined by the GAK board of directors. Storage capacity and the storage costs are specified in the service contract between the functional management department (IMAP) and the technical management department (R&I) of the BRP. Storage capacity is shown in table 6.10, costs are not given here for reasons of confidentiality, but the costs of storage include the costs of all data management activities performed by technical management (e.g. making backups etc.), not just the cost of the storage media.

Methods and techniques

Research question 2 (see page 32) asks to what extent the methods and techniques defined in chapter 5 are of use to the management of distributed data in distributed environments.

The case study found that eight out of eleven methods and techniques could be used successfully. The table of stakeholders could be used with partial success. The two methods and techniques that could not be used were the automation facilities matrix, and the activity model. During the case study it was found that the entity-relationship model was used to specify semantic requirements and preconditions, and that the network model was used (by the CODASYL data dictionary) to specify semantic properties of the BRP. The semantic model was used successfully in the case study to compare these two different kinds of data models. Other methods and techniques for data management, besides the E-R model, were not found, so it can be concluded that the methods and techniques defined in chapter 5 proved to be of value to the BRP data management organization.

The measures taken by data management need to be supported by methods and techniques to assure that once data are implemented, the data keeps satisfying the requirements and preconditions. The data management of the BRP does not use predetermined methods or techniques of its own, but uses the documentation it receives from the systems development department. This documentation is, however, inadequate for performing data management and is therefore seldom consulted when performing data management tasks. This means that data management personnel have to make their own interpretations of requirements and preconditions that have to be met by the BRP, and each makes their own form of procedure to meet them. During the case study use was made of available material supplemented by investigations into the properties of the BRP and additional interviews with data management personnel. More detailed conclusions regarding the role of the methods and techniques in this case study are given below.

List of stakeholders The list of stakeholders enabled the joint identification of all the stakeholders mentioned in the various documentation of the BRP. The list shown in table 6.1 gave the BRP data management a complete overview of all the

6.5 CASE RESULTS

185

stakeholders for the first time. The information systems and organizations mentioned in table 6.1 are both within the GAK organization and outside the GAK organization. By the vary nature of GAK, as an administration office for social security, most of the stakeholders of the BRP are outside the organization. Maintaining close relations with all the stakeholders mentioned in table 6.1 is a major undertaking for the functional management of the BRP considering their nationwide distribution. With regards to the information systems that are seen as stakeholders of the BRP, relationships should be established between the BRP functional management and the respective information systems management departments.

List of functionalities The list of functionalities enabled joint identification of all the functionalities mentioned in the various documentation of the BRP. The list shown in table 6.2 gave the BRP data management a complete overview of all the functionalities for the first time.

Functionality matrix The functionality matrix was used to demonstrate what functionalities are directed towards which stakeholders. The matrix shown in table 6.3 shows that functionalities are not equally distributed over the stakeholders. During the case study the functionality matrix was used to identify, for the BRP data management, which functionalities and stakeholders are the most important and which stakeholders are the least important. This gave the BRP data management a rationale for prioritizing its data management activities for the first time, and it gave an indication of the functional distribution of BRP data.

Create/use matrix The create/use matrix shows which stakeholders have the right to determine actual values of BRP data, and which stakeholders have the right to use the actual values of the BRP data. The create/use matrix was constructed in this case study by examining the documentation and interviews with data management personnel. The create/use matrix enabled identification of the stakeholders with most rights to the BRP data, and it enabled the identification of data utilized by the most stakeholders. This gave the BRP data management a concise view of all the influences on BRP data for the first time, these are important factors for the maintenance of confidentiality and the maintenance of the data definitions of the BRP. The create/use matrix also gave a second indication of the functional distribution of BRP data.

Semantic model Semantics models can be used to determine complex semantic relationships between data and display these relationships graphically. During this case study the semantic model was used to compare two different kinds of data models; the two entity-relationships models used to specify semantics requirements and preconditions, and the network model used by the CODASYL data dictionary to provide the semantic properties. The semantic model enables the identification of inconsistencies and contradictions in the requirements and preconditions. It was proved, using the semantic model, that the network model was not used properly and that many of the semantic properties that could be enforced by the data model were in fact enforced by application software. This means that the data management needs to maintain

a close relationship with the application management department to ensure that the application software implements the requirements and preconditions made on data correctly.

Table of properties A table of properties can be used to compare the actual properties of data with requirements and preconditions and to identify deviations, then using these deviations as a basis, data management can take appropriate measures. The emphasis in this case study was placed on the functional aspects of data management, and therefore property tables were drawn up only for the pragmatic and semantic characteristics of the BRP data. The property tables proved to be valuable for identifying deviations and comparing requirements and preconditions with actual properties. The properties of a CODASYL database management system can be described very well and utilized within data management using property tables, but conversion to the semantic data model simplifies greatly the identification of properties. Due to the large amount of detail in the specifications of some of the requirements and properties of the BRP data, the property tables can become very large.

Table of stakeholders A table of stakeholders shows the relationships between requirements and preconditions, in terms of data characteristics, the stakeholders that make the requirements and preconditions, and the data management tasks that ensure that these requirements and preconditions are met. During this case study, the table of stakeholders could be used partially to identify the relationships between pragmatic requirements and preconditions and the stakeholders of the BRP. The table of stakeholders shown in table 6.7 was useful as it demonstrated to the BRP data management that they should improve on their relationships with their customers [71]. No relationships could be found for the semantic requirements and preconditions, and it was not possible to draw up a table of stakeholders for semantic characteristics. As the activities of the BRP data management are unknown the relationships between activities and data characteristics could not be established. Without knowledge of these relationships data management is not able to determine the consequences of changes in the data environment, changes in stakeholders or the requirements and preconditions they impose, and the effect this has on the tasks performed by data management.

Automation facilities matrix An automation facilities matrix could not be used in this case study. An automation facilities matrix shows the physical distribution of automation facilities over several locations. A general distribution of the BRP data storage is given in table 6.10. The emphasis in this case study was placed on functional management, detailed technical configuration of the BRP automation facilities could not be established as the documentation was not available. Knowledge of the technical configuration is required to determine the properties of the empirical data characteristics such as reliability, repairability and recoverability.

Activity model Activity models could not be used in this case study. Tasks performed by the BRP data management are not laid down in procedures or other de-

6.5 CASE RESULTS

187

scriptions, this results in the handing over of tasks to other persons becoming even more difficult. The BRP data management documentation does provide instructions for functional management personnel to perform some of its tasks, but these are not specified using a particular method or technique. Activity models were applied successfully in the WWO case study and would be successful for the BRP data management. It should be possible to draw up activity models from the BRP documentation, this would greatly enhance the ability of the BRP data management to perform its tasks in an uniform and controllable way.

Table of RAWWE distribution The table of RAWWE distribution describes which organizational units have the responsibility (R), authorization (A), work (W) and expertise (E) to perform tasks in the BRP data management. The distribution of RAWWE shown in table 6.11 made it possible to identify the way tasks are performed by the different organizational units present at GAK and where improvements could be made. For example, it showed that a functionary at the functional management unit (AMU) was responsible for tasks which according to the FATO framework should be performed by a functionary in the technical management unit. This explained why the functionary lacked the proper authorizations (A) and expertise (E) to perform the tasks, and consequently could not perform (W) them properly. This functionary was transferred swiftly by GAK managers to the systems management unit (SM), given the right authorizations to access the databases and was soon taught by colleagues how to do the tasks.

Coordinating mechanisms Coordinating mechanisms describe how the different organizational units involved in the BRP data management, coordinate their tasks. Many tasks are performed on a daily basis, and functional and technical management of the BRP is performed by large organization with a great number of people. Separation of tasks is carried through a to great extent in the BRP data management organization, coordination is therefore imperative. The functional management unit (AMU) has documented its procedures in an extensive way achieving some form of standardization of work processes. The prime coordinating mechanism between the various units of the BRP data management organization is, however, mutual adjustment. The data management units contact each other on an ad hoc basis, for example when a functional manager encounters an error in a batch job he phones the systems development department to 'fix' this batch program. The prime coordinating mechanism of mutual adjustment is in accordance with the position of the custodian of the BRP in the support staff of GAK, as shown in the diagram in figure 6.2.

Chapter 7

Unemployment Benefit Case

The case study described in this chapter tests the framework and the methods and techniques in management of distributed legislative data, see figure 3.7, of a specific process administration: the information system unemployment benefit¹ (WWO). In chapter 6 the management instrumentation provided by chapters 4 and 5 in management of a distributed data infrastructure was tested in a case study. The two case studies, in combination, were used to test the management instrumentation in the two different approaches to data, data oriented and process oriented (see section 2.1.1), both present at GAK.

In this case study the framework and the methods and techniques in management of distributed legislative data were applied at the Dutch Joint Office for Social Security Administration² (GAK). One aspect of process data is the fact that the data is used to control one specific corporate process. This case study considered control of the execution of legislation concerning social security for people who become unemployed after being made redundant by their employer. Seasonal unemployment is an example of unemployed that is not dealt with by this information system. The second aspect of this case study was an examination of the manner in which data management tasks, performed on behalf of another information system, are suitable for data management of the WWO information system. The WWO data was to be managed by an organizational unit that already performed data management tasks for the information systems on behalf of the workers occupational disability act³ (WAO).

The main objective of this case study was to test the adequacy of the management instrumentation by its application in an analysis of the data management of the WAO and the manner in which it is applicable for to controlling, operating and maintaining the WWO data in accordance with requirements and preconditions. The case study provided suggestions for improvement of the WAO data management tasks to enable

¹In Dutch: Werkloosheidswet Ontslag (WWO)

²In Dutch: Gemeenschappelijk Administratie Kantoor (GAK)

³In Dutch: Wet op de Arbeidsongeschiktheid (WAO)

the implementation of the WWO data management. A full description of this case study can be found in the master's thesis of Arjan Groen [41].

This chapter is fashioned in the same manner as chapter 6. Section 7.1 positions the case study with respect to deconcentration of WWO data and decentralization of the WWO environment to enable cross case analysis in chapter 9, where the final research results are presented. Section 7.2 gives descriptions of the environment of the WWO and the methods and techniques that were used in this case study. Section 7.3 describes the data characteristics of the WWO and the methods and techniques that were used in this case study. Section 7.4 describes management of the WWO and the tasks and organizational units involved currently in performing data management on behalf of the WAO information system, and those that are intended to be involved in performing WWO data management tasks in the future. Section 7.5 gives the results of the application of the management instrumentation in this case and concludes in what way the management instrumentation is adequate for management of this specific process data and what adaptations are required in this specific situation.

7.1 Positioning the case

Positioning of the case study was achieved by following the guidelines given in section 1.4.7. Within the context of this thesis, distribution of the Unemployment Benefit (WWO) data is restricted to the geographical deconcentration of functionalities, automation facilities and the organizational management units, using the diagram for distribution of data shown in figure 4.4 at the end of chapter 4. The 'FAO diagram' for the WWO is shown in figure 7.1. Distribution of the environment is restricted within this thesis to the decentralization of the decisional powers over the WWO within the Joint Office for Social Security Administration (GAK) using the organizational diagram of Mintzberg [73]. The Mintzberg diagram for the WWO is shown in figure 7.2.

7.1.1 Deconcentration

Deconcentration of data of the WWO information system is described using the FATO framework with a diagram for graphical representation where functional distribution (F), the technical distribution (A) and organizational distribution (O) of data is shown. See also section 4.5.

Distribution of functionalities of WWO data is shown by the placement of 'F' segments in figure 7.1. There are thirty district offices within GAK, distributed throughout The Netherlands, where administrative activities on behalf of unemployment benefits are performed. There are three clusters of ten district offices represented by three clusters of ten 'F' segments at the left hand side of the 'FAO diagram', each cluster is linked to one of three regional service centres (RSCs). The head office of GAK uses the WWO information system for administration of unemployment legislation at the national level, this is represented by a single 'F' segment at the top right

7.1 POSITIONING THE CASE

191

hand side of the 'FAO diagram'. The WWO information system will supply mutations in personal data to the BRP system once it is fully operational, these mutations in personal data are supplied to the central service centre (CSC) and there is one 'F' segment at the CSC at the centre right hand side of the 'FAO diagram'.

Distribution of automation facilities of the WWO is shown by the placement of 'A' segments in figure 7.1. At each district office there is a 'front-end' computer with limited storage for temporary files, and ten of these 'front-end' computers are connected to one 'back-end' computer at a regional service centre. The thirty 'front-end' computers are represented by three groups of ten 'A' segments placed at the left hand side of the 'FAO diagram'. These front-end computers will not be involved in the case because the temporary files are not the subject of data management. Three regional service centres (RSC) in The Hague, Eindhoven and Hengelo, contain ten databases, one for every district office 'front-end' computer connected to the 'back-end' computer at the RSC, this is represented by three 'A' segments at the centre of the 'FAO diagram'. This means that each district office uses its own database, and data is not shared between district offices. The database are placed at regional service centres to simplify management and to make efficient use of the three 'back-end' computers. A database at the central service centre (CSC) in Amsterdam contains excerpts from the thirty 'district' databases and is used by the head office of GAK. This central database is represented by an 'A' segment at the CSC at the centre right hand side of the 'FAO diagram'.

Distribution of the management organization of the WWO is shown by the placement of 'O' segments in figure 7.1. Functional management of WWO data is performed by the account management unit (AMU) that is placed at the central service centre (CSC) in Amsterdam. The unit for technical management of the central database is also placed at the CSC. These two units are represented by an 'O' segment at the CSC at the right hand side of the 'FAO diagram'. Functional management is also performed by the project group WWO and the project group data management located at a short distance from the CSC. This separate location is represented by an 'O' segment placed below the CSC at the right hand side of the 'FAO diagram'. Technical management of thirty WWO databases is performed by organizational units located at the three RSCs, and they are represented by three 'O' segments placed at the three RSCs at the centre of the 'FAO diagram'.

7.1.2 Decentralization

Decentralization of the environment of Unemployment Benefit (WWO) data is described using the distribution of decisional powers over the user organization in which data management is performed. The data environment consists of stakeholders with a vested interest in the functionalities performed by the data. The power to decide over the functionalities and automation facilities of the WWO rests with the custodian of the WWO, see section 2.1.3. The level of decentralization of the WWO is determined by the location within the GAK organization where the power rests to make decision

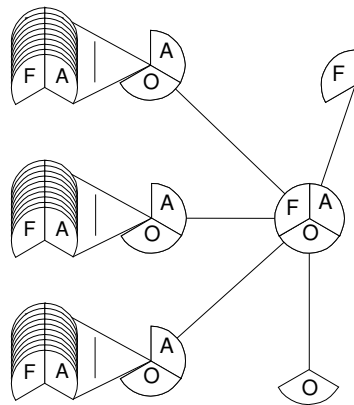


Figure 7.1: Deconcentration of WWO

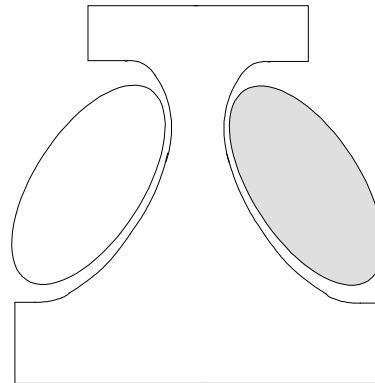


Figure 7.2: Decentralization of WWO

about data in the WWO.

The fact that the WWO is a process administration for use by the district offices of GAK (see figure 3.7) could lead to the expectation that divisional middle line management of GAK has a major influence in the decision making process around the WWO. This is not the case, custodianship of the WWO rests with the department Information management and Automation (I&A) of GAK which is a support unit for the execution of social security legislation at the operating core of GAK. The department I&A decides on the functionalities to be performed by the WWO as well as the automation facilities to be employed. The level of decentralization within the environment of the WWO is thereby characterized as a type D decentralization (see page 34), the Mintzberg diagram shown in figure 7.2 illustrates this type of decentralization.

7.2 Environment

The stakeholders and the functionalities of the WWO identified during the case study using the methods and techniques of the management instrumentation are discussed in section 7.2.1.

7.2 ENVIRONMENT

193

7.2.1 Stakeholders and functionalities

The stakeholders of the Unemployment Benefit (WWO) data are discussed in the first section. The stakeholders were identified during the case study using the list of stakeholders shown in table 7.1. The functionalities the WWO performs on behalf of the stakeholders are discussed in the second section. The functionalities were identified during the case study using the list of functionalities shown in table 7.2. The functional distribution of the WWO is discussed in the third section. The distribution the WWO functionalities over the stakeholders is identified using the functionality matrix shown in table 7.3.

Stakeholders

Stakeholders have an interest in the functionalities that are performed on their behalf by the WWO. There is no distinction between stakeholders that are organizational units or individuals within the GAK organization or outside of GAK, or stakeholders that are automated information systems that depend on the WWO in some form or other. Table 7.1 gives the stakeholders of WWO that were discovered using an analysis of the WWO documentation. There is no separate recognition of stakeholders in the documentation which only speaks of ‘users’ and further details of the stakeholders are lacking.

The board of directors of GAK are the owners of the WWO, and as such, have drafted up privacy regulations concerning the collection and distribution of personal data with regard to persons who are registered by GAK on behalf of social security administrations.

The three subject registers, the personal data register (BRP), the labour contracts register (BRDV) and the employers register (BRWG) supply personal data, data concerning labour contracts and data concerning employers to the WWO. Data that can be retrieved from the subject registers is allowed to be stored in the WWO.

The statistical research department receives statistical data concerning the performance of the processes for unemployment benefit administrations, e.g. the number of unresolved benefit applications or payments. The WWO information system is used by the functionaries at the GAK district offices and at the GAK head office to settle unemployment benefit claims. The WWO registers and archives the unemployment benefit applications made by a registered person.

Vacation funds receive an account of the the vacation coupons issued by GAK to persons receiving unemployment benefits. The pension funds are informed of the pension premium contributions and the pension statements made by the WWO information system. Some industries have separate unemployment regulations (e.g. fishing industry and inland waterways shipping) for which separate social funds are established which give reduced pay to unemployed persons instead of benefits. The WWO information system settles the reduced pay on behalf of the social funds.

The access management information system (TBS) controls and monitors access to the GAK technical infrastructure by employees of GAK and employees of other

B1	Board of directors GAK
B2	Personal data register
B3	Labour contracts register
B4	Employers register
B5	Statistical research department
B6	Functionaries at district offices
B7	Functionaries at head office
B8	Registered person
B9	Vacation funds
B10	Pension funds
B11	Social funds
B12	Access management information system
B13	Health insurance funds
B14	Employers
B15	Benefit payments information system
B16	Head office accounting
B17	Wage tax information system
B18	Insured persons administration
B19	District offices accounting
B20	Pension payments information system

Table 7.1: List of stakeholders WWO

social security agencies. The TBS controls, for each individual employee, access to specific computers, to specific information systems and the transactions employees are allowed to perform. Additional information on TBS can be found in [113].

The employers provide data for the WWO information system. Each month tapes with payment instructions are sent to the benefit payments information system (BETSY). The central accounts department at the GAK head office receives financial justification data directly from the WWO information system. The tax on wages information system (LBP) receives data from the WWO which are used to send annual statements to the tax offices. The insured persons administration consists of all the legislative information systems and subject registers present at GAK (see section 3.2.2), as the subject registers are already mentioned in the WWO documentation this is partially redundant.

The district office accounting departments receives justification data concerning the execution of payments, they return unsettled payments to the WWO information system. The pensions administration department (PUO) receives notification from the WWO when a person is registered by WWO and if pension premiums are paid in accordance with the unemployment benefit act.

7.2 ENVIRONMENT

195

F1	Person based registration of benefit rights, not traditional case based registration.
F2	Automated support of handling benefit applications.
F3	Monitor progress of benefit claim settlement.
F4	Cluster WWO capabilities logically on behalf of the functionaries.
F5	Facilitate time consuming tasks with regard to pension premiums payments and application and sign off to health insurance funds.
F6	Determine supplements to unemployment benefits.
F7	Automate determination of benefits and support the functionary in determining benefits.
F8	Calculate deductions in benefits when multiple social security legislations apply or when other partial benefits rights exist.
F9	Reduce double requests for data concerning personal, employers and labour contracts by being connected to the registers in the data infrastructure of GAK.
F10	Adjusting access rights for each district office.
F11	Reduce costs of handling unemployment benefit applications.

Table 7.2: List of functionalities WWO

Functionalities

The WWO documentation at GAK mentions the functionalities to be performed by the WWO information system separately, the functionalities given in the documentation were used to create the list of functionalities shown in table 7.2.

The WWO maintains a person oriented registration of unemployment benefit claims instead of the previously used case oriented registration. All unemployment benefit applications are registered and archived for each person. The WWO provides automated support of handling of unemployment benefits from the initial application by a person throughout the period of unemployment until the persons unemployment ends. Progress of the settlement of unemployment benefit applications is monitored automatically by the WWO. The various capabilities of the WWO are clustered in logical units on behalf of functionaries.

The WWO has to facilitate time consuming tasks with regard to pension premiums payments and application and sign off to the health insurance funds of the unemployed persons. Supplements to unemployment benefits are awarded by the WWO according to the various additional regulations issued by the different occupational associations.

The WWO automates the determination of benefits and provides support to the functionary in determining benefits. The WWO determines automatically the amount of unemployment benefit, bonuses, and additions according the specific regulations issued by the occupational associations. Settlement of advances on benefits or bonuses are supported, and rectifications of the benefits are calculated and spotted automatically. Determinations of gross and net benefits are performed in a single process. The functionaries have the capability to perform trial calculations. The WWO supports the calculation of the gross amounts of reclaiming of surplus benefits paid in the pre-

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20
F1								X												
F2																				
F3	X																			
F4							X	X												
F5							X	X												
F6																				
F7							X	X												
F8							X	X												
F9																				
F10																				
F11	X																			

Table 7.3: Functionality matrix WWO

vious fiscal year. Improvements in the calculations of deductions due to corrections prevents unwanted differences in benefits as a result of these deductions.

The WWO makes it possible to calculate deductions in benefits when multiple social security legislations apply or when other partial benefit rights exist. The WWO achieves a reduction in double requests for data concerning personal, employers and labour contracts by being connected to the subject registers in the data infrastructure of GAK. Access rights to WWO data and transactions can be adjusted for each district office. The final functionality of the WWO is to achieve a reduction in the cost of handling unemployment benefit applications.

Functional distribution

The functionalities performed by the WWO on behalf of particular stakeholders, indicate the importance of the WWO to the stakeholders and consequently the amount of support the stakeholder needs to receive from the functional data management organization. The documentation present at GAK did not indicate what functionality was built into the WWO on the request of which stakeholder. The distribution of WWO functionalities over the stakeholders is shown in the functionality matrix given in table 7.3. The functionality matrix is based on the documentation, this is however, highly inadequate with regard to relating stakeholders and functionalities. Functionalities are attributed to only four out of twenty stakeholders and four out of ten functionalities are unnecessary as none of the stakeholders require them.

Due to the fact that no rights have been defined, it was not possible to define a create/use table.

7.3 Data

The data characteristics, as defined in chapter 4, are used in the sections below to describe the requirements and preconditions of the WWO data. This shows how in practice the (future) data management organization of the WWO will implement the data characteristics. In this case study emphasis was placed upon functional aspects and technical aspects of data management of the WWO, and consequently conformity of properties of the data characteristics to the requirements and preconditions included data characteristics at the pragmatic, semantic, syntactic and empiric data levels. Data characteristics of the WWO at the pragmatic level are used in section 7.3.1 to structure the requirements and preconditions. Data characteristics of the WWO at the semantic level are used in section 7.3.2 to structure the requirements and preconditions. Data characteristics of the WWO at the syntactic level are used in section 7.3.3 to structure the requirements and preconditions. Data characteristics of the WWO at the empiric level are used in section 7.3.4 to structure the requirements and preconditions. When characteristics are missing from the sections below, they could not be found during the case study of WWO data management.

7.3.1 Pragmatic level

At the pragmatic level of data requirements and preconditions concerning the actual data values are attributed to seven data characteristics: correctness, completeness, actuality, timeliness, durability, confidentiality and auditability. See section 4.1.1 for more details.

The requirements and preconditions are defined using all of the data characteristics. A summary of the pragmatic data characteristics is shown in table 7.4, the number of requirements that are made on the pragmatic data characteristics is shown in the second column, the number of requirements that are fulfilled by properties of the WWO is shown in the third column, the number of additional properties of the WWO upon which no requirements are made is shown in the last column.

Correctness Checks on entered data should be performed as soon as possible, and signalled mistakes should be corrected as quickly as possible. Checks are performed using screen validations and field validations on the data entry screens of the WWO which send an alarm signal to the end-user as soon as invalid data is entered. Screen definitions are however, not the subject of data management but of application management. Consistency of data stored in the WWO is checked before any payment orders are issued, and contradicting data is avoided by using relationship checks and validity checks. The personal data stored by the WWO is verified with the authentic source, and are corrected on the request of the person registered. The WWO is required to store only the SoFi number of a registered person and use this number to access personal data stored in the BRP (see page 172). The WWO however stores, additionally, a person's name, native language, date when the person last changed address, and the person's telephone number. The WWO complies with the requirement

characteristic	requirements and preconditions	fulfilled by properties	additional properties
correctness	16	12	0
completeness	7	5	0
actuality	5	4	0
on-line timeliness	4	0	0
batch timeliness	6	0	0
durability	1	1	0
confidentiality	2	2	0
auditability	2	2	0
total	43	25	0

Table 7.4: Summary pragmatic characteristics

that employer data stored in the WWO is limited to the employer's identification number (for access of the BRWG database), the employer's place of business and the contract number (of the administration contract between GAK and the employer). The requirement that the WWO should only store the SoFi number for access to the labour contracts database (BRDV) is also met.

Completeness A special transaction tests, at the end of each day, if data processing of data stored in the WWO was performed correctly and if all the data processing was completed correctly. It is not known if the data supplied to the head office central accounting department is a complete overview of all financial transactions performed by the WWO, as required. Nor is it known how missing data is discovered and in what manner it is supplemented.

Actuality It is possible at any moment to obtain the most recent amounts of benefits, both gross amounts and net amounts, that have been transferred to an unemployed person. The district offices functionary has access to the most recent status of an unemployment benefit application, and to the most recent status of reclaiming of too much paid benefits. Actuality is influenced by the amount of updates that can be performed, if more updates occur within the specified period the database will start to lag behind the real situation. The documentation specifies the number of transactions that can be performed by a district office on particular relational database tables, as the WWO system was not fully operational, actual performance numbers were not available and there were no special measures built into the WWO to monitor the performance of these transactions.

