

CONTENTS

PREFACE xiii

1

INTRODUCTION

1.1	Data Quality Horror Stories	2	
1.2	Knowledge Management and Data Quality	7	
1.3	Reasons for Caring About Data Quality	10	
1.4	Knowledge Management and Business Rules		15
1.5	Structure of This Book	17	

2

WHO OWNS INFORMATION?

2.1	The Information Factory	26	
2.2	Complicating Notions	28	
2.3	Responsibilities of Ownership	31	
2.4	Ownership Paradigms	38	
2.5	Centralization, Decentralization, and Data Ownership Policies	38	
2.6	Ownership and Data Quality	43	
2.7	Summary	46	

3

DATA QUALITY IN PRACTICE

3.1	Data Quality Defined: Fitness for Use	47	
3.2	The Data Quality Improvement Program	49	
3.3	Data Quality and Operations	54	
3.4	Data Quality and Databases	55	
3.5	Data Quality and the Data Warehouse	61	
3.6	Data Mining	63	
3.7	Data Quality and Electronic Data Interchange	65	
3.8	Data Quality and the World Wide Web	67	
3.9	Summary	71	

4

ECONOMIC FRAMEWORK OF DATA QUALITY AND THE VALUE PROPOSITION

4.1	Evidence of Economic Impact	74	
4.2	Data Flows and Information Chains	76	
4.3	Examples of Information Chains	80	
4.4	Impacts	83	
4.5	Economic Measures	85	
4.6	Impact Domains	85	
4.7	Operational Impacts	87	
4.8	Tactical and Strategic Impacts	90	
4.9	Putting It All Together — The Data Quality Scorecard	93	
4.10	Adjusting the Model for Solution Costs	97	
4.11	Example	97	
4.12	Summary	98	

5

DIMENSIONS OF DATA QUALITY

5.1	Sample Data Application	102	
5.2	Data Quality of Data Models	102	

5.3	Data Quality of Data Values	113
5.4	Data Quality of Data Domains	116
5.5	Data Quality of Data Presentation	117
5.6	Data Quality of Information Policy	122
5.7	Summary: Importance of the Dimensions of Data Quality	123

6

STATISTICAL PROCESS CONTROL AND THE IMPROVEMENT CYCLE

6.1	Variation and Control	126
6.2	Control Chart	128
6.3	The Pareto Principle	128
6.4	Building a Control Chart	131
6.5	Kinds of Control Charts	132
6.6	Example: Invalid Records	135
6.7	The Goal of Statistical Process Control	137
6.8	Interpreting a Control Chart	138
6.9	Finding Special Causes	140
6.10	Maintaining Control	140
6.11	Summary	140

7

DOMAINS, MAPPINGS, AND ENTERPRISE REFERENCE DATA

7.1	Data Types	144
7.2	Operations	146
7.3	Domains	148
7.4	Mappings	154
7.5	Example: Social Security Numbers	159
7.6	Domains, Mappings, and Metadata	161
7.7	The Publish/Subscribe Model of Reference Data Provision	163
7.8	Summary: Domains, Mappings, and Reference Data	167

8

DATA QUALITY ASSERTIONS AND BUSINESS RULES

8.1	Data Quality Assertions	169	
8.2	Data Quality Assertions as Business Rules		171
8.3	The Nine Classes of Data Quality Rules	171	
8.4	Null Value Rules	171	
8.5	Value Manipulation Operators and Functions		174
8.6	Value Rules	174	
8.7	Domain Membership Rules	177	
8.8	Domain Mappings and Relations on Finite Defined Domains	179	
8.9	Relation Rules	186	
8.10	Table, Cross-Table, and Cross-Message Assertions		188
8.11	In-Process Rules	192	
8.12	Operational Rules	195	
8.13	Other Rules	197	
8.14	Rule Management, Compilation, and Validation		197
8.15	Rule Ordering	199	
8.16	Summary	201	

9

MEASUREMENT AND CURRENT STATE ASSESSMENT

9.1	Identify Each Data Customer	204	
9.2	Mapping the Information Chain	205	
9.3	Choose Locations in the Information Chain		207
9.4	Choose a Subset of the DQ Dimensions	208	
9.5	Identify Sentinel Rules	208	
9.6	Measuring Data Quality	209	
9.7	Measuring Data Quality of Data Models		210
9.8	Measuring Data Quality of Data Values	217	
9.9	Measuring Data Quality of Data Domains		219
9.10	Measuring Data Quality of Data Presentation		221
9.11	Measuring Data Quality of Information Policy		225
9.12	Static vs Dynamic Measurement	229	
9.13	Compiling Results	230	
9.14	Summary	230	