Timeliness The number of transaction that can be performed as mentioned for the actuality characteristic, are the preconditions under which the on-line response times and batch turnaround times are achieved. The maximum response time is required

7.3 DATA

199

to be ten seconds, whereas 90% of the on-line transactions must be completed within five second. The WWO should be available during working hours, between 7.30 and 18.00, to the on-line users. The performance tests made use of one fifth of the number of transactions and the results extrapolated, the response times during use under real conditions are unknown.

The documentation describes detailed requirements and preconditions for the turnaround times of the WWO batch processes. The annual batch processes to complete the annual accounts of unemployment benefits should be completed within two months after end of the year. The quarterly batch processes should be completed within one month after the end of the quarter. Monthly batch processes should be completed before the tenth of the next month. The weekly batch processes should be completed in 95% of the cases before the next Monday, the remaining 5% of the weekly batch processes may be delayed for at most one week. The daily batch processes should be completed in 95% of the cases before 7.30 the following morning, the remaining 5% of the daily batch processes may be delayed for one day. In reality, the district offices are free to determine when batch processes are started, without carefully determined scheduling it is not possible to guarantee that the batch processes are completed within the required period of time.

Durability Historical data are moved to another storage space of the WWO after six months when the unemployment benefit ends, where they are kept for two and half years before being removed.

Confidentiality The WWO databases are not allowed to contain personal data, but in actuality a person's name, native language, date when the person last changed address, and the person's telephone number are stored. The WWO is secured against unauthorized use and against loss or mutilation of data by the connections with the access management information system (TBS). The TBS limits access by GAK employees by using a user identification combined with password, and prevents access by unauthorized hardware and software.

Auditability The access authorizations can be adjusted by the district offices but should remain in compliance with the diversity of functions that can be accommodated by a functionary, and with the separation of functions and internal control requirements.

7.3.2 Semantic level

At the semantic data level, requirements and preconditions concerning data definitions are attributed to seven data characteristics: convertibility, relatibility, domain integrity, static integrity, insert integrity, modification integrity and securability. See section 4.1.2 for more details.

The requirements and preconditions are defined using all of the data characteristics. A summary of the semantic data characteristics is shown in table 7.5, the

characteristic	requirements and preconditions	fulfilled by properties	additional properties
convertibility	37	11	36
relatability	0	0	1
domain integrity	0	0	1
static integrity	0	0	0
insert integrity	0	0	1
modification integrity	0	0	1
securability	0	0	4
total	37	11	44

Table 7.5: Summary semantic characteristics

number of requirements that are made on the semantic data characteristics is shown in the second column, the number of requirements that are fulfilled by properties of the WWO is shown in the third column, the number of additional properties of the WWO upon which no requirements are made is shown in the last column.

Convertibility The requirements and preconditions of data at the semantic level are documented using a relational data model of the case tool Excelerator, and consists of 750 attributes and 100 compound types. The implementation of the relational database differs due to the fact that additional tables are defined, and within tables the number of columns and the definitions of columns differ from the specifications. The convertibility of the WWO is implemented by defining primary key constraints for each table, thus 100 constraints in total. Besides the thirty databases on behalf of the district offices a GAK head office database is also implemented. There are no separates requirements or preconditions made on the head office database, but the data model is completely different.

Relatability Requirements and preconditions on the characteristic relatability of data are not documented by the case tool. The relational database can implement relatability by using referential constraints on database tables. The referential constraints however, do not use existing indexes to enforce integrity but use lengthy sequential scans of database areas. Relatability was implemented using four SQL triggers for each referential constraint, in total 700 different triggers which required about 1560 rules of SQL code.

Domain integrity Requirements and preconditions on the data characteristic domain integrity are not documented. The Social Fiscal (SoFi) number is defined as an integer in the database, it should comply with the modulo-11 test that prevents errors in entry of SoFi numbers by refusing SoFi numbers that fail the modulo 11 test.

7.3 DATA

201

The modulo 11 test is implemented in application software and does not constitute a database property to be maintained by WWO data management.

Insert integrity Requirements and preconditions on the data characteristic insert integrity are not documented. New relational table entries in the database are stored with default values. Character fields are stored by default with a space character value, numerical fields are stored by default with a zero value. This prevents the storage of null values in the database which can cause errors in the evaluation of SQL queries.

Modification integrity Requirements and preconditions on the data characteristic modification integrity are not documented. Transactions use COMMIT and ROLL-BACK instructions to ensure that only valid updates of the database take place, this is however part of application software and does not constitute a database property to be maintained by WWO data management.

Securability Requirements and preconditions on the data characteristic securability are not documented. There are four views defined for the WWO database. These views combine district data with head office data, it is not known which stakeholder is allowed access to the views.

7.3.3 Syntactic level

At the syntactic data level, requirements and preconditions concerning data access structures and data storage structures are attributed to ten data characteristics: access speed, restructuring speed, data capacity, load factor, use factor, transformation, accuracy, concurrency, error detection, error correction. See section 4.2.1 for more details.

The requirements and preconditions are defined using all of the data characteristics. A summary of the syntactic data characteristics is shown in table 7.6, the number of requirements that is made on the syntactic data characteristics is shown in the second column, the number of requirements that are fulfilled by properties of the WWO is shown in the third column, the number of additional properties of the WWO upon which no requirements are made is shown in the last column.

Access speed Requirements and preconditions on the data characteristic access speed are not documented. To reduce fragmentation of the data storage, which slows down data access, null values are avoided by entering default values (see insert integrity).

Data capacity The documentation contains detailed calculations of the expected data capacity of the (thirty) databases. The thirty district offices are classified into four categories: small, medium, large and very large. For each category the present number of unemployment benefit cases is used to calculate the number of cases that needs to be stored in the databases during a period of three years (see durability). For

characteristic	requirements and preconditions	fulfilled by properties	additional properties
access speed	0	0	1
restructuring speed	0	0	0
data capacity	1	0	0
load factor	0	0	0
use factor	0	0	0
transformation	2	2	0
accuracy	0	0	1
concurrency	0	0	0
error detection	0	0	0
error correction	0	0	1
total	3	2	3

Table 7.6: Summary syntactic characteristics

small district offices this is 36,000 cases, for medium district offices 60,000 cases, for large district offices this 78,000 cases, and for very large district offices this is 108,000 unemployment benefit cases. Additional calculations yield the expected data capacity for the four categories of district offices: 1.6 Gigabytes, 2.7 Gigabytes, 3.5 Gigabytes and 4.8 Gigabytes. These requirements do not take into consideration the load factor that is required for good access speeds. For example, the databases of ten district offices are stored at the RSC Eindhoven. These ten databases are categorized into two small databases, five medium databases and three large databases, requiring at least 27.2 Gigabytes of data capacity.

Transformation The documentation contains requirements for the WWO to convert data from the Digital (DEC) mini computers to the ICL mainframe computers (see section 3.2.2), and to convert data to the file structures required by the statistical information system used by the statistical research department. Two batch processes are implemented which perform these two conversions.

Accuracy Requirements and preconditions on the data characteristic accuracy are not documented. The WWO databases store sums of money to cents accuracy, and stores percentages to one-hundredth percent accuracy.

Error correction Requirements and preconditions on the data characteristic error correction are not documented. The WWO databases use after image journals to restore a database up to the last transaction completed before an error occurred.

7.3 DATA

203

characteristic	requirements and preconditions	fulfilled by properties	additional properties
reliability	0	0	0
repairability	1	0	0
recoverability	1	0	0
access time	0	0	0
transfer rate	0	0	0
storage capacity	0	0	1
safety	1	1	0
retention time	0	0	0
storage costs	0	0	0
access pattern	0	0	0
fill factor	0	0	0
total	3	1	1

Table 7.7: Summary empiric characteristics

7.3.4 Empiric level

At the empiric data level, requirements and preconditions concerning physical data representations are attributed to eleven data characteristics: reliability, repairability, recoverability, access time, transfer rate, storage capacity, safety, retention time, storage costs, access pattern and fill factor. See section 4.2.2 for more details.

The requirements and preconditions are defined using all of the data characteristics. A summary of the empiric data characteristics is shown in table 7.7, the number of requirements that are made on the empiric data characteristics is shown in the second column, the number of requirements that are fulfilled by properties of the WWO is shown in the third column, the number of additional properties of the WWO upon which no requirements are made is shown in the last column.

Repairability Outage of the WWO information systems is permitted to last one complete day in the case of a severe malfunction. This period includes both the time needed to repair the hardware and to recover the data. The contingency procedures in hand at present are unlikely to guarantee this requirement.

Recoverability Outage of the WWO information systems is permitted to last one complete day in the case of severe a malfunction. This period includes both the time needed to repair the hardware and to recover the data. The contingency procedures in hand at present are unlikely to guarantee this requirement.

Storage capacity Requirements and preconditions on the data characteristic storage capacity are not documented, 28.7 Gigabytes of DASD storage capacity is installed

at each regional service centre. The amount of storage capacity installed at the central service centre is unknown.

Safety The spaces where automation facilities are placed need to be secured against unauthorized entry, as well as against fire and water damages. The halon fire extinguishers need to be replaced by water based fire extinguishers, consequently extra provision needs to be taken to provide protection against water damage. Unauthorized entry is prevented by the use of card-keys.

7.4 Management

Management of the data of the Unemployment Benefit information system (WWO) is performed by personnel of the department Information and Automation (I&A) at the regional service centres and at the central service centre. The WWO information system is to be performed in the near future by the organization which at present performs data management of the WAO information system. The management tasks performed by the WAO data management are to be used to perform management of the WWO data. The tasks which are performed by WAO data management are described in section 7.4.1. The organizational units involved in performing management tasks are described in section 7.4.2. The table of RAWE distributions shown in table 7.8 is used to show which management units are involved in performing management tasks with regards to the responsibility, authorization, work and expertise required for that task.

7.4.1 Management tasks

The management instrumentation is used to describe the current tasks of data management of the WAO information system. The activity models describe how the management tasks are performed. Using the principles of the management instrumentation defined in chapter 4 and chapter 5 the activity models were analyzed and changes were suggested to improve some tasks. To avoid any confusion, the descriptions in the sections below assume that the tasks are performed on behalf of the WWO. More detailed descriptions of the data management tasks can be found in section 4.3.

Managing authorization of data utilization A request to use the WWO is received on a form by functional management personnel. This request is judged for validity, if the person is a GAK employee and has an relationship with the WWO the request is granted. The user is entered into the access management information system (TBS), where access rights are assigned.

Managing actual data contents When data enter by a district office functionary passes the field validations and screen validations there is no erroneous input. A different functionary agrees the entered data, which then have to be authorized by

7.4 MANAGEMENT

205

a functionary from a different unit. This prevents a benefit from being granted or withheld unjustly. Each functionary has an agenda in the WWO information system which contains the actions which must be taken before a certain date. Getting through the agenda correctly and in time is the responsibility of the functionary. The functional manager checks the agendas and checks if actions are executed in time and are written off the agenda. When this is not the case the functional manager sends a reminder to the functionary. Once a month, just before the moment payments have to be made, the functional manager starts a batch process which checks the tally numbers to find database inconsistencies. In the case of an inconsistency, the database needs to be restored and checked again for inconsistencies until no more inconsistencies are found.

Ad hoc data provision A request for ad hoc provision of data must always be on a written form. The request is handled in the same manner as a request for a new user, after which the data is provided.

Establishing and documenting data names and a data naming standard

This task is not performed by functional management but by the WWO project group at the system development department and the data management group at the system development department.

Managing the application of data definitions This task is not performed by functional management but by the WWO project group at the system development department and the data management group at the system development department.

Consultancy Occasionally advise on data definitions is requested of the data management organization. The expertise to give recommendations on improvements in application of the data definitions rests with the systems development department.

Management of access structures and storage structures Using an analyze program a database or an area of a database can be checked for wrong parameters or other characteristics that can cause poor performance of the database, e.g. fragmentation. Technical management personnel run the analyze program but when errors show up on the printout, the system development department is called because they have the knowledge to make alterations to the database structures.

Availability of data Backup and restoration of the databases are performed at the request of functional management, e.g. when they find inconsistencies in the database, and are performed when malfunctions occur. The Storage Library System (SLS) tracks all the tapes in use and is used to find tapes that are free to be used in a backup or to find tapes with data that needs to be restored. When the tapes are loaded in the tape unit the backup or restore can be started. On completion the log files are checked to find if errors have occurred, and if necessary the error conditions are removed and backup or restore is started again. If successful the tapes are stored.

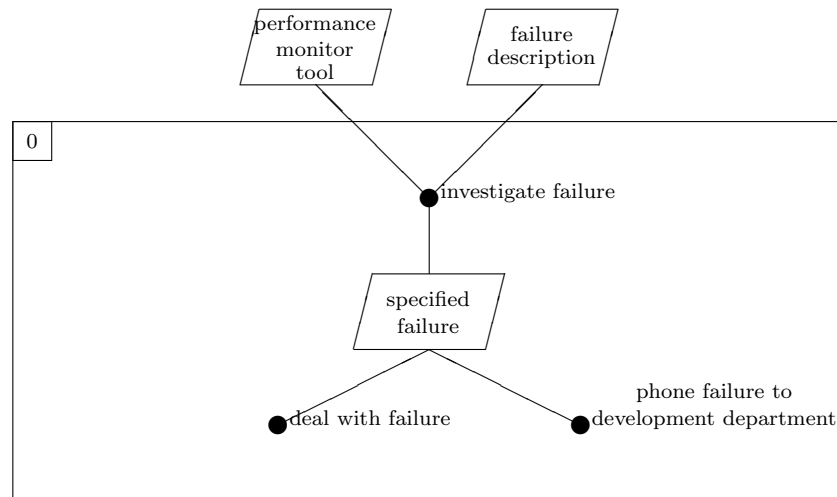


Figure 7.3: Activity model: failure recovery

Testing and evaluation During the case study no activities could be recognized regarding the testing and evaluation of the WWO databases. This means that WWO data management is not able to set standards for testing newly developed or modified WWO databases and that consequently the technical quality cannot be judged. The impact of the newly developed or modified databases can not be determined beforehand creating the risk of disruption of services.

Managing the storage of data The fill factor of DASD is monitored continuously, when predetermined thresholds are crossed an automatic warning is issued. What actions are taken to free space on the DASD are not known.

Safe keeping The most recent backup tapes are kept at the regional service centre, previous backup tapes are rotated between the three RSCs. The backup tapes of the central service centre are stored in the tape vault of the head office.

Reconstruction When a hardware failure is reported, the failure is investigated by examining the description of the failure and the reports of the performance monitoring tool. Relatively simple errors can be resolved by the technical management personnel. When the failure includes errors in the database management system, operating system, or any other complicated error the error is signaled to the systems development department by a telephone call. The activity model of the task reconstruction is shown in figure 7.3

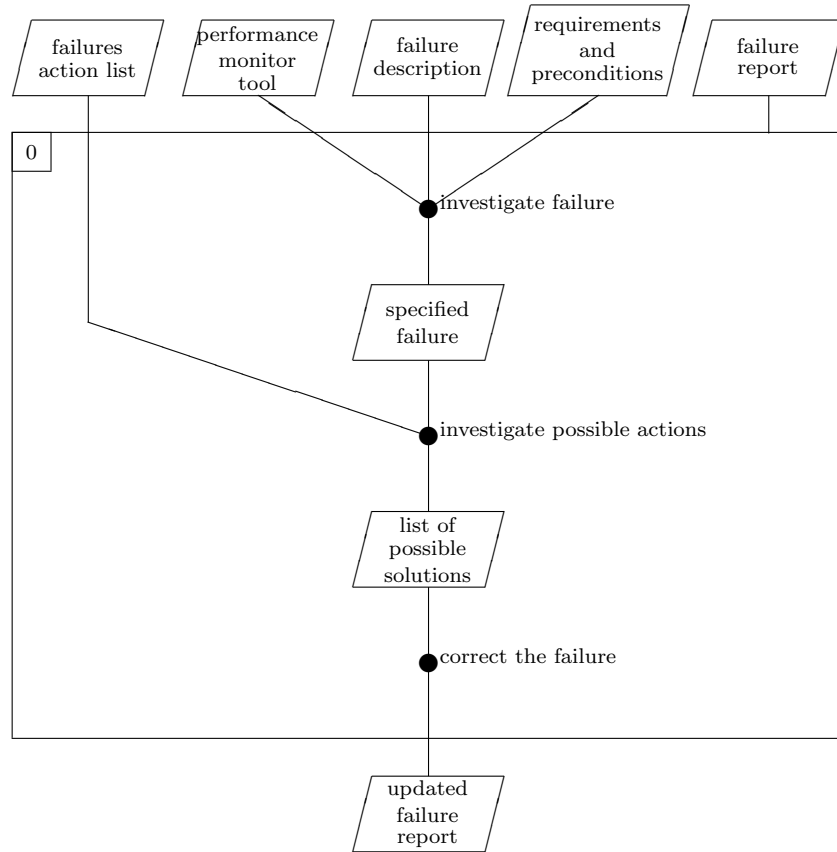


Figure 7.4: Revised activity model: failure recovery

The activity model was used in this case study to specify improvements to the way the data management task was performed [41]. In the proposed reconstruction task the failure is investigated by examining the failure description and the performance monitoring tool, based upon the requirements and preconditions and using previous failure reports to see if this failure has happened before. The detailed specification of the failure and the list of standard actions to be taken for specific failures are used to investigate the possible actions that can be taken to correct the problem. The failure is corrected using one of the possible solutions, the course taken to correct the failure is laid down in the failure report. The revised activity model of the task is shown in figure 7.4.

Destruction When a medium (tape or DASD) is replaced before it has failed, i.e. preventive maintenance, it is first initialized to remove the data before it is returned to the supplier. When a medium is replaced because it has broken down it is returned to the supplier without the data being removed.

7.4.2 Management units

The board of directors is the owner of all data contained in the WWO and of all the automation facilities that are deployed on behalf of the WWO, and as such has the competence to determine which functionalities are to be performed by the WWO and which automation facilities are to be deployed. This authority is given to the department Information and Automation (I&A) which acts as the custodian of the WWO. Within I&A, management of the WWO is divided into functional management performed by the department Information Management and Administrative Processes (IMAP), into application management performed by the department Systems Development (SD), and into technical management performed by the department Computer Centres and Infrastructures (R&I).

The department Information Management and Administrative Processes performs functional management of all the information systems and contains units for account management for each of the information systems. The unit Account Management and Functional Management (AMU-FB) is responsible for the support of the functionaries at the district offices and the head office and maintains contacts with the stakeholders outside the GAK organization.

The department Systems Development performs application management of all the information systems. The department consists of a number of portfolio holders of related applications who are responsible for the development and maintenance of the application software. Within a portfolio there are a number of project groups each having a single application (SD pgWWO). Traditionally the project groups have the expertise to determine functional specifications and technical specifications. The data management group (SD DM) supports the project groups in defining data models and the implementation and optimization of database structures of application software.

The department Computer Centres and Infrastructures performs technical management of the automation facilities and the technical infrastructures present at GAK. The department consists of service teams which perform computer operations and problem management. The service teams conclude agreements with the account managers of IMAP regarding the technical services they provide. The service team on behalf of the legislative information systems consists of a unit for each of the RSCs which manage the district office databases and the automation facilities at the RSCs. The central WWO database and the automation facilities at the CSC are managed by a unit from the service team for the GAK data infrastructure.

The distribution of management tasks with respect to responsibilities, authorizations, work and expertise of each of the management units involved is shown in table 8.4.

Tasks	Units	CSC	RSC (3×)	SD pgWWO	SD DM	AMU FB
Managing authorization of data utilization						RAWE
Managing actual data contents						RAWE
Ad hoc data provision				WE	WE	RAWE
Establishing and documenting data names and a data naming standard				WE	WE	RA
Managing the application of data definitions				E	E	RAE
Consultancy				E	E	
Management of access structures and storage structures	RAE	RA				
Availability of data	RAWE	RAWE				
Testing and evaluation						
Managing the storage of data	RAWE	RAWE				
Safe keeping	RAWE	RAWE				
Reconstruction	RAW	RAW	WE			
Destruction	RA	RA				

Table 7.8: Table of RAW distribution WWO

A separation of functions is introduced between organizations for technical management (CSC and RSC) and functional management (AMU-FB). The organization for application management also performs tasks of functional management, notably in the tasks fields data administration and database administration, which thwarts the intended separation of functions. The prime coordinating mechanism for WWO data management at GAK is mutual adjustment, this is in line with the organizational structure called ‘adhocracy’ by Mintzberg [73] and confirms the location of the custodian of the WWO as shown in the diagram in figure 8.2.

7.5 Case results

The main objective of this case study was to test the adequacy of the management instrumentation described in chapters 4 and 5 by applying it in a practical situation for analysis of data management of the WWO information system. The general model of data management was tested by applying the methods and techniques in the specific case situation. Analysis of the general model and what was observed in the specific management situation of the WWO enabled the identification of possible improvements for data management of the BRP and also of possible improvements for making the management instrumentation more adequate for use by WWO data management personnel. The fact that the management instrumentation enabled improvements to

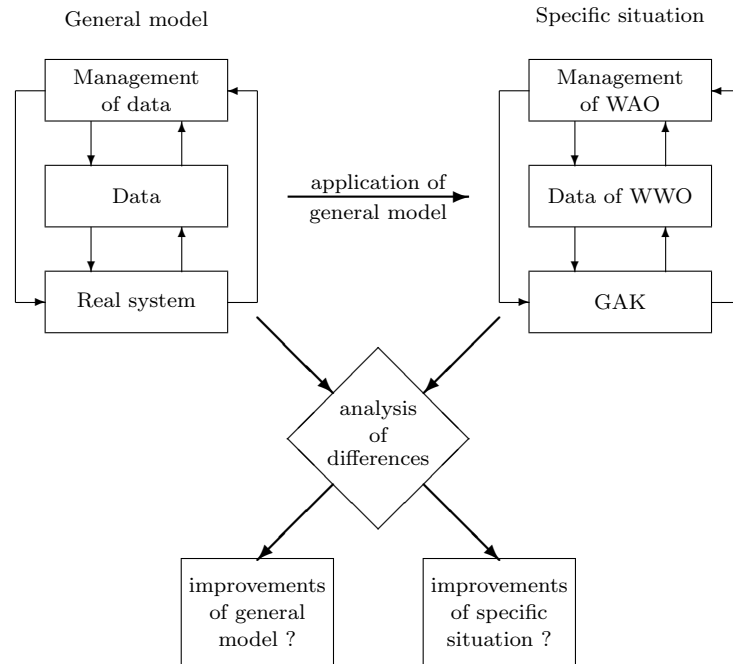


Figure 7.5: Method used to obtain case results WWO

be identified is an indicator of its applicability in this practical situation. The way of performing the case study and obtaining the case study results is shown in figure 7.5.

The WWO is a new information systems that was being introduced to the WAO management organization at the time of the case study and the conclusions and recommendations with regard to management of the WWO are given in section 7.5.1. Conclusions concerning the applicability of the management instrumentation in this case study are given in section 7.5.2

7.5.1 Conclusions: Unemployment Benefit

The central question for this case study was whether the management instrumentation was adequate in analyzing management of distributed data in distributed environments and for identifying bottle necks and providing suggestions for improvements. The conclusions of the WWO case study are described below in relation to the WWO environment and the WWO data and the WAO data management.

7.5 CASE RESULTS

211

Environment

The stakeholders of the WWO information system are not identified in the functional documentation, here only one stakeholder is mentioned: the 'user'. The management instrumentation enabled identification of twenty different stakeholders through a careful analysis of the functional documentation of the WWO information system. The study showed that this functional documentation was outdated even before the the WWO information systems was implemented, additional interviews with functional management were required. The most important stakeholders and functionalities are delineated in table 7.3:

1. The two most important stakeholders of WWO are:
 - (a) WWO functionaries at district offices GAK.
 - (b) WWO functionaries at head office GAK.
2. The three most important functionalities of WWO are:
 - (a) Alleviate time consuming tasks with regards to: pension premiums payments and applications and sign off from health insurance funds.
 - (b) Automated determination of benefits and support of the functionary determining benefits.
 - (c) To calculate deduction of benefits when multiple social security legislations apply or when other partial benefits rights exist.

The recommendations made in the case study with regard to the environment of the WWO emphasize the importance of close relationships between WWO data management and the stakeholders of the WWO. The functional documentation was outdated even before the WWO became operational and should be drawn up again by WWO data management. In the new documentation due attention should be paid to the the stakeholders, the functionalities and the relationships between stakeholders and the functionalities.

Data

The fact that no stakeholders of the WWO information system were defined, also means that no relationships between requirements and preconditions and such stakeholders could be found. Application of the management instrumentation found that requirements and preconditions were best defined at the pragmatic level of data and that they were least well defined at the syntactic and empiric level. The requirements and preconditions of data at the semantic level were documented using a relational data model from the case tool Excelerator, this consists of 750 attributes and 100 compound types. From table 7.5 it can be seen that all semantic requirements are made on the characteristic convertibility, and that the properties of implementation of the relational database differ due to the fact that additional tables are defined.

Requirements and preconditions on the characteristic reliability of data are not documented by the case tool. A bug in the relational database management system prevented the enforcement of reliability in the database and reliability had to be implemented using about 1560 lines of SQL code.

Data of the WWO information system is distributed over thirty databases placed in groups of ten databases at three regional service centres, one for each district office connected to the RSC, and one head office database placed at the central service centre. Requirements and preconditions are defined for the district databases, there are no requirements and preconditions for the head office database. The implementation of the head office database differs from the district office databases as it contains extracts from the thirty district databases.

The calculations of the data capacity requirements are not very realistic as a number of database management systems overheads are not included. The storage capacity at the regional service centres is inadequate with 28.7 GB available, as at least 27.2 GB is required this would lead to a load factor of 95% for the data without considering the additional data capacity required for 'after image' journals, and 'before image' journals etc.

The recommendations made in the case study with regard to the requirements and preconditions of the data of the WWO emphasize the fact that the relationship between the stakeholders and the requirements and preconditions form the basis of data management. The requirements and preconditions of the WWO must be drawn up for at least the semantic, syntactic and empiric data levels, as made by the stakeholders. Monitoring the properties of the thirty WWO databases, and one central database, should be implemented at the regional service centres and the central service centre.

Management

The fact that no stakeholders of the WWO information system were defined, also means that no relationships between management tasks, requirements and preconditions and stakeholders could be found. Tasks are not performed based upon requirements and preconditions and actions are undertaken based upon personal judgements. Performing management tasks involves several organizational units and coordination between them is based upon mutual adjustment on an ad hoc basis rather than structural agreements. For a number of management tasks the organizational unit that performs the tasks is not responsible for the results of performing the tasks, and the expertise required rests with a different organizational unit from that of the organizational unit that is responsible for the task and is authorized to perform it.

Analysis of the description of data management of the WWO information system and the general model provided by the management instrumentation gives a number of recommendations for management of WWO data:

- Define requirements and preconditions of data, and use these requirements and precondition as criteria for performing management tasks

7.5 CASE RESULTS

213

- Standardize work processes, outputs and skills of data management in addition to the current coordinating mechanisms of mutual adjustment and direct supervision.
- Place responsibilities, authorizations, work and expertise for a management task within the same organizational unit, and place different management tasks fields in different organizational units. This segregation of tasks reduces conflict of interest that could arise when the same organizational unit designs, implements, manages and modifies the data of the WWO information system.

The recommendations made in the case study with regard to the WWO data management organization emphasize the fact that many of the data management tasks are performed ad hoc in response to failures. Tasks should be designed to prevent failures rather than to solve failures. The definitions of the tasks should be based upon the requirements and preconditions made by the stakeholders. Correct allocation of responsibilities, authorizations, work and expertise should then be based on the tasks definition to ensure that management of both the regional databases and the central databases is possible. WWO data management should implement the testing and evaluation task field in which criteria set are for the documentation, data structures and algorithms provided by the systems development department.

7.5.2 Conclusions: Management Instrumentation

The observations made during the case study show in what respect the management instrumentation is adequate in this specific situation, and what kind of improvements of the management instrumentation are needed for this particular situation. Possibilities and ways to improve the management instrumentation are indicated by answering the two research questions stated in section 1.4.4.

Characteristics of data

Research question 1 (see page 31) asks to what extent the characteristics defined in chapter 4 are of importance for performing tasks of managing distributed data in distributed environments.

The case study did not find new data characteristics that were used by the WWO data management but that were not defined in the conceptual framework. It can therefore be concluded that the data characteristics defined in chapter 4 proved to be of value to the WWO data management.

The FATO framework prescribes that data management needs to be based on a characterisation of data to operate, control and maintain the data in accordance with requirements and preconditions. A description of how WWO data management implements these data characteristics in practice is given in section 7.3. The functional data requirements and preconditions were defined in the functional documentation of the WWO which was not up to date. This is surprising for a new information system, the

long development period (5 years) can be blamed for the fact that functionalities were outdated the moment the information system was delivered. The requirements and preconditions on automation facilities were poorly defined in the technical documentation, technical management should impose quality standards on the documentation delivered by the systems development department [41]. The following paragraphs give more detailed conclusions regarding the data characteristics used by the WWO data management at the pragmatic, semantic, syntactic and empiric data levels.

Pragmatic level All the characteristics defined at the pragmatic level of data were encountered in the data management of the WWO. The data characteristics are almost all specified qualitatively, instead of quantitatively making them hard to measure and making it difficult for the new WWO data management to meet the requirements and preconditions. The documentation of the pragmatic requirements and preconditions was out of date the moment the new information system was delivered. The WWO is in conflict with the functional specifications and the directives for use of the GAK data infrastructure in that it stores not only a person's Social Fiscal number (SoFi number) but also a number of other personal data. The case study showed that 42% of the requirements and preconditions were not met by the WWO at the point of delivery (see table 7.4). Using the table of stakeholders it could be shown that for 95% of the pragmatic requirements and preconditions it is not known which stakeholder requested them. For 95% of the stakeholders it was possible to identify what requirements and preconditions were made. There were no specific requirements and preconditions made on the central WWO database. The pragmatic data characteristics provided the future WWO data management organization with the opportunity to assess the requirements and preconditions made on the WWO at the level of actual data contents, and take action before the new information system became fully operational.

Semantic level Six out of seven characteristics defined at the semantic level of data were encountered in data management of the WWO. The data characteristic static integrity was not used to define the requirements and preconditions of WWO data. The semantic requirements and preconditions are laid down in an entity relationship data model stored in an case tool. The implemented data model is stored in the same case tool, despite this fact 70% of the requirements and preconditions are not fulfilled (see table 7.5). Due to the poor performance of the database management system, reliability had to be implemented using four database triggers for each database relation. These database triggers are considered to be a part of the application software and do not constitute a database property to be maintained by WWO data management. The securability of the databases is not clear, four views of the database are defined which shield access to certain data, but it is not clear which stakeholders have access to what data. The data model for the the central head office database is different from the main data model used for the thirty district offices databases. There are, however, no separate requirements and preconditions for the head office database and the properties had to be deduced from the examination of contents of the database

7.5 CASE RESULTS

215

data dictionary.

Syntactic level Five out of ten characteristics defined at the syntactic level of data were encountered in data management of the WWO. These data characteristics were: access speed, data capacity, transformation, accuracy, and error correction. There were just two requirements and preconditions made on a syntactic data characteristic, namely on transformation. There are an additional four data characteristics used for properties of the WWO database, of which only data capacity was specified numerically. Little was known of the actual properties because the database was not yet fully operational. The syntactic data characteristics demonstrated to the technical management of the WWO that much is as yet unknown about the WWO database. It is not possible to respond quickly and adequately to disturbances in database performance.

Empiric level Four out of eleven characteristics defined at the empiric level of data were encountered in data management of the BRP. These four data characteristics were: repairability, recoverability, storage capacity and safety. The characteristics repairability and recoverability are used to define the same requirement and precondition: in the case of a severe malfunction of the computer system, total system outage can amount to, at most, one day. During this period the malfunction has to be identified, repaired or replaced and the original data (in case of DASD outage) has to be recovered. Considering the fact that no tests were performed and the repair and recovery procedures are not well worked out, it is expected that this requirement and precondition will not be fulfilled. The halon fire extinguisher installations need to be replaced by water based fire extinguishers, consequently extra provision must be taken to provide protection against water damage to meet the requirement and precondition on the security of data. The small amount of empirical requirements and preconditions make it impossible for technical data management to control the WWO actively, and to operate and maintain it within the requirements and preconditions defined for it. The empirical data characteristics need to be defined, without them the technical management organization has to work without clear guidelines.

Methods and techniques

Research question 2 (see page 32) asks to what extent the methods and techniques defined in chapter 5 are of use to management of distributed data in distributed environments.

The case study found that seven out of the eleven methods and techniques could be used successfully. The four methods and techniques that could not be used were the create/use matrix, the semantic model, the table of stakeholders and the automation facilities matrix. During the case study it was found that the entity-relationship data model was used with the case-tool 'Excelerator', this data model failed however, to model the semantic requirements and preconditions and the semantic properties of the WWO. During the case study no other methods and techniques for data management

were found besides the E-R model, so it can be concluded that the methods and techniques defined in chapter 5 proved to be of value to the WAO data management organization.

The measures taken by data management need to be supported by methods and techniques to assure that once data are implemented, the data keep satisfying the requirements and preconditions. The current data management organization of the WAO, which is to perform data management of the WWO, does not use predetermined methods or techniques of its own, but uses the documentation it receives from the systems development department. This documentation is, however, inadequate to perform data management, as is clearly indicated by the fact that the functional documentation was five years old and out of date the moment the WWO system was delivered by the systems development department. The activity model shown in figure 7.3 gives a typical situation in which data management personnel responds to a failure event without reference to requirements and preconditions. The improved way of dealing with failures is shown in figure 7.4. During the case study use was made of available material supplemented with investigations into the properties of the WWO database and by interviews with the WAO data management personnel. More detailed conclusions regarding the role of the methods and techniques used in this case study are given below.

List of stakeholders The list of stakeholders enabled the identification of the stakeholders that were not recognized independently in the functional documentation of the WWO. The list shown in table 7.1 gave WAO data management a complete overview of all the stakeholders for the first time. The information systems and organizations mentioned in table 7.1 are within the GAK organization as well as outside the GAK organization. Maintaining close relationships with all the stakeholders mentioned in table 7.1 is a major undertaking for the functional management of the WWO considering their nationwide distribution. With regards to the information systems that are stakeholders of the BRP, relationships should be maintained between the WWO functional management and the respective information systems management departments. The list of stakeholders helped the WAO management department to obtain an impression of the stakeholders, organizations and information systems, they would be performing data management for in the near future.

List of functionalities The list of functionalities was taken directly from the functional documentation of the WWO. The list shown in table 7.2 gave the WAO data management a complete overview of all the functionalities it was expected to support.

Functionality matrix The functionality matrix was used to demonstrate what functionalities are directed towards which stakeholders. The matrix shown in table 7.3 shows that functionalities are directed mainly towards the functionaries at the district offices and the head office. The functionality matrix also shows that many stakeholders that were mentioned in the documentation as possible 'users' of the WWO do not have functionalities attributed to them. This can be explained by the fact that they

7.5 CASE RESULTS

217

were not identified as independent stakeholders. The functionality matrix shows, however, that three of the functionalities were not directed towards any stakeholder. The case study identified, for the WWO data management, which functionalities and stakeholders are most important. The case study gave the WWO data management a clear indication that a danger exists that many stakeholder will not be able to make effective use of the WWO because the functionalities that are to be supported on their behalf are not known.

Create/use matrix A create/use matrix could not be used in this case study. A create/use matrix shows which stakeholders have the right to determine actual values of the WWO data, and which stakeholders have the right to use the actual values of the WWO data. A create/use matrix was not possible because there were too many types of data (in excess of 100 tables) and because the stakeholders were not identified in the functional documentation, and therefore the relationships between data and stakeholders could not be identified. If a create/use matrix was made it would be more easy to define and maintain the requirements and preconditions of the data characteristics confidentiality, auditability and securability.

Semantic model A semantic model could not be used in this case study. Semantic models can be used to determine complex semantic relationships between data and display them graphically. The relational description of the WWO data in the case tool contains 750 attributed and 100 compound types, it was not feasible to draw a comprehensive semantic model for this amount of definitions. Furthermore, the properties of the semantic data characteristics are implemented using SQL statements in application software. Relatability is enforced using 1560 lines of SQL code, and management of this software is the responsibility of the application management organization, not the data management organization.

Table of properties A table of properties can be used to compare the actual properties of data with requirements and preconditions and to identify deviations. Based on these deviations data management can take appropriate measures. The property tables for all four levels of data, pragmatic, semantic, syntactic and empiric, were used to good effect to show the existing shortcomings in the requirements and preconditions and the actual properties of the WWO. The lack of adequate specifications of the requirements and preconditions on the functionalities and the automation facilities of the WWO, means that functional and technical management of the WWO cannot monitor the trends in service and performance of the WWO, and intervene when necessary. Failure to do so may cause unsatisfied customers (i.e. stakeholders) a problem made worse by the fact that the stakeholders are unknown to the WWO data management.

Table of stakeholders A table of stakeholders could not be used in this case study. A table of stakeholders shows the relationships between requirements and preconditions, in terms of data characteristics, the stakeholders that make the requirements and preconditions, and the tasks performed in data management to ensure that these

requirements and preconditions are met. The small amount of requirement and preconditions made on the WWO and the fact that the documentation needs to be brought up to date, means that the relationships between stakeholders, requirements and preconditions of the WWO could not be matched with data management tasks performed on behalf of the WAO information system. Without knowledge of these relationships the data management is not able to determine the consequences of changes in the environment of data, changes in stakeholders or the requirements and preconditions they impose, and the effect this has on the tasks performed by data management.

Automation facilities matrix An automation facilities matrix could not be used in this case study. An automation facilities matrix is used to show the physical distribution of automation facilities for several locations. As the WWO was not yet fully operational at the time of the case study the technical configuration of WWO automation facilities was not known.

Activity model Activity models are used to show how the tasks in management of the WWO data are performed. In this case study many tasks were performed by both functional management and technical management of the WAO data were analysed using activity models. The activity models enabled the identification of improvements in the way data management tasks were performed. The activity models were based on interviews with WAO data management personnel and were discussed with them later, an example of such an activity model is shown in figure 7.3. Analyses of the activity models was based on the FATO framework described in chapter 4 and on the definitions of information systems management and data management given on pages 22 and 25. This analysis led to improvements in the activity models to make the WAO data management tasks more suited to management of the WWO. An example of an improved activity model is shown in figure 7.4

Table of RAWWE distribution The table of RAWWE distribution describes which organizational units have the responsibility (R), authorization (A), work (W) and expertise (E) to perform tasks in BRP data management. The RAWWE distribution shown in table 7.8 made it possible to identify the way tasks are performed by the different organizational units present at GAK and where improvements could be made. As shown in table 7.8 many organization units are involved in performing a data management task. Table 7.8 shows that for some tasks the responsibilities and authorizations are separated from the work and expertise, or may even be completely absent. Table 7.8 also shows clearly the separation of functions between technical management of the WAO and functional management of the WAO.

Coordinating mechanisms Coordinating mechanisms describe how the different organizational units involved in present WAO data management, coordinate their tasks. Many tasks are performed on a daily basis, and functional and technical management of the WAO is performed by a large organization with a great number of people. Separation of tasks is carried through to a great extent in the WAO data management organization, and coordination is therefore imperative. The prime

7.5 CASE RESULTS

219

coordinating mechanism between the various units of the WAO data management organization is, however, mutual adjustment. This is shown in figure 7.3 in which a data management functionary contacts the systems development department when a failure is encountered which is too complicated for the functionary to deal with. The prime coordinating mechanism mutual adjustment is in accordance with the position of the custodian of the WWO in the support staff of GAK, as shown in the diagram in figure 7.2. In this respect the WAO data management organization agrees well with the WWO.

220

UNEMPLOYMENT BENEFIT CASE

Chapter 8

Moerdijk Process Database Case

The comprehensive tests of the management instrumentation described in chapters 4 and 5, within the Joint Office for Social Security Administration¹ (GAK), is described in chapters 6 and 7. As these two tests were carried out in the same setting, it was felt that an additional test in a completely different setting was needed to obtain a better understanding of the adequacy of the management instrumentation in real world situations. The case study described in this chapter concerns data management of the Moerdijk Process Database (MPD) used by the production plants of Shell Netherlands Chemicals at Moerdijk (SNC/M).

The setting of this case study differs in two aspects from the case studies performed at GAK. The first aspect of this case is the fact that the MPD is used in an industrial environment where data processing is not part of the primary production processes of the organization. At Moerdijk polymer semi-finished products are produced by three chemicals plants: the Moerdijk Ethylene Oxide and Derivatives plant (MEOD), the Moerdijk Lower Olefins plant (MLO) and the Moerdijk Styrene and Propylene Oxide plant (MSPO). The second aspect of this case is the fact the MPD differs from the databases at GAK in that it is not a transaction based database, but a real time database.

The method used in the case study differs from the previous two cases in that it tests the applicability of the management instrumentation with less rigour than in the cases within GAK. The case study consists of a short study of management of the Moerdijk Process Database (MPD), and is used to highlight those differences with GAK that accentuate the different application of the management instrumentation at Moerdijk. Experience gained from the previous cases enabled a more focussed approach to identifying bottle necks and produced suggestions for improving the man-

¹In Dutch: Gemeenschappelijk Administratie Kantoor (GAK)

agement instrumentation and as a result the case study was completed in just a few days. As the investigation was performed in less depth it was not possible to determine if a complete picture of management of the MPD was achieved and therefore suggestions for improvements in the management of the MPD are not given. It was possible however, to determine in what way the management instrumentation could be improved for those specific aspects where data management at Shell differs from GAK (see figure 8.4).

This chapter is fashioned in the same manner as chapters 6 and 7. Section 8.1 positions the case with respect to deconcentration of MPD data and decentralization of the MPD environment to enable cross case analysis in chapter 9, where the final research results are presented. Section 8.2 gives descriptions of the environment of the MPD with the methods and techniques used in this case. Section 8.3 describes data characteristics of the MPD. Section 8.4 describes management of the MPD. The conclusion in section 8.5 describes how specific situation differs from the previous two cases and what adaptations of the management instrumentation are required for this specific situation.

8.1 Positioning the case

Positioning of the case is achieved by following the guidelines given in section 1.4.7. Within the context of this thesis, distribution of the Moerdijk Process Database (MPD) is restricted to the geographical deconcentration of functionalities, automation facilities and the organizational management units, of the MPD using the diagram for distribution of data shown in figure 4.4 at the end of chapter 4. The 'FAO diagram' for the MPD is shown in figure 8.1. Distribution of the environment is restricted within this thesis to the decentralization of the decisional powers over the MPD within Shell Netherlands Chemicals at Moerdijk (SNC/M) using the organizational diagram of Mintzberg [73]. The 'Mintzberg' diagram for the MPD is shown in figure 8.2.

8.1.1 Deconcentration

Deconcentration of data of the Moerdijk Process Database is described using the FATO framework with a diagram for graphical representation where distribution of functionalities (F), distribution of automation facilities (A) and distribution of the management organization (O) are shown. See also section 4.5.

Distribution of functionalities of the MPD is shown by placement of 'F' segments in figure 8.1. At Moerdijk there are three plants where production data are stored in the MPD. These three plants are the Moerdijk Ethylene Oxide and Derivatives plant (MEOD), the Moerdijk Lower Olefins plant (MLO) and the Moerdijk Styrene and Propylene Oxide plant (MSPO). At each of the three plants there are several process control computers, PLCs (programmable logic controllers), data loggers and field instruments that supply real time process data to the MPD and also take data

8.1 POSITIONING THE CASE

223

from the MPD to perform some of the calculations required. Besides these automated systems, data is also used by operators and process technologists at each of the three plants. Three 'F' segments on the left hand side of the 'FAO diagram' indicate the three plants MEOD, MLO and MSPO. The central laboratory at Moerdijk supplies more accurate process data to the MPD. Due to exchange of raw materials and semi-finished products with plants at the Pernis site, process data is exchanged between Moerdijk and Pernis. The process data of the MPD is used at Rotterdam to analyze the economic performance of the plants, and finally, process data of the MPD is used by information systems at the plant management level and business level at Moerdijk. These four locations of use of the Moerdijk Process Database are indicated by four 'F' segments on the right hand side of the 'FAO diagram'.

Distribution of automation facilities of the MPD is shown by the placement of a single 'A' segment in figure 8.1. The Moerdijk Process Database is not a single database, but is implemented with three subsystems that are optimized towards specific utilization of the MPD. The Laboratory Information System (LIS) handles data of product samples at various stages of the production processes, these data are determined at a slower pace but are more accurate than real time determinations. The Moerdijk Information Control and Optimization System (MICOS) implements all real time database functionalities. The Application System Facility (ASF) implements analytical support and allows combination of data from MICOS and LIS, it is implemented using the Application System (AS) fourth generation system by IBM. The Laboratory Information System is implemented using a HP 3000 computer system, MICOS and ASF are implemented using an IBM ES/9000 computer system. These two computer systems are located at the 'bunker' at Moerdijk, shown by the single 'A' segment in the 'FAO diagram'.

Distribution of the management organization of the MPD is shown by 'O' segments in figure 8.1. Functional management is performed by the department Information and Organization (I/O) at Moerdijk which is located a short distance from the computer systems. The department information and organization is also involved in tasks of management of data structures of MICOS and ASF. This organizational unit is indicated by an 'O' segment at the bottom of the 'FAO diagram'. Technical management of the computer systems and therefore management of data storage is performed by Shell Common Information Services (SCIS) at Rijswijk. This organizational unit is indicated by an 'O' segment at the top of the 'FAO diagram'. The Laboratory Information System (LIS) is supplied by an external software company: Delaware. Some tasks of management of data structures of LIS are performed by Delaware, this is indicated by an 'O' segment on the right hand side of the 'FAO diagram'.

8.1.2 Decentralization

Decentralization of the environment of the Moerdijk Process Database is described using the distribution of decisional powers over the user organization in which data management is performed. The data environment consists of stakeholders with a

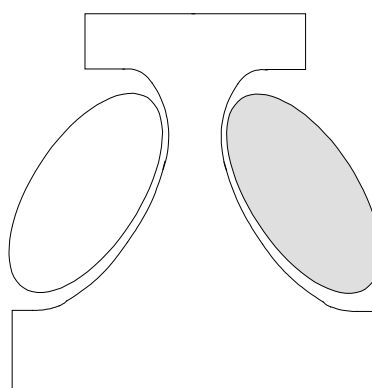
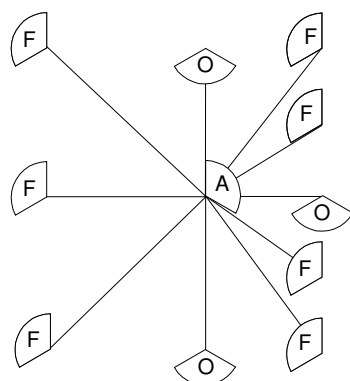


Figure 8.1: Deconcentration of MPD

Figure 8.2: Decentralization of MPD

vested interest in the functionalities performed by the data. The stakeholders of the MPD information system are all within the organizational boundary of Shell Netherlands Chemicals, though not all are at Moerdijk as is discussed in the previous section. The control over the functionalities and automation facilities of MPD data rests with the custodian of MPD, see section 2.1.3. The level of decentralization of MPD is determined by the location within the Shell organization where the power rests to make decisions about data in the MPD.

The fact that the MPD is used by a technical organization could lead to an expectation that the technostructure of Shell Netherlands Chemicals at Moerdijk (SNC/M) has a major influence in the decision making process around the MPD. This is not the case, custodianship of the MPD rests with the department Information and Organization (I/O) which is a support unit of SNC/M. All the computer systems at Moerdijk are owned by SNC/M, and all requests concerning functionalities and automation facilities of the MPD have to be approved by the department Information and Organization. The level of decentralization within SNC/M is thereby characterized as a type D decentralization (see page 34), the 'Mintzberg' diagram shown in figure 8.2 illustrates this type of decentralization.

8.2 Environment

The environment of the Moerdijk Process Database does not consist solely of people or organizational units; automated information systems both supply and use data kept by the MPD. In section 8.2.1 the architecture of systems at Moerdijk, and

8.2 ENVIRONMENT

225

other Shell locations, is therefore discussed first to show the role the MPD plays in production support at Shell Moerdijk. The stakeholders and the functionalities of the MPD identified during the case study using the methods and techniques of the management instrumentation are discussed in section 8.2.2.

8.2.1 Moerdijk systems architecture

The Moerdijk Process Database (MPD) was set up as a part of the information systems that support production at Moerdijk. The MPD is defined in the TIMMAS (Tactical Implementation Moerdijk Manufacturing Support) study [90]. The TIMMAS study makes a sharp distinction between process data and formal accounting data, each having their own specific source, use and requirements. Process data are the measured and calculated process variables as required and archived by the Moerdijk Process Database. Formal production accounting data are established according to formal accounting rules and criteria, with a frequency determined by business needs rather than production control requirements. Based upon (1) the above distinction between process data and formal production data, (2) the needs and requirements of the various user groups and applications, and (3) data transfer and consisting requirements, the TIMMAS study defines three functional levels:

- Local, and daily operation oriented functions which generate and use large volumes of process data (the “base” or production level)
- Planning, scheduling and plant management functions which operate on formal production data (the “upper” or business level)
- Intermediate functions which (1) translate plans and schedules in operating plans and detailed shift instructions and (2) establish formal production accounting data from process data to monitor actual performance and to provide top-down feedback (the “intermediate” level).

The information systems are integrated by sharing common data through common databases, and therefore all data is stored outside of, and independent of, the application software. The case study described in this chapter examines the Moerdijk Process Databases (MPD) which implements (most of) the base layer of the TIMMAS study. The MPD is not one database, it is the assembly of the Moerdijk Information Control and Optimization Program (MICOS) database, the data of the Laboratory Information System (LIS) and the tables that are part of the Application System Facility (ASF).

The TIMMAS study further concludes that key characteristics of base level data are simple structure and large volume, thus a relational database is not employed because the associated operating costs are high, this only pays for complex data. Process data have a simple structure of time series (or “tags”) of data, meaning that the strength of relational databases is not fully exploited, while the high cost of purchase and maintenance render it unattractive. Low operating costs and low

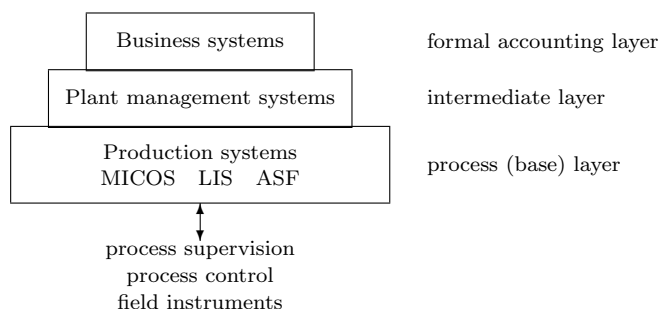


Figure 8.3: Systems architecture at Moerdijk

development and maintenance costs are key selection criteria in the TIMMAS study for the implementation of the Moerdijk Process Database.

For MICOS and related systems, simple database structures and database management systems will suffice. Access to the data still supports SQL through the Application System Facility (ASF) system. The Honeywell TDC 3000 field instrumentation systems of the Ethylene Oxide and Derivatives plant (MEOD), the Moerdijk Lower Olefins plant (MLO) and the Moerdijk Styrene and Propylene Oxide plant (MSPO) are connected to MICOS via an IBM token ring network. The automated sample analysis equipment installed at the laboratory is connected to a conversion system (LAS) which sends the laboratory data after conversion to LIS.

8.2.2 Stakeholders and functionalities

Documentation of the Moerdijk Process Database (MPD) is split into three components for the MICOS, LIS and ASF systems. Documentation of the LIS is maintained by the supplier of this system, Delaware, documentation of MICOS and ASF are maintained by the Organization and Information department present at Moerdijk. The documentation defines the different user groups and applications, in this thesis referred to as 'stakeholders', and the functions to be performed by the MPD. These are shown here using the methods and techniques defined in chapter 5; the list of stakeholders of the Moerdijk Process Database is shown in table 8.1, the list of functionalities of the MPD is shown in table 8.2 and the functionality matrix is shown in table 8.3.