10

DATA QUALITY REQUIREMENTS

10.1	The Assessment Process, Reviewed	235
10.2	Reviewing the Assessment	237
10.3	Determining Expectations	238
10.4	Use Case Analysis	240
10.5	Assignment of Responsibility	245
10.6	Creating Requirements	245
10.7	The Data Quality Requirements	248
10.8	Summary	250

11

METADATA, GUIDELINES, AND POLICY

11.1	Generic Elements	254
11.2	Data Types and Domains	256
11.3	Schema Metadata	260
11.4	Use and Summarization	265
11.5	Historical	266
11.6	Managing Data Domains	268
11.7	Managing Domain Mappings	269
11.8	Managing Rules	270
11.9	Metadata Browsing	275
11.10	Metadata as a Driver of Policy	276
11.11	Summary	276

12

RULE-BASED DATA QUALITY

12.1	Rule Basics	280
12.2	What Is a Business Rule?	281
12.3	Data Quality Rules Are Business Rules (and Vice Versa)	282
12.4	What Is a Rule-Based System?	283
12.5	Advantages of the Rule-Based Approach	284
12.6	Integrating a Rule-Based System	286

12.7	Rule Execution	287	
12.8	Deduction vs Goal-Oriented		289
12.9	Evaluation of a Rules System	291	
12.10	Limitations of the Rule-based Approach		294
12.11	Rule-Based Data Quality	295	
12.12	Summary	300	

13

METADATA AND RULE DISCOVERY

13.1	Domain Discovery	302	
13.2	Mapping Discovery	316	
13.3	Clustering for Rule Discovery		319
13.4	Key Discovery	326	
13.5	Decision and Classification Trees		327
13.6	Association Rules and Data Quality Rules		330
13.7	Summary	331	

14

DATA CLEANSING

14.1	Standardization	335	
14.2	Common Error Paradigms		340
14.3	Record Parsing	344	
14.4	Metadata Cleansing	348	
14.5	Data Correction and Enhancement		349
14.6	Approximate Matching and Similarity		354
14.7	Consolidation	364	
14.8	Updating Missing Fields	372	
14.9	Address Standardization		373
14.10	Summary	379	

15

ROOT CAUSE ANALYSIS AND SUPPLIER MANAGEMENT

15.1	What Is Root Cause Analysis?		382
15.2	Debugging the Process	385	

15.3	Debugging the Problem	389	
15.4	Corrective Measures — Resolve or Not?		391
15.5	Supplier Management	394	
15.6	Summary	396	

16

DATA ENRICHMENT/ENHANCEMENT

16.1	What Is Data Enhancement?	400	
16.2	Examples of Data Enhancement	400	
16.3	Enhancement Through Standardization		404
16.4	Enhancement Through Provenance	405	
16.5	Enhancement Through Context	406	
16.6	Enhancement Through Data Merging		407
16.7	Data Matching, Merging, and Record Linkage		412
16.8	Large-Scale Data Aggregation and Linkage		413
16.9	Improving Linkage with Approximate Matching		416
16.10	Enhancement Through Inference	420	
16.11	Data Quality Rules for Enhancement	421	
16.12	Business Rules for Enhancement	422	
16.13	Summary	423	

17

DATA QUALITY AND BUSINESS RULES IN PRACTICE

17.1	Turning Rules into Implementation	426	
17.2	Operational Directives	445	
17.3	Data Quality and the Transaction Factory		451
17.4	Data Quality and the Data Warehouse		455
17.5	Rules and EDI	456	
17.6	Data Quality Rules and Automated UIs		457
17.7	Summary	460	

18

BUILDING THE DATA QUALITY PRACTICE

18.1	Step 1: Recognize the Problem	463	
18.2	Step 2: Management Support and the Data Ownership Policy	464	
18.3	Step 3: Spread the Word	467	
18.4	Step 4: Mapping the Information Chain		468
18.5	Step 5: Data Quality Scorecard	470	
18.6	Step 6: Current State Assessment	472	
18.7	Step 7: Requirements Assessment	473	
18.8	Step 8: Choose a Project	474	
18.9	Step 9: Build Your Team	475	
18.10	Step 10: Build Your Arsenal	476	
18.11	Step 11: Metadata Model	479	
18.12	Step 12: Define Data Quality Rules	479	
18.13	Step 13: Archaeology/Data Mining	480	
18.14	Step 14: Manage Your Suppliers	481	
18.15	Step 15: Execute the Improvement	481	
18.16	Step 16: Measure Improvement	483	
18.17	Step 17: Build on Each Success	483	
18.18	Conclusion	484	

INDEX 485

BIBLIOGRAPHY 494