Stakeholders

The plant managers at Moerdijk use the Moerdijk Process Database to obtain data concerning plant performances. The advising technologists² at Moerdijk use the Application System Facility (ASF) as a decision support tool and to perform various types of analysis and improvements of the production processes at Moerdijk. ASF contains tables to enable the combination of real time data of the Real Time Plant Management System (RTPMS) with the laboratory data of LIS for production process analyses. During the period of such an analysis, the technologists can require large amounts of historical data. The operators at the three plants use data from the MPD to monitor the production processes on terminals where the real time data is superimposed onto the process diagram on the screen. The operators require this data 24 hours a day to monitor trends in process values and show calculated values that are not available on the instrumentation systems. Laboratory workers analyze samples of products and enter analysis data into the LIS component of the MPD. These determinations are slower than the real time measurements, but are more accurate. The laboratory workers enter their lab reports in LIS and consolidate the measurements, thereby guaranteeing the correctness of the data. Automated analysis equipment of the laboratory are connected to the LAS system where the measurements are converted and then send to LIS. The three production plants at Moerdijk (MEOD, MLO and MSPO) have process control computers, programmable logic controllers (PLCs), data loggers and field instruments that supply the MICOS database with real time data, and that download data from the MICOS database to perform some of the calculations required. The Honeywell TDC3000 instrumentation system is plant specific and data transfer between plants, e.g. tank gauges or calculated data, is therefore performed via the MICOS database.

The LOLA (Unload and Load³) system is a system at the process data level of the systems architecture (see figure 8.3) that uses the quality reports of LIS for shipment of finished products to customers. The Moerdijk suppliers provide quality reports about the raw materials they deliver, these reports are transferred from LOLA to LIS. The Shell plants at Pernis provide raw materials for Moerdijk and in return obtain finished products from Moerdijk. Quality reports of the raw material are fed from the LIS at Pernis directly into tables of the Application System Facility (ASF) at Moerdijk, in return quality report of finished products are sent from the LIS at Moerdijk to Pernis, and in addition the ASF tables concerning these finished products are sent to Pernis.

The systems at the intermediate level use the data gathered by ASF and processes it to make the data suitable for the business systems at the formal accounting level. The SILO system (Shell Integrated Logistics) is a business level system that uses stock reports provided by MICOS. The Chemical Finance System (CFS) is an information system used by the commercial department at Rotterdam (recently transferred to

²This is an internal Shell name for the chemical engineers employed at the Moerdijk plants

³In Dutch: Lossen en Laden

B1	Plant managers Moerdijk
B2	Moerdijk Advising Technologists
B3	Operators and Shift foremen
B4	Laboratory workers
B5	LAS
B6	Plant control systems MEOD
B7	Plant control systems MLO
B8	Plant control systems MSPO
B9	LOLA
B10	LIS Pernis
B11	Intermediate systems
B12	SILO
B13	Chemical Finance System

Table 8.1: List of stakeholders MPD

Pernis) which uses data from the MICOS database to determine plant performance, e.g. tons of raw materials, tons of finished products, energy consumption per plant, and the balances for each production unit, the three plants and for the Moerdijk site as a whole.

Functionalities

The functionalities of the Moerdijk Process Database (MPD) concern the three types of data that are used to control the production at Moerdijk: process data, process performance data and inventory data. Process data is actual and historic data, on-line as well as off-line, raw as well as reconciled, on the physical and chemical properties of process flow, storage quantities, operating conditions and process line ups. Process performance data is data on targeted and realized process yields, and on the degree of conformance of the actual process execution with models, recipes and specific instructions. Inventory data is data on non-product and non-feed stocks inventories such as process materials in use and intermediate storage.

Process computers interface the MICOS part of the MPD to the instrumentation systems at the plants. These instrumentation systems are hydrostatic tank gauges, compressor controllers, PLC's etc. The process computers scan the measured values of the instrumentation systems, convert them to engineering units and calculated some of the derived variables. This data is sent to MICOS every two minutes, where further processing and archiving of data takes place.

The Moerdijk Process Database transfers data to information systems at the intermediate level, as well as receiving data from these levels. The production planning and scheduling systems provide various unstructured data, e.g. shift instructions, to the MPD. Data is transferred from the MPD to information systems for inventory

F1	(process) data collection and archiving
F2	emission data collection and archiving
F3	derived value calculation and archiving
F4	simple& advanced reporting
F5	performance monitoring, reporting vs. targets
F6	constraint monitoring
F7	performance analysis
F8	manual entry and approval of laboratory data
F9	archiving and reporting on laboratory analyses
F10	flexible access to process data
F11	storage and archiving of event data of LIS and coherent processing with process data.
F12	end user computing with tools to produce graphs, reports and statistical analyses.

Table 8.2: List of functionalities MPD

management, standards and quality management and model management. Model management information systems comprise process models for analysis of plant performance, and possible improvements of plant performance. The results of the optimization models are transferred back to the MPD for improved operating modes (continuous production processes) and recipes (batch production processes).

The central laboratory at Moerdijk performs analyses of product samples and enters the resulting data in the LIS. The LIS draws up reports on the product analysis which are returned to the plants

Some of the calculations in MICOS require the more precise data from laboratory analyses, to this end a connection with LIS is made to gain access to this data. Laboratory data can also be made available through ASF (Application System Facility) tables. These tables contain LIS data that is transferred to ASF to enable process engineers to make cross reference analyses between process data from the MICOS database, which is also transferred to ASF, and product data from the LIS database.

The Application System Facility is a number of support tools for the AS 4th generation programming language for: calculations, statistical analysis, reporting, advanced control and optimization. Advising technologists can use this facility to define their own reports, perform ad hoc calculations and statistical analysis and other end-user computing activities. Operators and shift foremen can make use of the predefined reports and screens but cannot make their own [92].

Functional distribution

The distribution of functionalities performed by MPD over the stakeholders is shown in table 8.3. The table is derived from the documentation present at Moerdijk in

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13
F1	X	X	X										
F2	X	X	X										
F3	X	X	X										
F4	X	X	X										
F5	X	X	X										
F6	X	X	X										
F7	X	X	X										
F8		X	X	X									
F9		X	X	X									
F10		X	X										
F11		X	X										
F12		X	X										

Table 8.3: Functionality matrix MPD

which the functionalities of the MPD and its components (MICOS, LIS and ASF) are defined [90, 91]. Functionalities are assigned to four of the thirteen stakeholders that are identified at Moerdijk. The documentation speaks of three of these human stakeholders; operators, (assistant) shift foremen and advising technologists, as the “key user groups” in the base layer are of the TIMMAS architecture. The automated systems that are dependent on the MPD are consequently in danger of falling out of the attention of data management. The Laboratory Information Systems (LIS) is used by about fifteen people at the plants and the central laboratory, MICOS and ASF are used continuously by about ten people in the control room, ten to fifteen operators use it occasionally, the maximum number of people using MICOS and AS at any given time is around fifty.

8.3 Data

The data in the base layer is informal and consists of time series produced by the process computers. An essential data integrity requirement is that these time series are uncorrupted and uninterrupted. The data in this layer come in large volumes, but the data structure is relatively simple and straightforward. The documentation at Moerdijk defines requirements and preconditions of data verbally without the use of predetermined methods or techniques.

In the sections below the data characteristics, as defined in chapter 4, are used in the description of the requirements and preconditions of MPD data. This shows how in practice the data management organization of the MPD implements the theoretically defined data characteristics. Conformity of properties of the data characteristics to the requirements and preconditions is not established in this case study, and there-

8.3 DATA

231

fore property tables are not drawn up. In just a few instances properties could be identified and laid down. The relationships between requirements and preconditions, the stakeholders that make these requirements and preconditions, and the management tasks that work to fulfil them could not be established, and therefore no table of stakeholders was drawn up.

In section 8.3.1 data characteristics of the MPD at the pragmatic level are used to structure the requirements and preconditions. In section 8.3.2 data characteristics of the MPD at the semantic level are used to structure the requirements and preconditions. In section 8.3.3 data characteristics of the MPD at the syntactic level are used to structure the requirements and preconditions. In section 8.3.4 data characteristics of the MPD at the empiric level are used to structure the requirements and preconditions. When characteristics are missing from the sections below, they could not be found during the case study of MPD data management.

8.3.1 Pragmatic level

At the pragmatic level of data requirements and preconditions concerning the actual data values are attributed to seven data characteristics: correctness, completeness, actuality, timeliness, durability, confidentiality and auditability. See section 4.1.1 for more details.

Correctness The Honeywell TDC3000 field instrumentation systems of the MEOD, MLO and MSPO plants are connected with MICOS via an IBM token ring network. The TDC300 systems provide process variables that have a unique 'tag' name and are stored in the database as a time series with this tag name. Process variables include: flows, pressures, temperatures, levels and quality measurements. MICOS uses the process variables to calculate derived variables, derived variables include: energy consumption, tank contents, mass ratios and material balances.

Limits and target values can be defined with every tag name, these values can be downloaded from MICOS to the Honeywell TDC 3000 systems where they are used to trigger process alarms. The TDC3000 systems provides indications if a measurement is suspect and must be treated with care, thus contributing to the correctness of the measurements. Indications provided by the TDC3000 system are: 'A', the value is off scan and the last value is shown; 'H', the value is above the high limit; 'L,' the value is below the low limit; '?', the value is suspect or bad. The propagation rule used by MICOS for calculated tag values is that the calculation result is suspect ('?') if any of the values used in the calculation are suspect ('?') or off scan ('A').

The process variables and derived variables are used to calculate a number of averages and to produce a number of reports. There are 6 minute averages, hourly averages, shift averages, and daily averages. Within MICOS there are 150 schematics that contain process diagrams in which actual process conditions are displayed on terminals placed in various places at Moerdijk, and 380 reports that are printed automatically through a scheduling system or can be printed on demand. Exam-

ples are shift reports, daily plant summaries, tank reports, and energy consumption summaries.

Completeness Process values are stored in MICOS in 'tags', each tag has a unique name and stores the course of a single process variable relating to an item of interest in the plants. The Honeywell TDC3000 field instrumentation systems transmit 8800 process variables to MICOS where they are stored in the same number of tags. Process variables are categorized by MICOS into: continuously changing value or full value variables (FVV) of which 3000 are stored in MICOS; into discrete status variables (SV) of which 4200 are stored; into variables that shadow identical variables in a distributed control system (DV) of which 800 are stored; and into miscellaneous data type variables (600) and into various variables (200). Some examples of status variables are: pumps on/off, switch positions, buttons, and alarms. Based upon the process variables MICOS calculates 2200 derived variables which are also stored in tags, giving a total of 11000 tags to be maintained by MICOS. Flow compensations are essential for correct calculations and 690 are stored in MICOS. The LIS database receives on average 670 results a day while the original estimate was 540 results a day. The LIS also produces laboratory reports containing product sample data, which are returned to the plants. The LIS data are transferred daily to four ASF tables for samples, results, properties and products. The ASFtables are designed to contain 36,500 sample records, 255,500 result records and 2000 property records and 325 product records.

Actuality Process variables are two minute snapshots (or 'spot values') of production process conditions taken by the instrumentation systems at the plants. The Laboratory Information System (LIS) is a component of the MPD involved with the determination of product samples. These determinations are slower than real time measurements but they are more accurate. A small section of the laboratory data (1%) is sent to MICOS immediately the data is approved. Once a day all data is transferred to ASF where they are used to perform plant performance analysis and for data archiving using the History Archive and Retrieval Program (HARP) system.

Timeliness The scan rate for individual variables is 64 seconds, in this period 8800 process variables are read by MICOS from the instrumentation systems and 2200 derived variables are calculated. During normal use of the computer system, response times are less than a second. During intensive use standard displays and alarm displays respond in less than three seconds, customized displays in less than four seconds, and history displays respond in less than ten seconds.

Durability The MICOS system is implemented using the real time database system RTPMS (Real Time Plant Management System) produced by IBM. The RTPMS can store both real time data and historical data, however the process control systems generate too much data for the history mechanism of RTPMS. The HARP system produced by Hardy Software Systems is used to extend the archive capacity of MICOS. Process variables and derived variables are available for two days, after two days old

8.3 DATA

233

values are removed from the RTPMS and transferred to the HARP system. Six minute averages are available for one month, hourly averages are available for six months, shift averages are available for five days, and daily averages are available for five years. Product quality measurements provided by LIS are stored by LIS for 18 months.

Confidentiality All tags are readable for anyone that makes use of the MICOS system, modifications of tag values is separated for different user groups.

Auditability Process variable measurements are completely automated, the tag name in MICOS shows the plant unit number where the process variable was measured, and from this it can be deduced where the data value originated. Data in LIS is entered automatically or manually, all laboratory data is authorized by a second person (e.g. lab supervisor) for subsequent use in the MPD.

8.3.2 Semantic level

At the semantic data level, requirements and preconditions concerning data definitions are attributed to seven data characteristics: convertibility, relatability, domain integrity, static integrity, insert integrity, modification integrity and securability. See section 4.1.2 for more details.

Convertibility Process variables are identified with a tag name of up to 10 characters, which must be unique within the MICOS database. Tag names are defined according to Shell standards in the TDC3000 instrumentation systems and are copied by MICOS, derived tags are defined by MICOS. Each tag is accompanied by a 20 character description of the tag that is used to identify the data in reports and displays. A tag name consists of a unit number of up to three digits, an instrument or service identifier letter, an optional qualifier letter, and an up to five digit sequence number. Unit numbers identify production units inside the plants which could be pumps, tanks, reactors etc. Service identifiers specify what kind of process variable is being monitored. Qualifiers specify if the tag has a special use such as an override or a set point.

The Application System Facility (ASF) gives users the means to create their own data tables, screen and various other types of data. AS table naming conforms to a structure standard of eight characters: a three letter application code, a one letter table type identifier, and a four letter table suffix. An AS table can contain many different types of data, graphics, queries, reports, memos etc., and the table type identifier helps to differentiate between the different types of AS tables. A table can comprise several columns, and column names consist of the four letter table suffix from the table name, an optional four letter suffix of the table where column is a key field (foreign key) and up to seven additional letters to identify the column [93].

Relatability By their nature, tags are stored unrelated in the database, however, relationships between tags are implicitly defined by calculations that are performed, or by the models in the information systems at the intermediate level of the TIMMAS

architecture. AS tables can be defined by the users themselves, relations between tables are defined by using the same columns in two tables, the naming conventions in use at Moerdijk require the column name in the referring table (foreign key) to contain the suffix of the second table where this column is a key field [93].

Domain integrity All inputs received by MICOS are in standard engineering units, conversion is done by the gateway processors within the plants. Valid domains of AS tables are maintained by the data dictionary of AS where for each column of AS tables, the size and type (numerical or alpha-numerical) is stored.

Securability The Resource Access and Control Facility (RACF) determines access to application programs, RTPMS, TSO, AS and to specific data sets (files). Once access is obtained no further protections are employed with the exception that (manual) modification of tags is limited to specific user groups.

8.3.3 Syntactic level

At the syntactic data level, requirements and preconditions concerning data access structures and data storage structures are attributed to ten data characteristics: access speed, restructuring speed, data capacity, load factor, use factor, transformation, accuracy, concurrency, error detection, error correction. See section 4.2.1 for more details.

Restructuring speed The LIS makes use of a network database management system and fragmentation of the database is not considered a problem due to extensive caching of data. The RTPMS is a fixed data area and restructuring of the database is not performed. There is not much danger of fragmentation of the VSAM data sets (files) used to store AS tables, a possible danger of fragmentation exists with user defined screens but this is not considered to be a problem.

Data capacity The LIS database cannot extend dynamically, the LIS needs to be stopped whenever the database needs to be enlarged. Capacity enlargements are planned based upon weekly capacity reports. Sizing of the RTPMS is fixed, tags overwrite themselves, so there is no overflow of the storage area. Enlarging the storage area of RTPMS is only performed on a per project basis as it costs time, and thus loss of process data.

Transformation The LIS has had a few problems establishing data transfer from the HP3000 computer system to the IBM ES/9000 computer system. At present the SNA transport and the JES2 connection are performing adequately. Automated laboratory analysis equipment sends its data to the LAS computer where conversion takes place before the data is entered into LIS. All inputs received by the RTPMS from the TDC300 instrumentation system are transformed by the gateway processors within the plants.

8.3 DATA

235

Error correction The LIS can be recovered from minor errors using the transaction log files, it is possible to use the last full backup and perform a roll forward on the database. The LIS needs to be stopped to make backups, and to enable recovery of damaged data. The RTPMS database is not a transaction based system, this means that when the database is corrupted, it is only possible to restore the last backup, all data entered after this is lost. The AS tables are stored using the VSAM file system, and the AS needs to be stopped to make backups and to recover data from the backups.

8.3.4 Empiric level

At the empiric data level, requirements and preconditions concerning physical data representations are attributed to eleven data characteristics: reliability, repairability, recoverability, access time, transfer rate, storage capacity, safety, retention time, storage costs, access pattern and fill factor. See section 4.2.2 for more details.

Reliability SCIS guarantees availability greater than 99.5%, excluding maintenance windows, this means that the computer systems are required to be unavailable due to failures for less than 43 hours a year. The time required to repair and recover the computer systems is taken into account in calculating the availability figures.

Repairability There exists a maintenance contract between SNC/Moerdijk and SCIS by which a repair engineer has to be on-site within four hours, besides this there are no other binding agreements. The mean time required to replace a DASD (direct access storage device) unit depends on the number of paths used to connect it with the processing unit: a dual path installation is expected to take 2.9 hours, a four path installation is expected to take 3.6 hours. The installation of a tape unit is expected to take 2.5 hours.

Recoverability No recovery test was performed during acceptance and implementation of the ES/9000 computer system.

Storage capacity The HP3000 computer system is used by LIS along with other information systems, all together they have 125 MB internal memory available, the LIS database has 450 MB storage space available. The IBM ES/9000 computer system has 64 MB internal memory available, and 17 GB of DASD storage. Two DASD units are used for RTPMS, one for the real time log arrays and one for the disk resident database arrays, two DASD units are used for the HARP database, and one DASD unit is used for the Application System (AS).

Safety Safety of the ES/9000 computer system is maintained by RACF profiles. A RACF profile is identified by a profile name and contains a list of users and user groups that are authorized to access the resource related to this profile. These resources can be DASD volumes, tape cartridge volumes and datasets (files). The following user groups have been distinguished: system and/or other systems, RTPMS engineers,

Moerdijk advising technologists, operators, administration management, LIS users, CFS users, and specific (individual) users. Physical access to the devices kept in the 'bunker' at Moerdijk is constrained by use of access keys. Backups are kept in a vault at another building at Moerdijk.

Retention time Backups of the AS are kept seven days on-line. The durability of the tags stored in MICOS is defined in section 8.3.1 but the retention of storage is not specified independently.

8.4 Management

Management of the Moerdijk Process Database is performed at Moerdijk by personnel of the department Information and Organization (I/O), by Shell Common Information Services (SCIS) at Rijswijk, and by the supplier of LIS software Delaware. These organizational units of the MPD data management organization perform management tasks that are designed to operate, control and maintain the data contained in the MPD in agreement with requirements and preconditions set by the stakeholders of the MPD. How MPD data management implements the management tasks defined in chapter 4, is described in section 8.4.1. The organizational units involved in performing management tasks are described in section 8.4.2. A table giving RAWEDistribution is used in section 8.4.2 to show which management units are involved in performing management tasks with regards to the responsibility, authorization, work and expertise required for that task.

8.4.1 Management tasks

Management tasks are defined in chapter 4 and are grouped into four tasks fields. Management tasks placed within the task field actual data management are: managing authorization of data utilization, managing actual data contents, and ad hoc data provision. Management tasks placed within the task field data administration are: establishing and documenting data names and a data naming standard, managing the application of data definitions, and consultancy. Management tasks placed within the task field database administration are: management of access structures and storage structures, availability of data, and testing and evaluation. Management tasks placed within the task field storage management are: managing the storage of data, safe keeping, reconstruction, and destruction. More detailed descriptions of the data management tasks can be found in section 4.3.

Managing authorization of data utilization The Resource Access and Control Facility (RACF) tool is used to determine access of users to applications, RTPMS, TSO, AS, and specific data sets (files). RACF is also used for resetting passwords and invalidating user identifications. A user identification must be added to a database segment before it can be used for AS. A user is given a limit on the amount of space

8.4 MANAGEMENT

237

they are allowed to use. The total amount of space within the database segment is not checked and needs to be carefully administered. Monitoring authorizations and access, e.g. intrusions, is performed by SCIS and their reports are discussed in monthly consultations with the department Information and Organization (I/O). New user identifications are created in RACF by the I/O department after consultation with the 'plant teams' that represent the users.

Managing actual data contents Removing old tags from MICOS is provided for test runs of newly developed application software, and for temporary calculations. Sometimes old historical data is retrieved from tape to be used in long term process analyses.

Establishing and documenting data names and a data naming standard Tag names contained by MICOS are defined by the advising technologists at Moerdijk, most tags are taken from the old control computer that was previously used to monitor process variables. New tags are proposed by individual users and are approved by the plant consulting group. The management of AS table meta data is automated using a set of tools called the "meta data information system". It allows the advising technologists as well as AS data administration to search for all data definitions in the TIMMAS public libraries. Access to the AS public libraries depends on the user's authorization thus access to data is constrained in this way. Management of data definitions of LIS is performed by the supplier Delaware and remains outside the scope of the MPD data management organization.

Managing the application of data definitions Tag names have to conform to Shell standards, the use of tags by the plant instrumentation systems is monitored by close contact with the plant engineers. AS table names and column names are inspected regularly using listings provided by the "meta data information system". Management of data definitions of LIS is performed by the supplier Delaware and remains outside scope of the MPD data management organization.

Consultancy The database administrator (DBA) gives advice to the advising technologists with regards to the screens and AS tables they can define themselves using the the Application System Facility.

Management of access structures and storage structures Monthly reports that signal trends in growth of use and degradation of response times of the RTPMS database are sent to Moerdijk. Management of AS tables is supported by TSO accounting tools which are used to monitor resource use by the individual end users and to assess the load on the files. A database tool is used to reports weekly on the status of LIS data, but is only used ad hoc to investigate user complaints. When the database no longer performs well enough, the supplier Delaware is contacted for action, and because the LIS database is maintained by Delaware, diagnostics and modification of the database are performed by Delaware.

Availability of data Full backups and incremental backups secure the data of the RTPMS database. Full backups are performed once a week just before the start of the weekend, incremental backups are performed each day just after office hours, each day one or two system volumes are backup up. If new tags are added between full backups and incremental backups, database configuration has to be redone by hand.

System backups of the Application System (AS) are performed during lunch-hours because support personnel is available in case something goes wrong. System backups require AS to be closed down for a short time. Closing down AS during the night might affect overnight operations due to lack of support personnel. The system backups are stored 7 days on-line, the MVS/ESA Generation Data Group (GDG) tools automatically deletes old versions of backup files. User initiated backups of AS data are also possible. A user can backup an AS table directly using a pull down menu option, this should run on-line and the AS table must be able to be restored by the user restore program. Restores of AS table can be user initiated, the user makes a request for a restore from the system backups by selecting the appropriate pull down menu. This is not necessary for restore of a user initiated backup. The restore request is placed in a table for later processing by a scheduled restoring job. This job extracts the requested AS table from the system backup and puts it at the disposal of the user. Full backup of the LIS network database takes place once a week just before the weekend, incremental backups are made every evening.

Backups are not performed during the weekends, as support personnel are not available, so in the worst situation data added during the weekend is lost.

Testing and evaluation During the case study no activities could be recognized regarding the setting of testing standards, and evaluation of the MPD. This means that MPD data management is not able to set standards for testing newly developed or modified databases and that consequently the technical quality cannot judged. The impact of the newly developed or modified databases can not be determined beforehand creating the risk of disruption of services. The RTPMS database is an extremely stable database that is seldom modified. New releases of AS database software are obtained from IBM and the LIS database is maintained by Delaware.

Managing the storage of data The ES/9000 computer at Moerdijk is connected to the mainframe of SCIS in Rijswijk by a Remote Support Facility (RSF) tool, which enables SCIS to manage the ES/9000 out of Rijswijk. The Resource Management Facility (RMF) measures and reports on the performance and availability of a number of things: the ES/9000 computer; storage spaces and address spaces; JES jobs, individual or groups of jobs. RMF issues warnings about performance problems as they occur so that actions can be taken before the problem becomes critical. RMF also provides long term measurement of systems performance that can be used for tuning and capacity planning. Long term data collection and reporting on availability of systems hardware carried out using the tool System Availability Manager (SAM). The HP3000 computer system is managed by monitoring the number of input/output operations and number of users, only in problem situations are additional tools used

8.4 MANAGEMENT

239

to monitor access patterns etc.

Safe keeping The IBM ES/9000 computer system, and the HP3000 computer system are placed in a bunker at Moerdijk which is accessible by key-card. There are no fire extinguishers present at the bunker. Diverting operations to another location in case of a disaster is not practical because the process computers are connected physically to the bunker. The database backups are predetermined, at the beginning of the week a list is printed that contains the labels of the 3490 cartridges and DAT cassettes to be used for backups that week. In the morning of each working day help desk personnel store the backup tapes (type 3490 cartridges for the ES/9000 and DAT cassettes for the HP3000) in the tape vault, which is placed in a different building on the site, and takes the new set of cartridges and cassettes back to the bunker where they are loaded into the tape units. When the wrong cartridges or cassettes are loaded, backup is not performed and a warning is given which prompts the help desk to retrieve the correct cartridges or cassettes. During the weekend no backups are made, so tapes do not have to be stored in the vault.

Reconstruction In the event of a calamity that leads to damage of data storage medium, the correct tape set is retrieved from the tape vault and loaded into the tape unit. When a system disk needs to be recovered, the computer is shutdown and restarted stand alone, when a disk is recovered that is used by RTPMS, AS or LIS, only the program is stopped. Recovery of the LIS disks is followed by a roll forward of the database to obtain the most recent transaction. After recovery is completed the respective programs are started again.

Destruction DASD or tape cartridges used by the ES/9000 computer are first initialized to remove data before they are returned to IBM. After a crash of the disk, or if tape initialization is not possible, both are returned 'as is'. DAT cassettes used by the HP3000 computer are handed over to SCIS for destruction, the disks used by the HP3000 computer are handed over to an external destruction company.

8.4.2 Management units

Within Shell Netherlands Chemicals at Moerdijk (SNC/M) the department Information and Organization (I/O) is responsible for information systems management. The application portfolio is part of I/O, it contains the custodian of the information systems and the business information analyst for functional management of the information systems, referred to by Shell as "application management"⁴. The help desk is a second unit of I/O that is responsible for save storage of the backup tapes. The database administrators (DBA) of I/O monitor the databases of RTPMS and AS, and perform recovery procedures. The database of LIS is maintained by the supplier, Delaware. Physical storage of data is monitored and controlled via a remote link by

⁴Not to be mistaken with application management defined in section 1.2.3

Tasks	Units	Application management	Help desk	DBA	SCIS	Delaware
Managing authorization of data utilization		RAWE			WE	
Managing actual data contents		RAWE			WE	
Ad hoc data provision				RAWE		
Establishing and documenting data names and a data naming standard				RAWE		WE
Managing the application of data definitions						
Consultancy				RAWE		
Management of access structures and storage structures				RAWE	WE	WE
Availability of data				RAWE		
Testing and evaluation						
Managing the storage of data					RAWE	
Safe keeping			W		RA E	
Reconstruction			W		RAWE	
Destruction					RA	

Table 8.4: Table of RAW E distribution MPD

the Shell Common Information Services (SCIS) organization at Rijswijk. The distribution of management tasks with respect to responsibilities, authorizations, work and expertise of each of the management units involved is shown in table 8.4.

Within the department Information and Organization four people are involved in management of the Moerdijk Process Database, within Shell Common Information Services two people are involved, one for the IBM systems and another one for the HP systems. Despite such a small number of people to manage the MPD, care is taken that the custodian is not the same person as the one who initiates enhancements or other requests for funds. Besides this distinction, there is virtually not separation of functions, nor is the need felt for such a separation of functions. The prime coordinating mechanism for data management at Shell Moerdijk is mutual adjustment, this is in line with the organizational structure called ‘adhocracy’ by Mintzberg [73] and confirms the location of the custodian of the MPD as shown in the diagram in figure 8.2.

8.5 Case study results

The main objective of this case study was to test the adequacy of the management instrumentation in a different setting from the case studies performed at GAK. This case study does not involve the same depth of research into management of data as

8.5 CASE STUDY RESULTS

241

was performed at GAK. As a consequence no conclusions or recommendations are made concerning the management of data of the Moerdijk Process Database. The case study made use of material that was already available and the observations were supplemented with interviews with responsible data management personnel. The general model of management of data was tested by applying the methods and techniques offered by the management instrumentation in the specific case situation at Shell Moerdijk. Previous case studies performed at GAK provided sufficient indications of the adequacy of the management instrumentation, to test the management instrumentation further, an emphasis is placed on the contrasts between the industrial environment of Shell Moerdijk and the administrative environment of GAK. In this manner the case study helped to obtain a better understanding of the practical applicability of the management instrumentation in a wider range of data management settings. Analysis of the general model and what was observed in the specific situation of management of the MPD enabled identification of possible improvements for making the management instrumentation more adequate for use by MPD data management personnel. The way of performing the case study and obtaining the case study results is shown in figure 8.4. Conclusions concerning the applicability of the management instrumentation in this case study are given in section 8.5.1.

8.5.1 Conclusions: Management Instrumentation

The observations made during the case study show in what respect the management instrumentation is adequate in this specific situation, and what kind of improvements of the management instrumentation are needed in this particular situation. Possibilities and ways to improve the management instrumentation are indicated by answering the two research questions stated in section 1.4.4.

Characteristics of data

Research question 1 (see page 31) asks to what extent the characteristics defined in chapter 4 are of importance for performing tasks of managing distributed data in distributed environments.

No new data characteristics were found during the case study that were used by the MPD data management but that were not defined in the conceptual framework. It can therefore be concluded that the data characteristics defined in chapter 4 proved to be of additional value to the MPD data management.

The FATO framework prescribes that data management needs to be based on a characterisation of data to operate, control and maintain the data in accordance with requirements and preconditions. A description of how MPD data management implements these data characteristics in practice is given in section 8.3. The data characteristics were defined verbally in the documentation rather than using a predetermined model or framework. Documentation of LIS is maintained by the supplier Delaware, and without the specification of the LIS data characteristics, data manage-

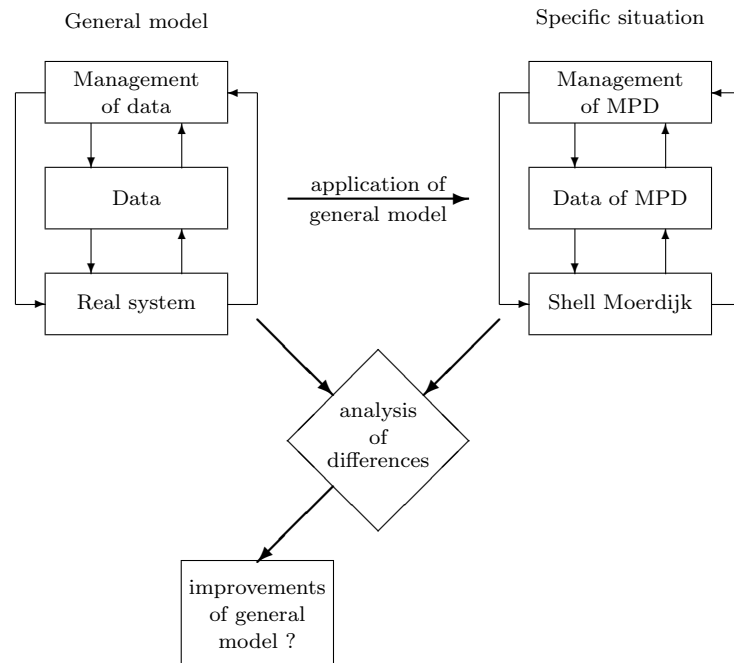


Figure 8.4: Method used to obtain case study results MPD

ment cannot be based on the characteristics of LIS and is in danger of failing to meet the requirements and preconditions. The following paragraphs give more detailed conclusions regarding the data characteristics used by the MPD data management at the pragmatic, semantic, syntactic and empiric data levels.

Pragmatic level All characteristics defined at the pragmatic level of data were encountered in data management of the MPD. The amount and detail of the characteristics is however much less than encountered in the GAK case studies. Supply of data is almost completely automated, and pragmatic characteristics are guaranteed before data is entered into the MPD, making it less necessary for MPD data management to specify them explicitly because it does not have to enforce the requirements and preconditions. This in contrast to data management at GAK where the quality of data from the various sources can not be guaranteed, and data management needs to take extensive measures to meet requirements and preconditions. The most stringent requirement is made on real time timeliness: if data cannot be read within the

8.5 CASE STUDY RESULTS

243

2 minute time slot it is lost; though even here the automated nature of Shell production processes and data supply to the MPD required little attention from MPD data management. The characteristics confidentiality and auditability are little employed, certainly compared to GAK, and mostly designed to prevent accidental errors. This is in contrast with GAK where one has to reckon seriously with the danger of fraud in social security benefits and with the protection of personal privacy, requiring extensive security measures.

Semantic level Four out of seven characteristics defined at the semantic level of data were encountered in data management of the MPD. The simple structure of the databases of the MPD make semantic characteristics less required. An additional benefit is constituted by the fact that all stakeholders of the MPD, both human and automated, are within the Shell organization and management of semantic, and pragmatic, requirements and preconditions is much easier compared to GAK where there are stakeholders in many different kinds of organizations outside of GAK. Tags are stored as an unrelated time series of data, in which relationships are only determined by the calculations that are performed on them. The tables of ASF are related by their column names only, so automated enforcement of relatability is not possible, and a danger of violating referential integrity exists. A special tool has been created to enforce the table naming rules set by MPD data management. The characteristics static integrity, insert integrity, and modification integrity are not employed by MPD data management. Securability is mainly implemented using RACF to limit access to application software, people with legitimate access are not considered a risk to the security of the data.

Syntactic level Four out of ten characteristics defined at the syntactic level of data were encountered in data management of the MPD. The characteristics are specified verbally, however, numerical specification is certainly possible. The stability and simple structure of the databases, and the automated data supply means that, compared to data management at GAK, MPD data management has less need to monitor the performance of the databases, and consequently requires less rigorously specified characteristics. Without clear cut requirements and preconditions for each of the ten syntactic characteristics the danger exists that the databases appear to be performing within generous boundaries of services and performance, while in fact they are not. Characteristics that are not monitored could be exhibiting a trend towards a level that may cause serious degradation of services and performance in the near future, taking data management by surprise.

Empiric level Six out of eleven characteristics defined at the empiric level of data were encountered in data management of the MPD. Although more numerical requirements and preconditions are specified compared with previous characteristics, the specifications are still predominantly verbal. The figures for availability are not solid as recoverability of the computer systems has not been tested and no binding agreements exist between SCIS and Moerdijk concerning the periods in which indi-

vidual failures are to be recovered. Only the total annual availability is specified. The fact that the MPD is not critical to the operations at Moerdijk might contribute to this, process analyses may be hindered by the loss of process data and process performance data, but actual production processes at Moerdijk are not effected. Sizing of storage capacity is critical for the RTPMS database to be able to meet the real time requirements, this importance is demonstrated by the detail with which storage capacity is specified. Physical safety at the empirical level is more strictly defined and maintained than at the other three levels of data, this can be attributed to the fact the computer systems are placed in a mini computer centre at Moerdijk, and that standard computer centre procedures are followed.

Methods and techniques

Research question 2 (see page 32) asks to what extent the methods and techniques defined in chapter 5 are of use to management of distributed data in distributed environments.

During the case study it was found that five of the eleven methods and techniques could be used successfully. The five methods and techniques that could be used in this case study were: the list of stakeholders, the list of functionalities, the functionality matrix, the table of RAWES distribution and the coordinating mechanisms. The six remaining methods and techniques that could not be used were: the create/use matrix, the semantic model, the table of properties, the table of stakeholders, the automation facility matrix and the activity model. No other methods or techniques for data management were found, so it can be concluded that the methods and techniques defined in chapter 5 proved to be of value to the MPD data management organization.

The measures taken by data management need to be supported by methods and techniques to assure that once data are implemented, the data keeps satisfying the requirements and preconditions. Data management of the MPD does not make use of predetermined methods and techniques, but gives a pragmatic implementation to data management in the traditional way of information systems management: ad hoc and reacting to unrelated events. Methods and techniques employed within MPD data management are not prescribed by a particular management method, but are selected based on individual needs and availability and capability of tools within the data management unit. The case study made use of available material supplemented by interviews. More detailed conclusions regarding the role of the methods and techniques in this case study are given below.

List of stakeholders The list of stakeholders enabled the joint identification of all the stakeholders mentioned in the various documentation of the MPD components. The information systems and user groups identified as stakeholders in table 8.1 are all within the Shell organization, in contrast to the various organizations where the stakeholders of the GAK can be found.

8.5 CASE STUDY RESULTS

245

List of functionalities The list of functionalities enabled the joint identification of all the functionalities mentioned in the various documentation of the MPD components. The list shown in table 8.2 may seem extensive, but the functionality matrix shows that all these functionalities are directed to only four out of thirteen stakeholders. This gives rise to the expectation that further examination will discover additional functionalities for the other nine stakeholders. In this manner the responsibilities of MPD data management will become clear.

Functionality matrix The functionality matrix is used to demonstrate which functionalities are directed towards which stakeholders. The matrix shown in table 8.3 proves that the functionalities provided by the MPD are directed mainly towards four 'human' stakeholders, thereby disregarding nine automated stakeholders out of thirteen stakeholders. Functionalities are of course still performed by the MPD on behalf of the nine automated stakeholders, otherwise they could not function, but the functionality matrix indicates that the automated systems are in danger of being ignored by MPD data management.

Create/use matrix A create/use matrix could not be used in this case study. Create/use matrixes show which stakeholders have the right to determine actual values of data contained in the MPD, and which stakeholders have the right to use the actual data values. A create/use matrix is defined in the TIMMAS study but concerns the relationships between the information systems and data in the three levels of the Moerdijk systems architecture (see figure 8.3). Create and use rights of MPD data are not defined, and this would be rather difficult with 11,000 different tag names, many AS tables which are created ad hoc by the users themselves, and LIS data which is maintained by Delaware. Together with the low priority placed on confidentiality and auditability as identified in previous paragraphs, this results in little benefit to be gained from a create/use matrix in this situation.

Semantic model A semantic model could not be used in this case study. Semantic models can be used to determine complex semantic relationships between data and display them graphically. Tags are not related to each other, except by their use in calculations, and the LIS data model is maintained by Delaware. AS tables are created ad hoc by end users according to naming conventions enforced by MPD data management. Without the use of the semantic model, it is difficult to analyze complex data definitions and maintain the integrity of semantic data characteristics.

Table of properties A table of properties could not be used in this case study. A table of properties can be used to compare the actual properties of data with the requirements and preconditions and identify deviations. It has been mentioned in previous paragraphs that there is little use of data characteristics by MPD data management, and because of this property tables can not be drawn up. Reports are made by MPD data management, but without objective judgement criteria for the figures they contain they are of little use. Data management cannot monitor the trends in service and performance of the MPD, which would be possible if the

property tables were employed to monitor and control the MPD directly based upon the data characteristics.

Table of stakeholders A table of stakeholders could not be used in this case study. A table of stakeholders shows the relationships between requirements and preconditions, in terms of data characteristics, the stakeholders that make the requirements and preconditions, and the tasks performed in data management to ensure that these requirements and preconditions are met. Little use of data characteristics, unknown relationships between requirements and preconditions and stakeholders, and tasks performed by data management, means that the table of stakeholders cannot be made. Without the knowledge of these relationships data management is not able to determine the consequences of changes in the environment of data, changes in stakeholders or the requirements and preconditions they impose, and the effect this has on the tasks performed by data management.

Automation facilities matrix An automation facilities matrix could not be used in this case study. An automation facilities matrix shows the physical distribution of automation facilities of several locations. As all the automation facilities are placed in the “bunker” at Moerdijk, this matrix is redundant.

Activity model Activity models could not be used in this case study. Tasks performed by MPD data management are not laid down in procedures or other descriptions, this results in the handing over tasks to other persons becoming even more difficult. The tasks data availability and safe keeping are specified in more detail than others, this can be explained by the fact that these tasks are performed regularly by data management and involve more than one person. Tasks which are not performed regularly are not so detailed. Activity models have been applied successfully in the WWO case study and would be successful for MPD data management.

Table of RAWE distribution The table of RAWE distribution describes which organizational units have the responsibility (R), authorization (A), work (W) and expertise (E) to perform tasks in MPD data management. In drawing up table 8.4 it becomes clear that most data management tasks are performed by the database administrators (DBA) of the MPD and by Shell Common Information Services (SCIS). It is notable that the database administrators performs tasks in the task fields ‘data administration’ and ‘database administration’, and that SCIS performs not only technical management tasks but is also involved in tasks of functional management.

Coordinating mechanisms Coordinating mechanisms are used to describe how the different organizational units involved in MPD data management, coordinate their tasks. As few requirements and preconditions are made on the MPD, and because of the few tasks performed; the functional and technical management of the MPD is performed by just a few people. To introduce separation of tasks in this small organization would be counter productive. Coordination is therefore less required compared with data management at GAK. The prime coordinating mechanism within MPD data

8.5 CASE STUDY RESULTS

247

management is mutual adjustment, which is in accordance with the position of the custodian of the MPD in the support staff of Shell Moerdijk, as shown in the diagram in figure 8.2.

248

MOERDIJK PROCESS DATABASE CASE

Chapter 9

Conclusion

In this thesis the issue of management of data in general and the issue of management of distributed data in distributed environments in particular is addressed. The research was aimed at providing a comprehensive definition of management of distributed data in distributed environments and providing management of data with a well tested instrumentation. To this end, three objectives for this research were defined:

1. To specify a framework containing all concepts applicable to management of distributed data in distributed environments.
2. To develop methods and techniques to manage distributed data in distributed environments.
3. To test the adequacy of the complete management instrumentation

The first research objective was met by definition of the FATO framework in chapter 4. The FATO framework consists of four components: the Functionalities of data, the Automation facilities of data, the Task fields of management and the Organization of management. All concepts pertaining to management of data and the relationships between the concepts are placed within this framework.

The second research objective was met by developing methods and techniques in chapter 5. These methods and techniques describe management of data at the systemological level of information systems [107] where the effect of the information system, here data management, on the real system, here data, is considered. They are derived from methods and techniques from Business System Planning (BSP) [50] and Information Systems work and Analysis of Change (ISAC) [62] modified to suit the specific requirements of management. The FATO framework and the methods and techniques together constitute the instrumentation for management of data in general, and management of distributed data in distributed environments in particular.

The third research objective was met by applying the management instrumentation in the three case studies described in chapters 6, 7 and 8. Two research questions were formulated to determine the adequacy of the instrumentation. The first research question addressed the practical applicability of the data characteristics defined within the FATO framework. The second research question addressed the practical applicability of the methods and techniques.

In this final chapter the research questions are reviewed and the applicability of the instrumentation for management of distributed data in distributed environments is shown. In section 9.1 the three case studies are positioned with regards to their respective distribution of data and of environment. After positioning the deconcentration and decentralization of the case studies with respect to each other, the results of the individual case studies are combined to describe the research findings in section 9.2. These findings are used to provide answers to the research questions in section 9.3. Ways to improve and extend the management instrumentation are discussed in section 9.4.

9.1 Positioning the case studies

Two practical situations of management of distributed data in distributed environments were observed during the preliminary surveys of the research, and three further practical situations were investigated during testing of the instrumentation using a case study approach. The preliminary surveys described in chapter 3 investigated the Dutch Ministry of Defence and the Dutch Joint Office for Social Security Administration¹ (GAK). Here a more intuitive description of distribution was used to gain insight into current practice of management of data at a general level. After development of the FATO framework for management of data, distribution of data was divided into three types in section 4.5:

Functional distribution of data is defined in this thesis as the distribution over several geographical locations of utilization of the actual contents and definitions of data for the knowledge, control and support of real systems.

Technical distribution of data is defined in this thesis as the distribution over several geographical locations of the exploitation of data structures and data storage that are implemented using available automation facilities.

Organizational distribution of data is defined in this thesis as the distribution over several geographical locations of the people and means employed in management of data, and also the distribution of responsibilities, authorizations, work and expertise required to perform management tasks over several geographical locations.

¹In Dutch: Gemeenschappelijk Administratie Kantoor (GAK)

9.1 POSITIONING THE CASE STUDIES

251

In this fashion distribution of data is related to the locations where data is utilized, exploited and managed, and can therefore be used within this thesis to describe the deconcentration of data using the diagram defined in section 4.5. Decentralization of the data environment is described within this thesis using the organizational position of the custodian within the organizational diagram defined by Mintzberg (see page 35).

These descriptions of deconcentration and decentralization of the data environment are employed in the three test case studies described in chapters 6, 7 and 8, to position the case studies to enable cross case analysis of the application of the management instrumentation defined in chapters 4 and 5. Two test case studies were performed at GAK in Amsterdam and one test case study was performed at Shell Netherlands Chemicals in Moerdijk, though less extensively than the case studies at GAK.

9.1.1 Deconcentration

Functional distribution

Distribution of the utilization of the actual data and the data definitions is shown by the placement of the 'F' segments in diagrams shown in figures 6.1, 7.1 and 8.1 for the Personal Data Register² (BRP), the Unemployment Benefit Information System³ (WWO) and the Moerdijk Process Database (MPD) respectively. In each diagram, functionalities show a large dispersion over many different stakeholders and locations.

The functionalities of the BRP are utilized by 27 different stakeholders, within GAK and outside of the GAK organization. These stakeholders range from individual registered persons to automated information systems at the regional service centres, and organizational units at the district offices.

The functionalities of the WWO are utilized by 20 different stakeholders within GAK and outside the GAK organization. The main stakeholders are the functionaries of GAK involved with administration of unemployment benefit applications, and the occupational associations and social security funds involved in payment of the unemployment benefits.

The functionalities of the MPD are utilized by 11 different stakeholders within the Shell organization. The main stakeholders of the MPD are the three production plants at Moerdijk, stakeholders outside Moerdijk are plants at Pernis and financial departments in Rotterdam.

Technical distribution

Distribution of the exploitation of automation facilities for implementation of data structures and data storage is shown by the placement of the 'A' segments in diagrams shown in figures 6.1, 7.1 and 8.1 for the BRP, WWO and MPD respectively. In these diagrams, dispersion of automation facilities vary between the three case studies.

²In Dutch: Basisregistratie Personen (BRP)

³In Dutch: Werkloosheidswet Ontslag (WWO)

The automation facilities of the BRP are placed at four locations. The main database is located at the central service centre (CSC) at Amsterdam, extracts for retrieval of personal data are located at three regional service centres (RSCs) in The Hague, Eindhoven and Hengelo.

The automation facilities of the WWO are placed at 34 locations, however, the thirty main databases are located at three regional service centres (RSCs) and the central service centre (CSC) contains the central database, the temporary files are located at thirty district offices but these are relatively small in size.

The automation facilities of the MPD use two different computer systems placed at the same location. The MPD consists of three separate systems: the laboratory information system (LIS) and the Moerdijk information control and optimization system (MICOS) and the Application System Facility (ASF). The LIS uses an HP 3000 computer system, the MICOS and ASF use an IBM ES9000 computer system.

Organizational distribution

Distribution of the organization of data management is shown by the placement of the 'O' segments in diagrams shown in figures 6.1, 7.1 and 8.1 for the BRP, WWO and MPD respectively. In these diagrams, dispersion of organizational units of management of data vary between the three case studies.

Organizational units that perform management tasks of the BRP can be found at 35 locations. These locations are the thirty district offices of GAK, three regional service centres, the central service centre, and the automation main office a short distance from the central service centre.

Organizational units that perform management tasks of the WWO can be found at 4 locations: the three regional service centres, the central service centre and the automation main office. It must be taken into consideration that at the time the case study was performed, the data management organization was still under construction.

Organizational units that perform management tasks of the MPD can be found at three locations. Most of the management tasks are performed at the main office building at Moerdijk, management tasks of the two computer systems are performed via remote monitoring and control by Shell Combined Information Services (SCIS) at Rijswijk. The Laboratory Information System was supplied by an external software supplier (Delaware) which is involved in the tasks involved with performance analysis and tuning of data structures of LIS.

9.1.2 Decentralization

The data custodian determines what functionalities should be realized by the data and what automation facilities should be employed on behalf of these functionalities. The decentralization of the data environment is shown by the location of the custodian within the 'Mintzberg' diagram of the data environment, for this research it is not important if the environment is contained within a particular organizational boundary,

or crosses organizational boundaries. In figures 6.2, 7.2 and 8.2 for the BRP, WWO and MPD respectively, the data custodian is located at the position of the support staff within the 'Mintzberg' diagram.

The BRP custodian and the WWO custodian are located within the department Information and Automation (I&A) of GAK. The MPD custodian is located within the department Information and Organization (I&O) of Shell Moerdijk. In this respect there appears to be little difference between decentralization of data in an environment where data processing is part of the primary production processes, and an environment where data processing is not. This means that differences in decentralization had no effect in the case studies and decentralization is therefore not considered to be a factor in evaluation of the management instrumentation in the sections below.

9.2 Research findings

Using the research findings cross case conclusions are drawn and the specific aspects of the case studies are taken into account. Each finding is preceded by general statements about the practical applicability of the framework and the methods and techniques that together constitute the instrumentation for data management. After the general statements, statements about case studies are made.

9.2.1 The FATO Framework

Functionalities

The functionalities of data are managed using characteristics at the pragmatic and semantics levels of data. In general, the pragmatic and semantic characteristics proved to be valuable for describing and analyzing requirements regarding data functionalities. The case studies yielded no new characteristics regarding functionalities that had not been incorporated into the framework. All the data characteristics at the pragmatic level are used for defining data requirements, but actual data properties are hard to establish. Fewer characteristics are used for defining requirements at the semantic level of data, and actual data properties do not meet these requirements very well.

Shell Moerdijk production processes provide a stable environment for the MPD, and as a consequence the functional requirements and preconditions of the MPD change little over time. These requirements are strict due to the real time nature of the MPD, but they remain stable over long periods of time and therefore require less attention from data management. The environment of the BRP is less stable, newly built and upgraded information systems of GAK are attached to the BRP and increasingly stakeholders from outside the GAK organization make use of the BRP. This leads to more and more complex requirements and preconditions for the BRP. The WWO is a newly introduced information system, and less requirements and preconditions and characteristics have been defined compared to the BRP. The

actual data properties of the WWO are in accordance with the requirements, this can be attributed to the fact that it is a more recent system.

Automation Facilities

The automation facilities of data are managed using characteristics at the syntactic and empiric levels of data. In general, the syntactic and empiric characteristics proved to be valuable in describing and analyzing requirements regarding automation facilities. The case studies yielded no new characteristics regarding automation facilities that had not been incorporated into the framework. Both the syntactic characteristics and empiric characteristics were little used in defining the data requirements, but of those characteristics that were used, the actual properties were easily measurable and they were in accordance with the requirements.

The stability of the Shell Moerdijk Production processes also leads to relatively little change in the technical requirements and preconditions of the MPD. The real time database is carefully tuned to process all the required data in time, the highly automated nature of the production processes is a benefit in this respect. The BRP is database that has grown out of its original technical specifications, the distribution principle and the increasing load on the database contribute to increasing difficulties in maintaining the properties of the characteristics in accordance with the requirements and preconditions. The (31) WWO databases are not fully operational and this shows in the presence of just a few requirements and precondition of characteristics at the syntactic and empiric level. It appears that such requirements and preconditions are not laid down until a system is exploited by technical data management.

Task Fields

Tasks of data management are clustered into task fields based upon the levels of data: pragmatic, semantic, syntactic or empiric. Thus the task field 'actual data management' is concerned with data characteristics at the pragmatic level, the task field 'data administration' is concerned with data characteristics at the semantic level, the task field 'database administration' is concerned with data characteristics at the syntactic level, and the task field 'storage management' is concerned with data characteristics at the empiric level. The case studies yielded no new management tasks that had not been incorporated into the framework. Tasks that were performed within data management are linked with the requirements and characteristics of data. Not all requirements and characteristics at the pragmatic and semantic levels of data are employed in the management tasks of the task fields: actual data management and data administration. Those requirements and characteristics at the syntactic and empiric levels of data that *are* defined within data management, are employed in the management tasks of the appropriate task fields: database administration and storage management.

Management of the MPD consists of few tasks, but because of the stable environment of the MPD and the few requirements and preconditions made of the MPD,

9.2 RESEARCH FINDINGS

255

these few tasks suffice. The management of the BRP is more elaborate with tasks being performed at many different places in the GAK organization. Tasks performed within the task field 'actual data management' are laid down in handbooks, other tasks are left to individual functionaries involved with the management of BRP. The management task of the WWO are deduced from management task performed on behalf of the comparable WAO⁴ information system. The WAO information system is comparable with the WWO information system, both in requirements and preconditions and in technical realization with relational databases for different district offices of GAK. Despite the fact that the WAO information is fully operational, management tasks are not documented and are left to the personal judgement of individual functionaries.

Organization

Organizational units involved in performing data management tasks are separated into two forms of management: functional management and technical management. This separation of functions in the organization of data management is also found in the case studies. Sometimes the allocation of tasks is not in accordance with the FATO framework. These tasks were also identified as problematic by the data managers in the case studies, and improvements were made by bringing the allocation of tasks into accordance with the framework.

The management organization of the MPD consists of just a few people, where a strict separation of functions would be counter productive and functional and technical management are not strictly separate units. Separate organizational units are however used for the specific tasks of optimization and tuning of the LIS database (by Delaware) and for the task field 'storage management' of both the HP computer systems and the IBM computer systems of the MPD (by SCIS Rijswijk). The management organization of BRP consists of many people and organizational units, and here strict separation of functions is employed. Organizational units are dedicated to functional management or technical management. A few discrepancies were found in the allocation of tasks that were technical in nature, but were performed by a functionary within functional management. The management organization of the WWO had not yet been established but was considered in the case study to be identical to the management organization of the comparable WAO information system. The responsibilities, authorizations, work and expertise involved in performing management tasks are not allocated within the same organizational unit, and this leads to reduced effectiveness when performing management tasks.

Research finding 1 *The case studies did not provide elements that were not already defined in the FATO framework, nor did the case studies provide refutations of elements defined in the FATO framework.*

⁴In Dutch: Wet op de Arbeidsongeschiktheid (WAO)

9.2.2 Methods and techniques

The methods and techniques described in this thesis proved to be a valuable tool to specify how data management should be performed and to identify improvements in the case studies BRP and WWO. The case study MPD was used to identify any specific problems when using the methods and techniques in a non-administrative organization, where a real-time database was used. In none of the case studies were difficulties or particular differences found that could lead to a less effective use of the methods and techniques, with two exceptions where thorough testing of the methods and techniques was not possible. These exceptions were:

- It was not possible to establish the relationships between the stakeholders in the data environment, the requirements made on data by the environment, and the management tasks that are performed on behalf of these requirements and the environment. The 'table of stakeholders' (see section 5.3.4) was meant to describe these relationships, but it proved difficult for people involved in data management to identify these relationships.
- It was not possible to provide accurate descriptions of coordination between different management tasks and to validate such descriptions with the people involved with performing management tasks. Such coordination is achieved in an 'ad hoc' fashion, and further methods and techniques to describe and analyze it are required.

Research finding 2 *The methods and techniques provided by this thesis are valuable for specification of the functions to be performed by data management and identify possibilities for improvement.*

9.2.3 Applicability of the instrumentation

The management instrumentation proposed in this thesis consists of a FATO framework and several methods and techniques. The FATO framework is perfected into a generic model which provides a complete and concise definition of the operational level of data management. The methods and techniques which are traditionally used during planning and development are translated to meet the specific needs of utilization and exploitation. The instrumentation is used in the case studies to model the observations made in three specific situations. Any other organization can take the FATO framework and the methods and techniques to model its own specific situation.

The management instrumentation supports personnel at the operational level of data management to operate, control and maintain data in accordance with the specified requirements and preconditions. The instrumentation enables specification of activities which are performed within data management tasks. The instrumentation also enables the identification of the various stakeholders of the data, which requirements and preconditions are made by each individual stakeholder, and which management activity is performed to meet these requirements and preconditions. Data

9.3 ANSWERS TO THE RESEARCH QUESTIONS

257

management personnel is able to take appropriate actions on behalf of the stakeholders, and to assess the implications of changes in the stakeholders, requirements and preconditions, or the activities within the management tasks.

As the operational level of data management can now be modelled more accurately, the instrumentation can be used to provide the tactical level of data management with a means for diagnosis of data management operations. The analysts in the techno-structure can use the instrumentation to provide quality assurance, and the support staff can use the instrumentation to provide accurate costing of data management. The instrumentation also enables middle line managers to provide directions for the operational level, and to obtain the feed back information they require.

Diagnosis of the operational level can be performed by following the same approach as the case studies Personal Data Register (BRP) and Unemployment Benefit (WWO) described in chapters 6 and 7 (see figures 6.5 and 7.5). The general model is applied in the specific situation by using the methods and techniques to describe the operational level of data following the four 'FATO' phases as was discussed in section 1.4.6 concerning the case study design. These descriptions are compared with the arrangements prescribed in the general model to identify possibilities for improvement of the specific situation. The improvements of the operational level of data management can be concerned with three aspects of the general model. First, in the specific situation the operational level of data management is implemented in a way that is not according to the placement of data management aspects of the FATO framework. Second, in the specific situation the relationships between the environment (the real system), the data, and data management are not implemented in the manner prescribed by the general model (and the data management paradigm). And third, the way in which certain parts of the FATO framework are implemented leave much to be desired; for example, the requirements and preconditions on data characteristics are poorly specified or are poorly met by the actual properties of the data characteristics, or the activities within certain data management tasks are not correctly implemented. Using the methods and techniques of the management instrumentation, alternatives can be proposed and analyzed before being implemented.

9.3 Answers to the research questions

The answers to the research questions are qualified with respect to the observations made within the case studies. Cross case analysis takes into account the differences and similarities of the individual case studies to ensure the non specific nature of the statements.

9.3.1 Characteristics of data

The first research question addressed the practical applicability of the characteristics defined within the FATO framework. These characteristics are positioned by

the FATO framework at four levels of data: the pragmatic level, the semantic level concerned with functionalities (F) of data, the syntactic level and the empiric level concerned with the automation facilities (A) of data. Data management tasks (T) were positioned within four related task fields: actual data management, data administration, database administration, and storage management. These task fields were then positioned within two organizations (O): functional management and technical management. The FATO approach was first introduced by Looijen [58] on behalf of exploitation of automation facilities in computing centres, but has been extended and enhanced in the field of data management by this research. Of all theoretically possible characteristics, maybe just a few may be of interest for performing management in practice, the first research question repeated below concerns this issue:

What characteristics of data are of interest for performing tasks of managing distributed data in distributed environments?

The research findings show a clear difference in this respect, between functional data management and technical data management.

Functional data management follows to some extent the characteristics of data defined at the pragmatic and semantic level (see table 4.2). The case studies Personal Data Register (BRP) and Unemployment Benefit (WWO) show that all of the pragmatic data characteristics can be used to express the data requirements and preconditions. In contrast, the only pragmatic requirement and precondition made of the Moerdijk Process Database (MPD) concerned timeliness of data. The Moerdijk Process Database (MPD) is a real time database that receives data from automated production processes (PLC's, Data loggers etc.) that indicate suspect values when supplied to the MPD, when data is not supplied by a particular (e.g.) PLC to the MPD within the designated time slot, it is lost. The high quality of data supply requires less management activities to maintain the pragmatic characteristics and therefore less requirements and preconditions are made. In all the case studies functional management had difficulties in ascertaining that the actual properties of the pragmatic data characteristics matched the requirements and preconditions.

Semantic data characteristics are used to a lesser extent to express data requirements and preconditions. Within the BRP case all but the characteristics 'modification integrity' and 'securability' were used, but in the case of WWO only the requirements and preconditions of the characteristic 'convertibility' were defined. The MPD case showed an intermediate result.

Technical management of data hardly follows the data characteristics defined at the syntactic and empiric level (see table 4.2). The BRP case focussed on functional management, but investigation of the Service Level Agreement (SLA) revealed that the only technical requirements of the BRP concern 'storage capacity' and 'storage costs'. Discussions with technical managers of BRP indicated that they only monitor the characteristics 'load factor' and 'storage capacity'. The technical management of the WWO was investigated extensively, but also here, just a few syntactica and empiric characteristics were used. Requirements were made only for 'transformation' and

9.3 ANSWERS TO THE RESEARCH QUESTIONS

259

‘repairability’, ‘recoverability’ and ‘safety’. Properties of additional data characteristics were measured, namely: ‘access speed’, ‘data capacity’, ‘accuracy’ and ‘storage capacity’. Interviews with technical managers of the MPD showed a similar picture.

Research answer 1 *The characteristics of data defined in chapter 4 are valuable for functional management, data characteristics are less valuable for technical management.*

9.3.2 Methods and techniques

The second research question addresses the practical applicability of the methods and techniques developed in chapter 5. These methods and techniques were selected from BSP and ISAC and they describe data management at the systelological level. Other methods exist for network management; the Telecommunications Management Network (TMN) model produced by CCITT[43], and computing centre management; the IT infrastructure library (ITIL) produced by CCTA [15], but these methods are limited to tasks and organization of management of specific automation facilities and do not concern data management. They do not provide the techniques necessary to document and analyze data management at the systelological level. The methods and techniques described in chapter 5 were originally meant for development of information systems, and were modified to suit the specific needs of data management. Whether these modifications are sufficient to make them suited for data management is the subject of the second research question repeated below:

What methods and techniques can be made available to management to operate, control and maintain distributed data in accordance to the requirements and preconditions of a distributed environment and the characteristics of the distributed data?

The methods and techniques were used extensively to describe and analyze management of data in the case studies BRP and WWO. In both case studies there was little difficulty in making the models using the proposed methods and techniques. These models proved to be useful to provide suggestions for improvements in management of both the BRP and WWO. The two exceptions were the table of stakeholders and coordination mechanisms:

- The table of stakeholders is hard to make when relationships between stakeholders (i.e. the environment), requirements and preconditions for data and management tasks are unknown.
- There are no models available for coordination mechanisms, we have to rely on verbal descriptions. Direct supervision can be shown using organization charts (or organigram).

Research answer 2 *The methods and techniques described in chapter 5 are valuable for management of distributed data in distributed environments.*

9.4 Future research

The research findings and the answers to the research questions show possible enhancements of the instrumentation for management of distributed data in distributed environments. Possibilities to improve the management instrumentation are discussed in the first section. Ways to extend the instrumentation are discussed in three sections concerning extending the instrumentation in the state model, coping with external influences and constraints for management of data, and finally with internal influences between different components of data management.

9.4.1 Improving the instrumentation

The research findings and answers to research questions showed how the three case studies gave no cause to provide additions to the FATO framework. The results of the case studies showed how the FATO framework has defined characteristics and management task that do not appear in the case studies. The FATO framework appears in this respect to be quite complete for use within the states Utilization and Exploitation (see sections 1.2.2 and 1.3.2).

The methods and techniques to complete the management instrumentation were chosen from BSP and ISAC to describe management of data at a systelological level. Two limitations of methods and techniques were described in section 9.3.2. Future research should focus on new models to describe the relationships between stakeholders, requirements and preconditions for data and management tasks, and on models for coordination between units of a (distributed) management organization. The improved methods and techniques for data management should be closely allied with the methods and techniques in use for development.

Future research should further include the infological, datalogical and technological levels of management of distributed data in distributed environments. These levels are more suited to the use of more formal techniques as found in operational research and the use of simulation techniques. Such formal methods and techniques are required for the proper implementation of automated systems to support data management tasks, e.g. intelligent data dictionary systems.

9.4.2 Beyond the steady state

The FATO framework concentrates on task fields and characteristics of data in the states Utilization (U) and Exploitation (E) of the state model shown in figure 1.2. Future research should include the state Maintenance (M). Data characteristics related to maintenance, management tasks performed in the state Maintenance (M), and the project nature of the organization are all different. These differences effect the complete FATO framework when applied to data management in the state Maintenance (M). The extended state model shown in figure 1.3 shows how M2 maintenance, of a innovative nature, will lead to new steady states of Utilization (U') and Exploitation

9.4 FUTURE RESEARCH

261

(E'). The transition of data management into these new states should be the subject of future research.

9.4.3 Internal influences

Internal influences are those influences that occur within data management, however, figure 4.2 shows how, as well as external influences, internal influences exist within the paradigm for data management. These relationships between the data environment, data and management are addressed by the FATO framework, but there are internal influences within data and management that should be the subject of future research. Examples include the relationships between the data characteristics and relationships between data management tasks. These relationships are situationally dependent and can be modelled using simulation techniques such as were used by De Wijs [109].

9.4.4 External influences

External influences are those influences that enable new possibilities, or impose constraints upon organizing data management and that enhance or limit the possibilities for implementing its task fields. The research presented in this thesis was conducted in a more or less "independent" situation where data management could be implemented without external influences, based solely upon the user requirements and preconditions and the characteristics of the object of management.

During the interviews, those involved with data management indicated they had not implemented parts of the management instrumentation proposed in this thesis due mainly to organization and economic constraints. Six external influences of data management were identified in section 4.4.1. Besides the economic constraints, future researchers should investigate data management with respect to 'interorganizational information systems' (IIS) such as described by Wierda [108], where organizational boundaries are crossed and organizational influences are of importance to data management. Future research should also be linked to research into international networks, currently undertaken by Van den Broek [9], and investigate data management with respect to 'international' information systems, where national boundaries are crossed. Social and political influences, and in particular legal influences, are of importance to data management.

Bibliography

- [1] Jan Achterberg. *Informatiemanagement*. PhD thesis, Vrije Universiteit Amsterdam, 1986. (in Dutch).
- [2] J.A. Bakker. Inleiding databeheer. Lecture notes a315, Delft University of Technology, 1992. (in Dutch).
- [3] J.J.A. Bakker. *Operationele informatieverzorging*. H.E. Stenfert Kroese B.V., Leiden, 1987. (in Dutch).
- [4] J.H. ter Bekke. *Semantic data modelling*. Prentice-Hall, 1992.
- [5] M.E. van Biene-Hershey. *Auditing the Technical EDP Organization*. NGI Sectie EDP-Auditing, Paulus Potterstraat 40, 1071 DB Amsterdam, 1985.
- [6] A. van den Bogaart. Gegevensbeheer in decentrale organisaties. Master's thesis, Katholic University of Nijmegen, May 1992. (in Dutch).
- [7] C. Brevoord. *De informatiehuishouding*. Kluwer Bedrijfs wetenschappen, Deventer, 1991. (in Dutch).
- [8] C. Brevoord and H. Gorter de Vries. *Externe administratieve integratie*. Stenfert Kroese, Leiden, 1981. (in Dutch).
- [9] Floris van den Broek. Het beheer van internationale netwerken. Research proposal, 1994. (in Dutch).
- [10] J. Bruijns. Eigendom en beheer van informatiesystemen. *Informatie*, 21(2):77–84, 1979. (in Dutch).
- [11] Bas K. Brussaard and Peter A. Tas. Information and organization policies in public administration. In S.H. Lavington, editor, *Information Processing*, pages 821–826. North-Holland Publishing Company, 1980.
- [12] A. Burns, J. McDermid, and J. Dobson. On the meaning of safety and security. *The computer journal*, 35(1):3–15, 1992.

- [13] R. Bushoff and J.A. Oosterhaven. Information strategy planning. *Informatie*, 29(3):228–238, 1987.
- [14] Pei Cao, Swee Boon Lim, Shivakumar Venkataraman, and John Wilkes. The TickerTAIP parallel RAID architecture. Technical Report HPL-OSR-93-25, Hewlett-Packard Laboratories, 1993. (also published as HPL-92-151).
- [15] Central Computer and Telecommunications Agency. *IT Infrastructure Library*. Central Computer and Telecommunications Agency, Norwich, England, 1990.
- [16] Central Computer and Telecommunications Agency. *IT Infrastructure Library, an introduction*. Central Computer and Telecommunications Agency, Norwich, England, 1992.
- [17] S. Ceri and G. Pelagatti. *Distributed databases, principles and systems*. McGraw-Hill, 1984.
- [18] Micheal J. Cerullo and Virginia Cerullo. Approaches to organizing computer facilities. *Information Age*, 11(3):165–169, July 1989.
- [19] T.C.E. Cheng and V. Kanabar. On some issues of information resource management in the 1990s. *Information Resources Management Journal*, 5(1):21–33, Winter 1992.
- [20] R. Clark, S. Holloway, and W. List. *The security, audit and control of databases*. Avebury Technical, Aldershot, England, 1991.
- [21] C.J. Date. Distributed databases: a closer look. In *Relational Database Writings 1989–1991* [22], chapter 33, pages 485–516.
- [22] C.J. Date. *Relational Database Writings 1989–1991*. Addison-Wesley, 1991.
- [23] C.J. Date. *An introduction to Database Systems*, volume 2. Addison-Wesley, 1985.
- [24] C.J. Date. *An introduction to Database Systems*, volume 1. Addison-Wesley, 5th edition, 1990.
- [25] E. Dee. Distributed database theory. Information Systems Engineering Paper P2, Central Computer and Telecommunications Agency, November 1988.
- [26] G.P.A.J. Delen and M. Looijen. *Beheer van informatievoorziening*. SDM-reeks. Cap Gemini Publishing, P.O. Box 3027, 2280 GA Rijswijk, The Netherlands, 1992. (in Dutch).
- [27] G.P.A.J. Delen and D.B.B. Rijsenbrij. Kwaliteitsattributen van automatiseringsprojecten in informatiesystemen. *Informatie*, 32(1):46–55, 1990. (in Dutch).

BIBLIOGRAPHY

265

- [28] Cory Devor, Ramez Elmasri, James A. Larson, Saeed Rahimi, and James P. Richardson. Five-schema architecture extends DBMS to distributed applications. In *Tutorial: Distributed database management* [57], pages 606–612.
- [29] Digital Equipment Corporation. *VAX DBMS database design guide*, September 1989.
- [30] W. Durell. *Data administration*. McGraw-Hill, 1985.
- [31] G.M. Er. Is a database a corporate resource? *Information Age*, 11(2):106–111, April 1989.
- [32] E.K.C. Esseling and H. van Nimwegen. *Administratieve processen, aanpak en technieken t.b.v. vastlegging, analyse en ontwerp*. Kluwer Bedrijfswetenschappen, Deventer, 1989. (in Dutch).
- [33] E.B. Fernandez, R.C. Summers, and C. Wood. *Database Security and Integrity*. Addison-Wesley Publishing Company Inc., Reading, Massachusetts, 1981.
- [34] Gert Florijn, Arthur ter Hofstede, Norbert van Oosterom, Gerard Wijers, Cees van Miltenburg, Guus Vreven, and Anne van den Dool. Gegevensdefinitiebeheer in een gedecentraliseerde organisatie. Technical Report 91/09, SERC, August 1991. (in Dutch).
- [35] Bas C. van Fraasen. *The scientific image*. Clarendon Press, Oxford, 1980.
- [36] S.B. Geller. Care and handling of computer magnetic storage media. Technical Report 500-101, National Institute of Standards and Technology, Gaithersburg, MD 20899, June 1983.
- [37] Gemeenschappelijk Administratie Kantoor. Verslag over het boekjaar 1953. Annual report, Gemeenschappelijk Administratie Kantoor, 1954. (in Dutch).
- [38] Gemeenschappelijk Administratie Kantoor. Verslag over het boekjaar 1991. Annual report, Gemeenschappelijk Administratie Kantoor, 1992. (in Dutch).
- [39] A.J. van de Goor. *Computer architecture and design*. Electronic systems engineering series. Addison-Wesley Publishing Company, 1989.
- [40] A.J. van de Goor. *Testing semiconductor memories*. John Wiley & Sons, 1991.
- [41] A.J. Groen. Beheer van gedistribueerde gegevens bij het GAK. Master's thesis, Delft University of Technology, August 1994. (in Dutch).
- [42] V. Grover and J.T.C. Teng. How effective is data resource management? *Journal of Information Systems Management*, 8(3):16–20, Summer 1991.

- [43] L. van Hemmen. TMN-model, beschrijving en gebruik voor SIG-RNS. Technical report, RCC Apeldoorn, afdeling RNS, P.O. Box 9105, 7300 HN, Apeldoorn, March 1994. (in Dutch).
- [44] A.E. Hoagland and J.E. Monson. *Digital magnetic recording*. John Wiley & Sons, Inc., 2nd edition, 1991.
- [45] B.A.A. Hopstaken and A. Kranendonk. *Informatieplanning, puzzelen met beleid en plan*. Stenfert Kroese, Leiden, 1988. (in Dutch).
- [46] G.E. Houtekamer. *Measuring and modelling disk I/O subsystems*. PhD thesis, Delft University of Technology, Stevinweg 1, 2628CN Delft, 1989.
- [47] G.E. Houtekamer and H.P. Artis. *MVS I/O subsystems*. J. Renade IBM series. McGraw-Hill, Inc., 1991.
- [48] J.J.P. van Hulzen and R.P. de Moel. Gedrags-/prestatiekaracteristieken van informatievoorziening. *Informatie*, 29(3):254–261, 1987. (in Dutch).
- [49] D.S. Hussain and K.M. Hussain. *Information management; organisation, management and control of computer processing*. Prentice-Hall, Inc., 1992.
- [50] IBM. Planning for distributed information systems. Technical Report GE20-0655-1, International Business Machine Corporation.
- [51] Ivan F. Jackson. *Corporate Information Management*. Prentice-Hall, 1986.
- [52] E.O.J. Jans. *Informatieverzorging, zorgen en verzorgen*. Samson, Alphen aan den Rijn, 1979. (in Dutch).
- [53] E.O.J. Jans. *Grondslagen van de administratieve organisatie*. Samson, Alphen aan den Rijn, 14th edition, 1987. (in Dutch).
- [54] R.H. Johnson and R.D. Johnson. *DASD: IBM's direct access storage devices*. J. Renade IBM series. McGraw-Hill, inc., 1992.
- [55] Wiebren de Jonge and Ardie Schijf. B-bomen kritisch bekeken. *Informatie*, 32(5):455–467, 1990. (in Dutch).
- [56] Börje Langefors and Bo Sundgren. *Information systems architecture*. Mason/Charter Publishers, Inc., 1975.
- [57] James A. Larson and Saeed Rahimi. *Tutorial: Distributed database management*. IEEE Computer Society Press, Silver Spring, MD, 1985.
- [58] M. Looijen. *Management en organisatie van automatiseringsmiddelen*. Kluwer Bedrijfswetenschappen, Deventer, 1989. (in Dutch).

BIBLIOGRAPHY

267

- [59] M. Looijen. *EBM - Een Beheer Methodiek*. Lansa Publishing, P.O. Box 176, 2260 AD Leidschendam, The Netherlands, 1991. (in Dutch).
- [60] M. Looijen. *Beheer van informatiesystemen*. Kluwer Bedrijfswetenschappen, Deventer, 1995. (in Dutch).
- [61] M.B. Louweret. Het spel om de gegevens: een analyse van basisregistraties en macht. *Informatie en Informatiebeleid*, 10(4):36–42, winter 1992. (in Dutch).
- [62] M. Lundeborg, G. Goldkuhl, and A. Nilsson. *De ISAC-methodiek*. Automatische informatieverwerking. Samson Uitgeverij, Alphen aan den Rijn, 2nd edition, 1987. (in Dutch).
- [63] C.E. Mackenzie. *Coded character sets, history and development*. Addison-Wesley, 1990.
- [64] R.N. Maddison. *Information system methodologies*. Wiley Heyden Ltd., 1986.
- [65] S.T. March. Information resource management: integrating the pieces. *Database*, 32(Summer):27–38, 1992.
- [66] J. Martin. *Strategic data-planning methodologies*. Prentice-Hall, Inc., 1982.
- [67] J. Martin. *Managing the data-base environment*. Prentice-Hall, Inc., 1983.
- [68] C. Dennis Mee and Eric D. Daniel, editors. *Computer data storage*, volume 2 of *Magnetic recording*. McGraw-Hill, inc., 1988.
- [69] R. Mercken. De relatie tussen gegevensbeheer en interne audit. *Management en organisatie van automatiseringsmiddelen*, (2):22–37, 1993. (in Dutch).
- [70] R.J. Mersel. Rapportage stuurgroep GAK verkennend onderzoek gegevensbeheer. Internal report, Gemeenschappelijk Administratie Kantoor, 1992. (in Dutch).
- [71] R.J. Mersel. Rapport functioneel beheer van de basisregistratie personen (BRP). Internal report, Gemeenschappelijk Administratie Kantoor, 1994. (in Dutch).
- [72] R.J. Mersel and J. de Boer. Gedistribueerde databases in de praktijk: De BasisRegistratie Personen bij het GAK. *IT Management Select*, 1(1):58–67, 1995. (in Dutch).
- [73] Henry Mintzberg. *Structures in fives: designing effective organisations*. Prentice Hall, Englewood Cliffs, 1983.
- [74] R.P. de Moel and J.J.P. van Hulzen. De weg van eisen naar eigenschappen voor informatievoorziening. *Informatie*, 30(2):141–151, 1988. (in Dutch).

- [75] K.IJ. Mollema. Informatiekwaliteit en EDP-audit. *Informatie*, 33(7/8):482–485, 1991. (in Dutch).
- [76] P.B. Morssink and A. Kranendonk. *De voorkant van het automatiseren*. H.E. Stenfert Kroese B.V., Leiden/Antwerpen, 5th edition, 1985. (in Dutch).
- [77] NGI-commissie beroepsontwikkeling. *Taken en functies in de informatica: een voorstel tot ordening*. Kluwer Bedrijfswetenschappen, Deventer, 1993. (in Dutch).
- [78] W.P.E. Notermans. AAW-Voorzieningen. B.Sc. graduation report, Hogeschool Heerlen, 1993. (in Dutch).
- [79] Patrick D.T. O’Conner. *Practical reliability engineering*. John Wiley & Sons Ltd., 1985.
- [80] T.W. Olle, H.G. Sol, and A.A. Verrijn Stuart, editors. *Information design methodologies: a comparative review*, Amsterdam, 1982. North-Holland.
- [81] N. van Oosterom, G. Florijn, A. ter Hofstede, G. Wijers, C. van Miltenburg, A. van den Dool, and G. Vreven. Gegevensdefinitiebeheer in een gedecentraliseerde organisatie. *Informatie*, 35(2):128–139, 1993. (in Dutch).
- [82] E. Ortner. Informationsmanagement: wie es entstand, was es its und wohin es sich entwickelt. *Informatik Spektrum*, (14):316–327, 1991. (in German).
- [83] M. Tamer Özsu and P. Valduriez. *Principles of distributed database systems*. Prentice-Hall, London, 1991.
- [84] H.J. Pels. *Gegevens: modellering, analyse en beheer*. Samson Bedrijfsinformatie, Alphen aan den Rijn, 1992. (in Dutch).
- [85] R. Planche. *Data driven systems modelling*. Prentice Hall International / Mas-son, 3rd edition, 1992.
- [86] B.R. Rosen and M.H. Law. Guide to data administration. Technical Report 500-173, National Institute of Standards and Technology, Gaithersburg, MD 20899, October 1989.
- [87] Chris Ruemmler and John Wilkes. Modelling disks. Technical Report HPL-93-68, Hewlett-Packard Laboratories, 1993.
- [88] M. Satyanarayanan. *Modelling storage systems*. Number 5 in Computer architecture and design. UMI research press, 1986.
- [89] E.A. Van Schaik. *A management system for the information business*. Prentice Hall, Inc., 1985.

BIBLIOGRAPHY

269

- [90] Shell Nederland Chemie B.V. Moerdijk. TIMMAS study. Moerdijk Rapport SCM 011/90, Shell Nederland Chemie B.V. Moerdijk, 1990.
- [91] Shell Nederland Chemie B.V. Moerdijk. AS Facility, global functional specification. Moerdijk Rapport SCM 012/91, Shell Nederland Chemie B.V. Moerdijk, 1991.
- [92] Shell Nederland Chemie B.V. Moerdijk. Projectdossier AS Facility. Internal report, Shell Nederland Chemie B.V. Moerdijk, 1991.
- [93] Shell Nederland Chemie B.V. Moerdijk. Technical documentation AS Facility. Internal report, Shell Nederland Chemie B.V. Moerdijk, 1992.
- [94] H.G. Sol. Kennis en ervaring rond het ontwerpen van informatiesystemen. *Informatie*, 27(3):184–192, 1985. (in Dutch).
- [95] Ralph H. Sprague, Jr. and Barbera C. McNurlin. *Information systems management in practice*. Prentice-Hall international, Inc., 1986.
- [96] R. Stamper. *Information in business and administrative systems*. B.T. Batsford, London, 1973.
- [97] R.W. Starreveld, E.J. Joëls, and H.B. de Mare. *Bestuurlijke informatieverzorging*, volume 1. Samson, Alphen aan den Rijn, 2nd edition, 1985. (in Dutch).
- [98] R.J. Streng. *Dynamic modelling to assess the value of electronic data interchange*. PhD thesis, Delft University of Technology, P.O. Box 356, 2600 AJ Delft, The Netherlands, 1993.
- [99] Toby J. Teorey. *Design of database structures*. Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1982.
- [100] A.L. Tharp. *File organization and processing*. John Wiley & Sons, 1988.
- [101] J. in 't Veld. Produktiestructuren en informatiebehoeften. *Informatie*, 27(6):554–559, June 1985. (in Dutch).
- [102] E. Veldhuizen. Data beheer en controle bij geavanceerde informatiesystemen. In R. Paans and R.A.C. Thomas, editors, *Data: Beheer en controle*, pages 522–542, Paulus Potterstraat 40, 1071 DB Amsterdam, The Netherlands, May 1985. Nederlands Genootschap voor Informatica (NGI), NGI Sectie EDP-Auditing. (in Dutch).
- [103] H. Völlmar. Eigendom en beheer van gegevens. *Informatie*, 23(6):381–386, 1981. (in Dutch).
- [104] H. Völlmar. De gegevensbankbeheerder. *Informatie*, 24(1):43–48, 1982. (in Dutch).

- [105] Ria van Waes. *Architectures for information management*. PhD thesis, Free University of Amsterdam, Amsterdam, 1991.
- [106] H.M. Walker. Administering a distributed data base management system. *ACM SIGMOD Record*, 12(3):86–99, April 1982.
- [107] R.J. Welke. Current information system analysis and design approaches: framework, overview, comments en conclusions for large complex information system education. In R.A. Buckingham, editor, *Education and Large Information Systems*, pages 149–167, The Hague, Netherlands, April 1977. IFIP TC3/TC8, North Holland, Amsterdam.
- [108] Folkert W. Wierda. *Developing interorganizational information systems*. PhD thesis, Delft University of Technology, P.O. Box 356, 2600 AJ Delft, The Netherlands, 1991.
- [109] C. de Wijs. *Information systems management in complex organizations*. PhD thesis, Delft University of Technology, June 1995.
- [110] M.P. Williamson. The 3480 type tape cartridge: potential data storage risks, and care and handling procedures to minimize risks. Technical Report 500-199, National Institute of Standards and Technology, Gaithersburg, MD 20899, November 1991.
- [111] Robert K. Yin. *Case study research*, volume 5 of *Applied Social Research Methods*. SAGE Publications Ltd., 28 Banner Street, London EC1Y 8QE, England, 1st edition, 1984.
- [112] P.J. Zeldenrust. Op weg naar gedecentraliseerd functioneel beheer. Master's thesis, Delft University of Technology, August 1992. (in Dutch).
- [113] B. Zonneveld. Access manager op ICL computers. Master's thesis, Delft University of Technology, August 1993. (in Dutch).

Author Index

A

Achterberg, Jan 263
Artis, H.P. 266

B

Bakker, J.A. 263
Bakker, J.J.A. 263
Bekke, J.H. ter 263
Biene-Hershey, M.E. van 263
Boer, J. de 267
Bogaart, A. van den 263
Brevoord, C. 263
Broek, Floris van den 263
Bruijns, J. 263
Brussaard, Bas K. 263
Burns, A. 263
Bushoff, R. 264

C

Cao, Pei 264
Central Computer and
Telecommunications Agency 264
Ceri, S. 264
Cerullo, Micheal J. 264
Cerullo, Virginia 264
Chastain, Clark L. 268
Cheng, T.C.E. 264
Clark, R. 264

D

Daniel, Eric D. 267
Date, C.J. 264
Dee, E. 264
Delen, G.P.A.J. 264
Digital Equipment Corporation 265

Dobson, J. 263

Dool, A. van den 265, 268
Durell, W. 265

E

Er, G.M. 265
Esseling, E.K.C. 265

F

Fernandez, E.B. 265
Florijn, G. 265, 268
Fraasen, Bas C. van 265

G

Geller, S.B. 265
Gemeenschappelijk Administratie
Kantoor 265
Goldkuhl, G. 267
Goor, A.J. van de 265
Gorter de Vries, H. 263
Groen, A.J. 265
Grover, V. 265

H

Hemmen, L. van 266
Hoagland, A.E. 266
Hofstede, A. ter 265, 268
Holloway, S. 264
Hopstaken, B.A.A. 266
Houtekamer, G.E. 266
Hulzen, J.J.P. van 266, 267
Hussain, D.S. 266
Hussain, K.M. 266

272

AUTHOR INDEX

I

International Business Machine
Corporation 266

J

Jackson, Ivan F. 266
Jans, E.O.J. 266
Joëls, E.J. 269
Johnson, R.D. 266
Johnson, R.H. 266
Jonge, Wiebren de 266

K

Kanabar, V. 264
Kranendonk, A. 266, 268

L

Langefors, Börje 266
Larson, James A. 265, 266
Law, M.H. 268
Lim, Swee Boon 264
List, W. 264
Looijen, M. 264, 266, 267
Louweret, M.B. 267
Lundeberg, M. 267

M

Mackenzie, C.E. 267
Maddison, R.N. 267
March, S.T. 267
Mare, H.B. de 269
Martin, J. 267
McDermid, J. 263
McNurlin, Barbera C. 269
Mee, C. Dennis 267
Mercken, R. 267
Mersel, R.J. 267
Miltenburg, C. van 265, 268
Mintzberg, H. 267
Moel, R.P. de 266, 267
Mollema, K.IJ. 268
Monson, J.E. 266
Morssink, P.B. 268

N

NGI-commissie beroepsontwikkeling
268
Nilsson, A. 267
Nimwegen, H. van 265
Notermans, W.P.E. 268

O

O'Conner, Patrick D.T. 268
Olle, T.W. 268
Oosterhaven, J.A. 264
Oosterom, N. van 265, 268
Ortner, E. 268
Özsu, M. Tamer 268

P

Paans, R. 268, 269
Pelagatti, G. 264
Pels, H.J. 268
Planche, R. 268

R

Rahimi, Saeed 265, 266
Ram, Sudha 268
Rijsenbrij, D.B.B. 264
Rosen, B.R. 268
Rossen, G. van 268
Ruemmler, C. 269

S

Satyanarayanan, M. 269
Schaik, E.A. Van 269
Schijf, Ardie 266
Shell Nederland Chemie B.V. Moerdijk
269
Sol, H.G. 268, 269
Sprague, Ralph H. Jr. 269
Stamper, R. 269
Starreveld, R.W. 269
Streng, R.J. 269
Summers, R.C. 265
Sundgren, Bo 266

T

Tas, Peter A. 263

AUTHOR INDEX

273

Teng, J.T.C. 265
Teorey, Toby J. 269
Tharp, A.L. 269
Thomas, R.A.C. 268, 269

V

Valduriez, P. 268
Veld, J. in 't 269
Veldhuizen, E. 269
Venkataraman, Shivakumar 264
Verrijn Stuart, A.A. 268
Völlmar, H. 270
Vreven, G. 265, 268

W

Waes, Ria van 270
Walker, H.M. 270
Welke, R.J. 270
Wierda, Folkert W. 270
Wijers, G. 265, 268
Wijs, C. de 270
Wilkes, J. 264, 269
Williamson, M.P. 270
Wood, C. 265

Y

Yin, Robert K. 270

Z

Zeldenrust, P.J. 270
Zonneveld, B. 270

Subject Index

A

acceptance 22, 23, 52, 65, 68, 80, 235
access pattern 59, 118, 121, 125,
203, 235
access speed 53, 114, 115, 116, 124,
172, 175, 183, 201, 202, 215,
234, 259
access structures 53, 64, 65, 114–116,
124, 172, 173, 175, 201, 205,
234, 236, 237
access time 55, 118, 120, 125, 128,
173, 203, 235
accuracy 51–53, 59, 61, 67, 68, 73,
114, 117, 124, 201, 202, 215,
234, 259
actual data management 33, 107,
122, 122, 181, 236, 254, 255,
258
actuality 52, 53, 109, 110, 122, 159,
162, 171, 197–199, 231, 232
adequacy 27, 28, 30, 31, 32, 36, 51,
59, 60, 79, 143, 145, 146,
177, 178, 181, 189, 190, 209,
210, 213, 216, 217, 221, 240,
241, 249, 250
application management 24
auditability 51–53, 109, 111, 122,
159, 163, 197, 199, 217, 231,
233, 243, 245
authorization
distribution, 35, 141, 174, 177,
208, 240, 250

automation facility 20, 24, 25, 33–35,
48, 49, 53, 57, 59, 78, 79, 81,
92, 101, 107, 114, 122,
126–129, 131–134, 139, 142,
148, 167, 168, 172, 176, 177,
182, 184, 186, 191, 192, 204,
208, 214, 215, 217, 218, 224,
244, 246, 249, 250, 252, 254,
258, 259
deconcentration, 35, 98, 133, 146,
190, 222

C

characteristics 19–22, 25, 26, 31, 32,
36, 39, 52, 55, 60, 67, 85,
108, 109, 111, 113–115, 118,
122, 130, 136, 139, 140, 142,
145, 146, 153, 159, 163, 166,
172, 175, 181–184, 186, 222,
225, 230, 231, 233–235,
241–243, 245, 246, 250, 253,
254, 257–261
completeness 41, 51–53, 72, 109, 122,
159, 161, 197, 198, 231, 232
concurrency 53, 69, 114, 117, 124,
201, 234
confidentiality 51, 109, 111, 122,
159, 163, 184, 185, 197, 199,
217, 231, 233, 243, 245
convertibility 111, 112, 113, 123,
163, 164, 199, 200, 211, 233,
258
correctness 41, 51–53, 61, 109, 122,
159, 160, 197, 227, 231

276

SUBJECT INDEX

D

DASD *see* direct access storage device

data

decentralization, 19, 20, 35, 56, 77

deconcentration, 19, 19, 20, 56, 59, 84, 144, 146, 190, 222, 251

distribution, 19, 20, 25–28, 30–37, 39, 47, 56, 57, 60, 62, 63, 65, 69, 74, 79–85, 91, 95, 101–105, 107, 108, 135, 136, 139, 143, 145, 146, 159, 167, 178, 180, 181, 184, 189, 190, 193, 210, 213, 215, 222, 241, 244, 249–251, 258–260

data administration 33, 69, 73, 76, 78, 107, 122, 123, 177, 209, 237, 246, 254, 258

data capacity 114, 115, 124, 172, 175, 183, 201, 202, 212, 215, 234, 259

data management 25

database 20, 25, 42, 47, 48
deconcentration, 56, 58
distribution, 56–58, 60, 62–64, 66, 116, 122

database administration 33, 51, 69, 72, 107, 122, 124, 124, 177, 209, 236, 246, 254, 258

database administrator 67, 70–72, 237, 239, 240, 246

database management system
distribution, 62, 63, 67, 135, 168

database system
distribution, 30, 56

DBA *see* database administrator

decentralization 34, 45, 59, 67, 75, 76, 78, 81, 100, 126, 131

horizontal, 46

parallel, 46

selective, 46

vertical, 46

deconcentration 45, 59, 75, 76, 126, 133, 134, 146, 222

development 19–21, 22, 49–53, 59, 65–67, 72, 74, 79–81, 83, 91, 99, 101, 102, 108, 127, 131, 135–139, 143, 163, 168, 175, 176, 184, 187, 205, 206, 208, 213, 214, 216, 219, 226, 256, 259, 260

direct access storage device 114, 118, 120, 125, 126, 173, 175, 176, 203, 206, 208, 215

distribution 65, 75, 79, 81, 99, 102, 143, 190

business processes, 138

exponential, 119, 120

mutations, 96, 154, 155, 159, 162, 170, 171

query processing, 62

RAWE, 140, 141, 175, 187, 204, 218, 236, 244

retrieval, 159, 172

transaction processing, 62, 65

domain integrity 53, 111, 113, 123, 163, 166, 199, 200, 233, 234

durability 109, 110, 122, 159, 162, 197, 199, 201, 231, 232, 236

E

empirics 33, 48, 107, 114, 118, 122, 125, 132, 136, 140, 142, 159, 172, 173, 182, 184, 186, 197, 203, 211, 212, 214, 215, 217, 231, 235, 242–244, 254, 258

environment 26, 31–33, 34, 37, 39–41, 44, 45, 55, 60, 79, 80, 83, 84, 90, 92, 99, 101–103, 122, 129, 130, 137, 142, 149, 179, 180, 186, 189, 192, 210, 211, 218, 221, 223, 224, 241, 246, 253, 254, 256, 257, 259, 261

SUBJECT INDEX

277

decentralization, 34, 77, 80, 144,
146, 148, 149, 190–192, 222,
223, 251–253

deconcentration, 34, 80

distribution, 25–28, 30–37, 39,
40, 44, 45, 79–85, 90, 91,
101–105, 107, 108, 133, 135,
136, 142, 143, 145, 146, 179,
181, 184, 190, 210, 213, 215,
222, 241, 244, 249, 250,
258–260

error correction 54, 55, 114, *118*,
124, 183, 201, 202, 215, 234,
235

error detection 55, 114, *117*, 124,
175, 183, 201, 234

expertise

distribution, 35, 141, 174, 177,
208, 240, 250

exploitation 19, 21, 22, *23*, 24–27,
35, 50, 65, 67, 69–71, 78–80,
101, 110, 132, 133, 135–137,
250, 251, 254, 256, 258, 260

F

FATO 33, 35, 37, 81, 104, *107*, 108,
122, 126, 132, 136, 137,
142–144, 146, 182, 187, 190,
213, 218, 222, 241, 249, 250,
255–258, 260, 261

fill factor 118, *121*, 125, 175, 203,
206, 235

functional

decentralization, 45

distribution, 35, 102, 132, *133*,
138, 142, 146, 149, 155, 185,
190, 193, 196, 222, 229, 250,
251

functional management *24*, 27, 78,
99, 107, 131, 133, 134, 137,
142, 147, 172, 174–177, 181,
182, 184–187, 191, 204, 205,
208, 209, 211, 216, 218, 223,
239, 246, 255, 258, 259

functionality 23, 24, 33, 34, 78, 81,
102, 104, 107, *109*, 122, 126,
132–134, 137–139, 142,
146–149, 153, 155, 156, 168,
169, 176, 180, 185, 191–193,
195, 196, 208, 211, 214, 216,
217, 223–226, 228, 230, 244,
245, 249, 251–253, 258

deconcentration, 35, 133, 146,
190, 222

distribution, 138

G

geographical

decentralization, 45

H

hardware 19, 21, 22, 25, 47, 48, 52,
62, 66, 68, 69, 71, 74, 75,
109, 114, 116, 199, 203, 206,
238

deconcentration, 75

I

implementation *22*, 23, 33, 35,
47–49, 52, 53, 56–58, 60, 62,
63, 65, 67, 69–71, 73, 76, 80,
88, 90, 94, 95, 99–101, 107,
126, 130, 133, 136, 137, 139,
159, 166–170, 172, 174, 175,
182–184, 186, 190, 197,
200–202, 208, 211–214, 216,
217, 223, 225, 230, 232, 235,
236, 241, 243, 244, 250, 251,
257, 260, 261

information system 19, 20, *21*

information systems management 20,
22, 25

278

SUBJECT INDEX

paradigm, 21, 25
insert integrity 111, 113, 123, 163,
166, 199, 201, 233, 243
instrumentation 28, 30, 31–34, 36,
37, 81, 83, 104, 143, 145,
146, 149, 177, 178, 181, 189,
190, 192, 204, 209–213, 221,
222, 225, 240, 241, 249–251,
253, 256, 257, 260, 261

J

Job Entry System 234, 238

L

load factor 53, 114, 115, 116, 121,
124, 172, 173, 175, 183, 201,
202, 212, 234, 258
local autonomy 20, 56, 62, 84, 85,
91, 102, 103, 134

M

maintenance 22, 23, 24, 25, 27, 42,
47–49, 65, 67–72, 74, 80, 92,
176, 185, 208, 225, 235, 260
methods and techniques 20, 26–28,
30, 32, 33, 36, 37, 79, 81, 83,
102, 104, 135–140, 142, 143,
145, 146, 159, 172, 177, 184,
189, 190, 192, 209, 215, 216,
222, 225, 226, 230, 241, 244,
249, 250, 253, 256, 257, 259,
260
modification integrity 111, 113, 123,
163, 166, 199, 201, 233, 243,
258

O

operational level 25, 27, 65–67, 70,
75, 80, 81, 104, 129, 137,
256, 257
organization
decentralization, 20, 34, 46, 56
organizational

distribution, 35, 132, 134, 141,
143, 146, 147, 190, 191, 222,
223, 246, 250, 252

P

people 19, 21, 22, 25, 34, 35, 46, 47,
59, 91, 95, 109, 111, 134,
137, 138, 140, 142, 146, 155,
167, 179, 187, 189, 218, 224,
230, 240, 243, 246, 250, 255,
256
planning 19–21, 22, 23, 40, 50, 65,
66, 68, 71, 72, 74, 80, 81, 87,
110, 135, 136, 139, 141, 143,
154, 155, 225, 238, 249, 256
policy 19, 22, 23, 65, 69, 80, 87, 103,
150, 155, 180
pragmatics 33, 48, 107, 109, 122,
132, 136, 140, 142, 159, 180,
182, 186, 197, 211, 214, 217,
231, 242–244, 253, 254, 258
procedures 19, 21, 22, 25, 47, 51, 57,
59, 66, 68, 70–73, 75, 76, 78,
99, 109, 125, 129, 130, 142,
163, 184, 186, 187, 203, 215,
239, 244, 246
property 52, 58, 61, 109, 119, 136,
137, 139, 140, 142, 145, 159,
162–164, 166, 172, 180–186,
197, 200, 201, 203, 211, 212,
214–217, 228, 230, 232,
244–246, 253, 254, 257–259

R

real system 20, 21, 24, 35, 40, 41, 44,
47, 48, 80, 109–111, 113,
133, 135, 137, 138, 140, 141,
250, 257
recoverability 52, 118, 120, 125, 186,
203, 215, 235, 243, 259
relatability 111, 112, 113, 123, 163,
164, 166, 199, 200, 212, 214,
217, 233, 243

SUBJECT INDEX

279

reliability 19, 41, 52, 53, 56, 58, 69,
118, 119, 125, 173, 186, 203,
235
repairability 118, 119, 125, 186, 203,
215, 235, 259
requirements and preconditions
19–22, 25–27, 30, 32, 33,
39–42, 45, 47, 50–53, 60,
70–72, 76, 78, 80, 90, 100,
103, 108, 109, 113, 122–125,
130–133, 136–140, 145, 149,
159, 160, 162–164, 166, 172,
173, 180–186, 189, 197,
199–203, 207, 211–218, 230,
231, 233–236, 241–246,
253–261
responsibility
distribution, 35, 141, 174, 177,
208, 240, 250
restructuring speed 114, 115, 124,
201, 234
retention time 71, 118, 121, 125,
203, 235, 236

S

safety 48, 118, 121, 125, 173, 203,
204, 215, 235, 244, 259
securability 111, 113, 123, 163, 166,
199, 201, 214, 217, 233, 234,
243, 258
semantics 33, 48, 51, 52, 107, 109,
111, 113, 114, 122, 123, 132,
136, 138–140, 142, 159, 163,
164, 166, 180, 182–186, 197,
199, 200, 211, 212, 214, 215,
217, 231, 233, 242–245, 253,
254, 258
software 19, 21, 22, 24, 25, 47–49, 53,
55, 56, 66, 68, 69, 71, 74, 75,
109, 111, 114, 116, 125, 135,
166, 175, 176, 181, 185, 186,
199, 201, 208, 214, 217, 223,
225, 232, 236–238, 243, 252

deconcentration, 75
stakeholder 26, 33, 34, 40, 45, 86, 90,
102, 103, 109, 111, 122–125,
129, 131–133, 137–140, 142,
145, 146, 148, 149, 155, 156,
159, 160, 169, 172, 176,
179–186, 191–193, 196, 201,
211–218, 223, 225–227, 229,
231, 236, 243–246, 251, 253,
256, 259, 260
distribution, 102
static integrity 111, 113, 123, 163,
166, 199, 214, 233, 243
storage
deconcentration, 56, 60
storage capacity 55, 98, 118, 120,
121, 125, 173, 184, 203, 204,
212, 215, 235, 244, 258, 259
storage costs 49, 59, 103, 118, 121,
125, 128, 184, 203, 235, 258
storage management 25, 27, 33, 107,
122, 125, 236, 254, 255, 258
storage structures 53, 64, 65, 70,
114–116, 124, 172, 175, 201,
205, 237
strategic level 25, 27, 34, 46, 65–67,
72, 75, 77, 80, 81, 103, 127,
129, 130, 141, 155, 180
syntactics 33, 48, 51, 107, 114, 118,
121, 122, 124, 132, 136, 139,
140, 142, 159, 172, 181–183,
197, 201, 211, 212, 214, 215,
217, 231, 234, 242, 243, 254,
258

T

tactical level 25, 27, 65–67, 75, 81,
129, 257
task field 22, 24, 25, 27, 33, 67, 69,
70, 81, 104, 107, 122,
123–126, 129, 131, 132, 142,
181, 213, 236, 246, 254, 255,
258, 260, 261

280

SUBJECT INDEX

technical

distribution, 35, 102, 132, *133*,
139, 142, 146, 147, 190, 191,
218, 222, 223, 246, 250, 251

technical management 24, *25*, 27, 78,
79, 99, 101, 107, 131–134,
137, 147, 175–178, 181, 184,
187, 191, 205, 206, 208, 214,
215, 217, 218, 246, 255, 258,
259

timeliness 51, 52, 109, *110*, 122, 159,
162, 197, 198, 231, 232, 242,
258

transfer rate 54, 118, *120*, 125, 203,
235

transformation 114, *116*, 124, 201,
202, 215, 234, 258

U

use factor 98, 111, 114, *116*, 124,
201, 234

utilization 19–22, *23*, 24–27, 32, 35,
39, 45, 48, 50, 54, 57, 58, 65,
67, 69, 78, 80, 83, 84, 109,
111, 115, 121–123, 125, 131,
133, 135–137, 142, 145, 167,
169, 174, 181, 183, 185, 186,
204, 223, 236, 250, 251, 256,
260

distribution, 167, 168

V

Virtual System Access Method 234,
235

W

work

distribution, 35, 141, 174, 177,
208, 240, 250

Samenvatting

Beheer van gespreide gegevens in gespreide omgevingen

Beheer van gegevens

Het beheer van gegevens kent een lange historie met als gevolg dat er een groot aantal opvattingen over het gegevensbeheer is ontstaan. Met de toenemende penetratie van automatisering en informatie-technologie binnen het bedrijfsleven en de overheid, worden zij in toenemende mate afhankelijk van geautomatiseerde informatiesystemen voor de besturing en verbetering van hun produkten en diensten. Als gevolg hiervan komt het beheer steeds meer in de belangstelling te staan en is men op zoek naar een systematische aanpak ter vervanging van de traditionele wijze van beheer.

Elke organisatie kan geabstraheerd worden in een informatiesysteem (IS) en bijbehorend reëel systeem (RS). Een geautomatiseerd informatiesysteem is een geheel van apparatuur met bijbehorende basisprogrammatuur en toepassingsprogrammatuur, gegevens, procedures en personen (de componenten van het informatiesysteem) voor het kennen en/of besturen/ondersteunen van reële systemen. Binnen het reële systeem vinden de bedrijfsprocessen plaats die bestuurd worden vanuit het informatiesysteem op basis van de representaties die door de gegevens gemaakt worden. De omgeving van gegevens is gedefinieerd als de belanghebbers binnen het reële systeem die baat hebben bij het succesvol functioneren van de bedrijfsprocessen en die vanwege de toenemende afhankelijkheid steeds hogere eisen en randvoorwaarden aan de gegevens stellen. Elke component van het informatiesysteem bezit bepaalde specifieke eigenschappen die getypeerd worden door statische en dynamische karakteristieken. De statische karakteristieken van gegevens worden in eerste instantie vastgelegd tijdens de toestanden 'informatiebeleid en -planning' en 'ontwikkeling' uit de levenscyclus van een informatiesysteem. Tijdens de toestanden 'gebruik' en 'exploitatie' vertonen gegevens voorts dynamische karakteristieken die tezamen met de statische gegevenskarakteristieken noodzaken tot beheeractiviteiten om te garanderen dat op ieder moment

de gegevenskarakteristieken voldoen aan de gestelde eisen en randvoorwaarden.

Het beheer van gespreide gegevens in gespreide omgevingen is gedefinieerd als de instandhouding van de gespreide gegevens overeenkomstig de eisen en randvoorwaarden gesteld vanuit het gespreid gebruik en rekening houdend met de karakteristieken van de gespreide gegevens. Spreiding van gegevens wordt verdeeld in twee typen: deconcentratie en decentralisatie, deze beïnvloeden zowel gegevenskarakteristieken als eisen en randvoorwaarden.

Deconcentratie van gegevens heeft betrekking op de spreiding van de gegevens over meerdere geografische lokaties binnen de organisatie. Deconcentratie van gegevens kan gebaseerd zijn op twee principes. Het eerste principe heet 'partitionering'; een bepaald gegevensobject (bijvoorbeeld een personeelsrecord) heeft één fysieke representatie en wordt op precies één plaats opgeslagen, een ander gegevensobject (bijvoorbeeld een klantenrecord) heeft één fysieke representatie die op een andere plaats opgeslagen wordt. Hierdoor ontstaat de situatie dat alle personeelsgegevens op de ene plaats zijn opgeslagen en alle klantgegevens op een andere plaats. Het tweede principe heet 'replicatie'; een bepaald gegevensobject heeft meerdere fysieke representaties die op meerdere lokaties opgeslagen kunnen worden. Bij beide principes kunnen de gegevenskarakteristieken per lokatie verschillen omdat het gebruik per lokatie kan verschillen en omdat de geëxploiteerde automatiseringsmiddelen per lokatie eveneens kunnen verschillen.

Decentralisatie van gegevens heeft betrekking op de spreiding van bevoegdheden betreffende gegevens over meerdere onderdelen van de organisatie. Zowel in het openbaar bestuur als in het bedrijfsleven heeft de afgelopen jaren een toename in de lokale autonomie van organisatie-eenheden plaatsgevonden. Lokale autonomie houdt in dat organisatie-eenheden een eigen verantwoordelijkheid en bevoegdheid met betrekking tot gegevens verkrijgen. Het houdt tevens in dat autonome eenheden eigen eisen en randvoorwaarden aan gegevens zullen stellen die gericht zijn op hun specifieke situatie. In vele gevallen worden gegevens ook buiten de organisatie-eenheid gebruikt door andere onderdelen van de organisatie en wellicht worden gegevens zelfs buiten de organisatie gebruikt. Als gevolg hiervan leidt decentralisatie tot verschillende eisen en randvoorwaarden aan gegevens die bovendien strijdig met elkaar kunnen zijn.

Gezamenlijk leiden deze twee typen van spreiding van gegevens tot een complexe situatie waarin het beheer zich moet inspannen voor de instandhouding van gegevens in overeenstemming met een grote variëteit aan eisen en randvoorwaarden vanuit het gebruik en een grote variëteit aan gegevenskarakteristieken.

Onderzoeksdoelstelling

Het onderzoeksprogramma naar het beheer van informatiesystemen is onder meer gericht op de ontwikkeling van toepassingsonafhankelijke methoden en technieken ten behoeve van het beheer van informatiesystemen en de toepassing hiervan in specifieke situaties. Binnen dit onderzoeksprogramma vormen de definitie van beheer en

SAMENVATTING

283

het taken-toestandenmodel belangrijke uitgangspunten, alsmede het drievoudig model van beheer. Het onderzoek naar het beheer van gespreide gegevens in gespreide omgevingen is binnen dit onderzoeksprogramma gepositioneerd door zich in de eerste plaats te richten op het beheer van de gegevenscomponent van het informatiesysteem. In de tweede plaats richt het onderzoek zich op de toestanden 'gebruik' en 'exploitatie' van gegevens, waar het beheer zorgt voor de instandhouding van een dynamisch evenwicht tussen eisen en randvoorwaarden en de gegevenskarakteristieken. Op de derde plaats richt het onderzoek zich op de ondersteuning van personeel op het operationele niveau van de organisaties die verantwoordelijk zijn voor het functioneel en technisch beheer van gegevens.

Het onderzoek had tot doel om een instrumentarium te ontwikkelen waarin zich een conceptueel raamwerk bevindt en waarin methoden en technieken geplaatst worden ter ondersteuning van het gegevensbeheer op operationeel niveau. Het onderzoek was er voorts op gericht om dit instrumentarium te toetsen in een drietal situaties die betrekking hebben op het beheer van gespreide gegevens in gespreide omgevingen. Ten einde dit doel binnen de gegeven onderzoeksperiode te bereiken zijn drie doelstellingen geformuleerd:

1. Het ontwikkelen van een conceptueel raamwerk als onderdeel van het instrumentarium, dat alle relevante aspecten bevat die betrekking hebben op het beheer van gespreide gegevens in gespreide omgevingen.
2. Het ontwikkelen van methoden en technieken als onderdeel van het instrumentarium, ter ondersteuning van het beheer van gespreide gegevens in gespreide omgevingen.
3. De empirische toetsing van de adequaatheid van het instrumentarium, conceptueel raamwerk en methoden en technieken, binnen de praktijk van het beheer van gespreide gegevens in gespreide omgevingen

Vanuit deze doelstellingen heeft allereerst een uitgebreide literatuurstudie plaatsgevonden ten einde een beeld te verkrijgen van zowel de spreiding van omgevingen, als de spreiding van gegevens en het gegevensbeheer. De literatuurstudie geeft een overzicht van de geldende opvattingen rond het onderzoeksgebied en geeft aan op welke wijze het onderzoek een bijdrage kan leveren aan het onderzoeksgebied. Theoretische verkenning is echter niet voldoende; eigen waarnemingen zijn noodzakelijk om inzicht te verkrijgen in de praktijk van het gegevensbeheer. Hiertoe zijn twee verkennende onderzoeken gepleegd: één bij het Gemeenschappelijk Administratie Kantoor (GAK) en één bij de Centrale Organisatie van het Ministerie van Defensie.

Verkenning van literatuur en praktijk

Uit het onderzoek naar literatuur op het gebied van beheer van gespreide gegevens in gespreide omgevingen komt een gefragmenteerd beeld naar voren. Publikaties zijn

voornamelijk gebaseerd op individuele praktijkervaringen in verschillende gebieden en gezien vanuit verschillende invalshoeken. Er wordt een groot aantal karakteristieken van gegevens gedefinieerd en een groot aantal taken en functionarissen in het beheer. Een heldere samenhang tussen de eisen en randvoorwaarden vanuit de omgeving, de karakteristieken van de gegevens en de taken in het beheer ontbreekt. De literatuur geeft geen uniform beeld welke beheertaken op het operationeel, het tactisch en het strategisch niveau van beheer uitgevoerd dienen te worden. Er worden geen duidelijke richtlijnen aan het gegevensbeheer op het operationeel niveau gegeven, tevens reikt de literatuur geen methoden en technieken aan voor de ondersteuning van het operationeel niveau.

De twee verkennende praktijkonderzoeken geven aan dat er geen gelijke tred is geweest in de spreiding van gegevens binnen organisaties en de wijze waarop het gegevensbeheer hiermee omgaat. Het ontbreekt aan inzicht en een samenhangende zienswijze betreffende het gegevensbeheer. Er zijn geen duidelijke richtlijnen aangetroffen voor het beheer van gespreide gegevens in gespreide omgevingen. Tevens zijn er geen voorgeschreven methoden en technieken aangetroffen ter ondersteuning van het operationeel niveau van het gegevensbeheer. Deze constatering laten zien dat er behoefte is aan instrumenten om de vaak historisch gegroeide situatie, alsook nieuwe situaties, 'beheerbaar' te maken.

Onderzoeksvragen

Gegevenskarakteristieken spelen een belangrijke rol binnen het beheer, zij zijn de 'procesvariabelen' waar het beheer zich naar richt bij het uitvoeren van de beheertaken. De literatuur geeft een groot aantal karakteristieken die vanuit verschillende invalshoeken gedefinieerd worden. Binnen het ontwikkelde conceptueel raamwerk worden alle relevante gegevenskarakteristieken op eenduidige wijze gedefinieerd. Empirische toetsing moet vaststellen in welke mate de gedefinieerde karakteristieken adequaat zijn binnen het gegevensbeheer. Dit leidt tot de eerste onderzoeksvraag:

Welke gegevenskarakteristieken zijn van belang voor het uitvoeren van taken binnen het beheer van gespreide gegevens in gespreide omgevingen?

Zowel literatuur als de onderzochte praktijksituaties reiken geen methoden en technieken aan die binnen het gegevensbeheer gebruikt kunnen worden. Binnen het onderzoek worden methoden en technieken ontwikkeld ter ondersteuning van het operationeel niveau van het gegevensbeheer. Empirische toetsing moet vaststellen in welke mate deze methoden en technieken adequaat zijn binnen het gegevensbeheer. Dit leidt tot de tweede onderzoeksvraag:

Welke methoden en technieken kunnen ter beschikking gesteld worden aan het beheer voor de instandhouding van gespreide gegevens overeenkomstig de eisen en randvoorwaarden gesteld vanuit een gespreide omgeving en de karakteristieken van de gespreide gegevens?

Het instrumentarium

Het instrumentarium voor het beheer van gespreide gegevens in gespreide omgevingen bestaat uit twee onderdelen. Het eerste onderdeel is het FATO-raamwerk dat binnen het onderzoek ontwikkeld is als conceptueel raamwerk dat op generieke wijze verband legt tussen de relevante aspecten van het gegevensbeheer. Het tweede onderdeel van het instrumentarium bestaat uit de methoden en technieken die binnen het onderzoek ontwikkeld zijn en die toegepast worden binnen het FATO-raamwerk.

Het FATO-raamwerk

Het FATO-raamwerk is genoemd naar de vier onderdelen van het raamwerk: Functionaliteiten, Automatiseringsmiddelen, Taakvelden en Organisatie. Het wordt gebruikt om alle relevante aspecten van het beheer van gespreide gegevens in gespreide omgevingen te positioneren. Functionaliteiten hebben betrekking op het gebruik van de gegevens door de belanghebbers die de omgeving vormen. De gegevenskarakteristieken gerelateerd aan het gebruik van de functionaliteiten, worden geplaatst op het pragmatische beschouwingsniveau (betreffende gebruik of verwerking) en het semantische beschouwingsniveau (betreffende betekenis of verwijzing) van gegevens. Automatiseringsmiddelen hebben betrekking op de exploitatie van de technische realisatie van gegevens. De gegevenskarakteristieken gerelateerd aan de exploitatie van automatiseringsmiddelen, worden geplaatst op het syntactisch beschouwingsniveau (betreffende vorm of structuur) en het empirische beschouwingsniveau (betreffende medium of signaal) van gegevens.

De instandhouding van gegevens, overeenkomstig de eisen en randvoorwaarden en de karakteristieken van de gegevens, wordt gerealiseerd door het uitvoeren van taken in het beheer. De gegevenskarakteristieken hebben betrekking op het gebruik van de functionaliteiten en de exploitatie van automatiseringsmiddelen. Het FATO-raamwerk geeft een fundamentele onderbouwing voor de clustering van de verschillende beheertaken tot taakvelden, door deze clustering te baseren op de hierboven genoemde vier beschouwingsniveau's van gegevens. Op deze wijze wordt een verband aangebracht tussen de gegevenskarakteristieken van de functionaliteiten en automatiseringsmiddelen en de hierbij betrokken taakvelden. Het taakveld 'inhoudelijk gegevensbeheer' betreft de instandhouding van gegevens op het pragmatische beschouwingsniveau; het taakveld 'gegevensdefinitiebeheer' betreft de instandhouding van gegevens op het semantische beschouwingsniveau; het taakveld 'gegevensstructuurbeheer' betreft de instandhouding van gegevens op het syntactische beschouwingsniveau; en het taakveld 'fysiek gegevensbeheer' betreft de instandhouding van gegevens op het empirische beschouwingsniveau.

De uitvoering van de benodigde beheertaken vereist een effectief en efficiënt opererende beheerorganisatie. Binnen een beheerorganisatie dient een functiescheiding aangebracht te worden die rekening houdt met de verschillen in benodigde expertise, de inhoudelijkheid van de taken en het mogelijk belangenconflict. Om deze rede-

nen wordt het gegevensbeheer gesplitst in twee beheervormen, te weten: functioneel beheer en technisch beheer. Het functioneel beheer is gericht op ondersteuning van het gebruik van de gegevens en omvat taken binnen de taakvelden 'inhoudelijk gegevensbeheer' en 'gegevensdefinitiebeheer'. Het functioneel beheer verschaft optimale ondersteuning aan het gebruik van de gegevens en is gericht op het maximaliseren van de effectiviteit van gegevens door middel van functionele expertise. Het technisch beheer is gericht op de exploitatie van de gegevens en omvat taken binnen de taakvelden 'gegevensstructuurbeheer' en 'fysiek gegevensbeheer'. Het technisch beheer verschaft optimale exploitatie van de aanwezige automatiseringsmiddelen en is gericht op het maximaliseren van de efficiëntie van de gegevens door middel van technische expertise.

De spreiding van gegevens wordt aan de hand van het FATO-raamwerk ingedeeld naar de drie invalshoeken: functioneel, technisch en organisatorisch. De vierde invalshoek 'taakvelden' wordt buiten beschouwing gelaten omdat taakvelden binnen de drie overige invalshoeken aan de orde komen.

Functionele spreiding is gerelateerd aan de spreiding van het gebruik van de functionaliteiten van gegevens. Dit heeft consequenties voor taakvelden binnen het functioneel beheer omdat inhoudelijkheid en definities van gegevens per gebruiksomgeving kunnen verschillen. Technische spreiding is gerelateerd aan de spreiding van de geëxploiteerde automatiseringsmiddelen. Dit heeft consequenties voor taakvelden binnen het technisch beheer omdat structuur en vastlegging van gegevens per technische implementatie kunnen verschillen. Organisatorische spreiding is gerelateerd aan de spreiding van de uitvoering van de beheertaken over meerdere beheereenheden. Dit heeft consequenties voor de coördinatie van beheertaken omdat de uitvoering van taken, de benodigde bevoegdheden en verantwoordelijkheden en de aanwezige expertise per lokatie kunnen verschillen.

Methoden en technieken

Het gegevensbeheer is te beschouwen als een informatiesysteem voor het kennen en/of besturen/ondersteunen van gegevens. Deze beschouwingswijze maakt het mogelijk om methoden en technieken voor informatiesystemen te gebruiken voor het beschrijven van het gegevensbeheer. De methoden en technieken die binnen het onderzoek gebruikt zijn richten zich op de systelologische aspecten betreffende de resultaten van activiteiten van de gegevensbeheer-organisatie op de gegevenskarakteristieken (het 'waartoe'). Er is gekozen voor ruim bekende methoden en technieken: "business system planning" (BSP) en "information systems work and analysis of change" (ISAC). Dit heeft het voordeel dat tijdens de toetsing van het instrumentarium de bekendheid met de methoden en technieken validatie van de opgestelde modellen bij het betrokken beheerpersoneel kan vereenvoudigen. Om ondersteuning te geven aan de levenscyclus van het informatiesysteem binnen het toestandenmodel, dienen methoden en technieken binnen het beheer van informatiesysteem zoveel mogelijk aan te sluiten bij methoden en technieken die binnen beleid en planning en ontwikkeling van informatiesystemen gebruikt worden. Doordat het beheer zijn eigen kenmerken heeft,

is enige aanpassing van deze methoden en technieken echter noodzakelijk.

Empirische toetsing

De praktische bruikbaarheid (de adequaatheid) van het ontwikkelde instrumentarium moet door empirische toetsing vastgesteld worden. Deze toetsing heeft plaatsgevonden in drie verschillende casus-studies. Het criterium dat gebruikt is om de resultaten van deze drie casus-studies te beoordelen wordt weergegeven door de onderstaande vraag:

Is het mogelijk om door toepassing van het instrumentarium knelpunten in specifieke situaties van gegevensbeheer aan te geven, en kunnen met behulp van het instrumentarium aanbevelingen voor verbeteringen in het beheer van gespreide gegevens in gespreide omgevingen gedaan worden?

De empirische toetsing heeft plaatsgevonden binnen het Gemeenschappelijk Administratie Kantoor (GAK) en binnen Shell Chemie Moerdijk.

Casus Basisregistratie Personen (BRP)

Binnen de casus BRP is het instrumentarium gebruikt om een analyse te geven van het beheer van infrastructurele persoonsgegevens binnen het GAK.

Toepassing van het instrumentarium maakt het mogelijk om eenduidige specificatie van de BRP te verkrijgen. Het beheer heeft inzicht verkregen in het grote aantal belanghebbers bij de BRP en in de verschillende functionaliteiten die ondersteund moeten worden per belanghebbende. De BRP bestaat uit een hoofd-database met twee kopie-databases geplaatst in een 'centraal service centrum' en één kopie-database geplaatst op een drietal 'regionale service centra'. Het functioneel beheer heeft pragmatische en semantische eisen en randvoorwaarden vastgelegd die gelden voor de persoonsgegevens in alle databases. Het functioneel beheer heeft de middelen en bevoegdheden om beheer te plegen op de hoofd-database, maar het ontbreekt aan middelen en bevoegdheden om functioneel beheer te plegen op de kopie-databases. Met het instrumentarium zijn beheertaken geïdentificeerd die niet geplaatst zijn in overeenstemming met het FATO-raamwerk; dit heeft tot aanpassing van de bestaande situatie geleid met plaatsing van beheertaken overeenkomstig het FATO-raamwerk.

Casus Werkeloosheidswet Ontslag (WWO)

Binnen de casus WWO is het instrumentarium gebruikt om de bestaande beheerorganisatie rond het informatiesysteem WAO voor te bereiden op de introductie van het nieuwe informatiesysteem WWO.

Niet alle belanghebbers worden binnen de functionele documentatie van het WWO informatiesysteem als zodanig herkend: er is alleen sprake van "gebruikers". Met

toepassing van het instrumentarium zijn twintig verschillende groepen belanghebbers onderkend, uiteenlopend van individuele gebruikers en informatiesystemen binnen het GAK tot organisaties buiten het GAK, die ondersteuning behoeven vanuit het functioneel gegevensbeheer. Het is gebleken dat de functionele documentatie reeds verouderd was op het moment van opleveren van het WWO informatiesysteem. Met behulp van het instrumentarium is het functioneel gegevensbeheer in staat correcte functionele documentatie op te stellen om op basis hiervan ondersteuning te bieden aan het gebruik van het WWO informatiesysteem. Het blijkt dat het gegevensbeheer weliswaar eisen en randvoorwaarden vastgesteld heeft, maar dat de uitvoering van activiteiten binnen het gegevensbeheer niet op gegevenskarakteristieken gebaseerd is en dat de beheeractiviteiten op een ad hoc wijze uitgevoerd worden. Omdat de beheeractiviteiten niet gebaseerd zijn op de gegevenskarakteristieken, zijn de karakteristieken niet in overeenstemming met de eisen en randvoorwaarden. Uit toepassing van het instrumentarium is voorts gebleken dat het gegevensbeheer voor de uitvoering van de beheertaken een beroep moet doen op de ontwikkelingsorganisatie. Aangetoond is dat de voor deze taken benodigde expertise binnen de ontwikkelingsorganisatie is gebleven en niet is overgedragen aan de beheerorganisatie. Met behulp van het instrumentarium zijn verbeteringen aangegeven die het gegevensbeheer in staat stellen alle beheertaken in het vervolg zelf uit te voeren.

Casus Moerdijk Process Database (MPD)

Binnen de casus MPD is het instrumentarium gebruikt om een analyse te geven van het gegevensbeheer binnen een industriële productie-organisatie in vergelijking met het gegevensbeheer binnen een administratieve organisatie.

De MPD-functionaliteiten zijn uitsluitend toegekend aan gebruikers. Hierdoor is ondersteuning van het gebruik van MPD-gegevens binnen de geautomatiseerde productiesystemen van Shell Chemie Moerdijk niet mogelijk. Het gebruik van MPD-gegevens is stabiel vanwege het feit dat de gegevens voornamelijk aangeleverd en afgenomen worden door geautomatiseerde productieprocessen. Deze stabiliteit leidt tot eenvoudige eisen en randvoorwaarden op het pragmatisch en semantisch niveau van gegevens. Het real-time aspect van de gegevens behoeft dankzij het stabiele gebruik geen extra beheerinspanning. Doordat alle belanghebbers van de MPD zich binnen de Shell organisatie bevinden is het functioneel beheer eenvoudiger vergeleken met het GAK waar veel belanghebbers zich buiten de organisatie bevinden. De eisen en randvoorwaarden aan de syntactische en empirische gegevenskarakteristieken worden binnen het technisch beheer op informele wijze gespecificeerd. Het gegevensbeheer van de MPD maakt geen gebruik van voorgeschreven methoden en technieken voor de uitvoering van gegevensbeheer activiteiten. De stabiele en eenvoudige beheersituatie binnen Shell Chemie Moerdijk maakt het mogelijk om met een kleine beheerorganisatie alle benodigde beheertaken uit te voeren.

Conclusies

De eerste onderzoeksvraag heeft betrekking op het belang van de gedefinieerde gegevenskarakteristieken voor de uitvoering van taken binnen het beheer van gespreide gegevens in gespreide omgevingen. Uit de casus-studies BRP, WWO en MPD blijkt dat het functioneel beheer eisen en randvoorwaarden op het pragmatisch en semantisch niveau van gegevenskarakteristieken vastlegt overeenkomstig het FATO-raamwerk. Doordat het functioneel beheer de activiteiten binnen de beheertaken niet baseert op gegevenskarakteristieken, is de instandhouding van de gegevens niet in overeenstemming met deze eisen en randvoorwaarden. Uit de casus-studies BRP, WWO en MPD blijkt verder dat het technisch beheer een gering aantal eisen en randvoorwaarden vastlegt op het syntactisch en empirisch niveau van gegevenskarakteristieken. Doordat het technisch beheer de activiteiten binnen de beheertaken wel baseert op de gegevenskarakteristieken, is de instandhouding van de gegevens wel in overeenstemming met deze eisen en randvoorwaarden. Binnen de casus-studies werden geen karakteristieken aangetroffen die niet in het FATO-raamwerk gedefinieerd zijn. Als zodanig is er geen aanleiding om het FATO-raamwerk bij te stellen en luidt het antwoord op de eerste onderzoeksvraag:

De karakteristieken van gegevens zoals gedefinieerd in hoofdstuk 4 zijn waardevol voor het functioneel beheer en, in mindere mate, waardevol voor het technisch beheer.

De tweede onderzoeksvraag heeft betrekking op de methoden en technieken die aan het beheer van gespreide gegevens in gespreide omgevingen ter beschikking gesteld kunnen worden. De methoden en technieken zoals gedefinieerd in hoofdstuk 5 blijken een waardevol hulpmiddel in het onderkennen van knelpunten en het aangeven van verbeteringen in het beheer. Met betrekking tot de twee methoden en technieken 'belangengroepentabel' en 'coördinatiemechanismen' wordt opgemerkt dat zij binnen de casus-studies niet toegepast konden worden. De belangengroepentabel die de relaties tussen de belanghebbers, de eisen en randvoorwaarden en gegevensbeheer-activiteiten zou moeten tonen, kon niet gemaakt worden omdat het beheerpersoneel niet in staat was deze relaties aan te geven. Het was evenmin mogelijk om modellen te maken van de coördinatiemechanismen binnen het gegevensbeheer omdat hiervoor geen technieken beschikbaar zijn. Ondanks deze twee uitzonderingen luidt het antwoord op de tweede onderzoeksvraag:

De methoden en technieken zoals beschreven in hoofdstuk 5 zijn waardevol voor het beheer van gespreide gegevens in gespreide omgevingen.

Het onderzoek stelt het beheer van gespreide gegevens in gespreide omgevingen een instrumentarium ter beschikking dat bestaat uit het FATO-raamwerk en uit methoden en technieken. Het FATO-raamwerk is een generiek model dat kernachtige en bruikbare definities geeft van het operationeel niveau van gegevensbeheer. De methoden en technieken zijn vertaald van beleid, planning en ontwikkeling naar de

specifieke behoeften van gebruik en exploitatie van gegevens. Het instrumentarium ondersteunt het beheerpersoneel op het operationeel niveau in de instandhouding van gegevens overeenkomstig eisen en randvoorwaarden. Tevens stelt het instrumentarium het personeel in staat om vast te stellen welke eisen en randvoorwaarden door welke belanghebbers worden gesteld.

Mogelijke toepassingen van het instrumentarium

Op het tactisch niveau van gegevensbeheer is het instrumentarium te gebruiken om een diagnose-stelling van het gegevensbeheer op operationeel niveau mogelijk te maken. Deze diagnose-stelling kan onder meer gebruikt worden ten behoeve van kwaliteitsborging of de bepaling van beheerkosten. De diagnose-stelling kan binnen iedere beheersituatie uitgevoerd worden door de onderstaande stappen te volgen.

Het generiek model voor gegevensbeheer wordt toegepast in een specifieke situatie door met behulp van de methoden en technieken een beschrijving te geven van het operationeel niveau van gegevensbeheer. Deze beschrijving wordt opgesteld door het doorlopen van de vier fasen: functionaliteiten, automatiseringsmiddelen, taakvelden en organisatie, zoals beschreven in paragraaf 1.4.6. Uit de vergelijking van deze beschrijvingen met de inrichting van het gegevensbeheer zoals dat voorgeschreven is door het generiek model, volgen aanbevelingen tot het verbeteren van de specifieke situatie (zie ook figuren 6.5 en 7.5). Deze aanbevelingen kunnen betrekking hebben op drie aspecten van het generiek model. Ten eerste kan er aanbevolen worden om de plaatsing van de verschillende onderdelen van het gegevensbeheer in overeenstemming te brengen met de positionering van het gegevensbeheer binnen het FATO-raamwerk. Ten tweede kan er aanbevolen worden om de verbanden tussen de omgeving (het reële systeem), de gegevens, en het gegevensbeheer in overeenstemming te brengen met het generiek model. Ten derde kan er aanbevolen worden om de invulling van onderdelen van het FATO-raamwerk aan te passen. Voorstellen tot verbeteringen kunnen met behulp van methoden en technieken geanalyseerd worden voordat tot daadwerkelijke implementatie overgegaan wordt.

Curriculum Vitae

Rob Mersel was born in Rotterdam on the third of May 1966. He attended the 'Tooropschool voor R.K. MAVO' in Rotterdam where he received his diploma in 1982. Afterwards he attended the 'Sint-Laurenscollege' in Rotterdam where he received his 'VWO' diploma in 1985. In 1985 he started his study of informatics (which is derived from the French contraction *informatique* of *information automatique*) at Delft University of Technology. His master's thesis focussed on the redesign of a production control database at ICI Netherlands BV in Rozenburg using semantic data modelling techniques. He was granted his masters degree in April 1991

He began his research project for this thesis in May 1991. As part of a cooperation between the Delft University of Technology and the Joint Office for Social Security Administration (GAK) he performed much of his research at GAK from 1992 until 1994. During 1991-1992 he was a parttime teacher of database systems at the 'Hogeschool voor Economische Studies' (HES) Rotterdam. He has coached several students during their master's project and he has participated in several courses on information systems management, and lectured on the subject of data management.

At present Rob Mersel is working for InterProm b.v. Veenendaal as a consultant in management and security of open, multi-vendor IT-infrastructures. InterProm focusses on IT management organizations, service management architectures, processes and technology.