

Data Quality

THE ACCURACY DIMENSION

Jack E. Olson

*Understanding the concepts of
accurate data is fundamental to
improving the ways we collect
and use data.*

MORGAN KAUFMANN PUBLISHERS

An Imprint of Elsevier

SINGAPORE SYDNEY TOKYO

CONTENTS

Foreword	vii
Preface	xv

PART I

Understanding Data Accuracy	1
---------------------------------------	---

CHAPTER 1 The Data Quality Problem	3
--	---

1.1 Data Is a Precious Resource	3
1.2 Impact of Continuous Evolution of Information Systems	5
1.3 Acceptance of Inaccurate Data	8
1.4 The Blame for Poor-Quality Data	9
1.5 Awareness Levels	10
1.6 Impact of Poor-Quality Data	12
1.7 Requirements for Making Improvements	14
1.8 Expected Value Returned for Quality Program	15
1.9 Data Quality Assurance Technology	16
1.10 Closing Remarks	22

CHAPTER 2 Definition of Accurate Data	24
---	----

2.1 Data Quality Definitions	24
2.2 Principle of Unintended Uses	27
2.3 Data Accuracy Defined	29
2.4 Distribution of Inaccurate Data	32
2.5 Can Total Accuracy Be Achieved?	34
2.6 Finding Inaccurate Values	35
2.7 How Important Is It to Get Close?	40
2.8 Closing Remarks	41

CHAPTER 3 Sources of Inaccurate Data	43
3.1 * Initial Data Entry.	44
3.2 Data Accuracy Decay.	50
3.3 Moving and Restructuring Data	52
3.4 Using Data	62
3.5 Scope of Problems.	63
3.6 Closing Remarks.	64

PART II

Implementing a Data Quality Assurance Program . . . 65

CHAPTER 4 Data Quality Assurance.	671
4.1 Goals of a Data Quality Assurance Program	681
4.2 Structure of a Data Quality Assurance Program.	691
4.3 Closing Remarks.	781

CHAPTER 5 Data Quality Issues Management

5.1 Turning Facts into Issues	
5.2 Assessing Impact	81
5.3 Investigating Causes.	81
5.4 Developing Remedies.	91
5.5 Implementing Remedies	
5.6 Post-implementation Monitoring	
5.7 Closing Remarks.	10

CHAPTER 6 The Business Case for Accurate Data 101

6.1 The Value of Accurate Data	101
6.2 Costs Associated with Achieving Accurate Data	101
6.3 Building the Business Case	
6.4 Closing Remarks.	11

PART III

Data Profiling Technology. 11

CHAPTER 7 Data Profiling Overview	121
7.1 Goals of Data Profiling	122
7.2 General Model	123
7.3 Data Profiling Methodology	130
7.4 Analytical Methods Used in Data Profiling	136
7.5 When Should Data Profiling Be Done?	140
7.6 Closing Remarks	141
CHAPTER 8 Column Property Analysis	143
8.1 Definitions	143
8.2 The Process for Profiling Columns	152
8.3 Profiling Properties for Columns	155
8.4 Mapping with Other Columns	167
8.5 Value-Level Remedies	169
8.6 Closing Remarks	171
CHAPTER 9 Structure Analysis	173
9.1 Definitions	173
9.2 Understanding the Structures Being Profiled	187
9.3 The Process for Structure Analysis	188
9.4 The Rules for Structure	193
9.5 Mapping with Other Structures	210
9.6 Structure-Level Remedies	212
9.7 Closing Remarks	213
CHAPTER 10 Simple Data Rule Analysis	215
10.1 Definitions	216
10.2 The Process for Analyzing Simple Data Rules	220
10.3 Profiling Rules for Single Business Objects	225
10.4 Mapping with Other Applications	230
10.5 Simple Data Rule Remedies	232
10.6 Closing Remarks	235
CHAPTER 11 Complex Data Rule Analysis	237
11.1 Definitions	237
11.2 The Process for Profiling Complex Data Rules	238

11.3	Profiling Complex Data Rules.	240
11.4	Mapping with Other Applications.	244
11.5	Multiple-Object Data Rule Remedies.	245
11.6	Closing Remarks.	245
c H A p r E R i 2	Value Rule Analysis.	246
12.1	Definitions.	246
12.2	Process for Value Rule Analysis.	247
12.3	Types of Value Rules.	249
12.4	Remedies for Value Rule Violations.	252
12.5	Closing Remarks.	253
c H A P T E R 13	Summary.	255
13.1	Data Quality Is a Major Issue for Corporations.	255
13.2	Moving to a Position of High Data Quality Requires an Explicit Effort	256
13.3	Data Accuracy Is the Cornerstone for Data Quality Assurance....	257
A P P E N D I X A	Examples of Column Properties, Data Structure, Data Rules, and Value Rules.	260«
A.1	Business Objects.	260 j
A.2	Tables.	260
A.3	Column Properties.	263-1
A.4	Structure Rules.	266
A.5	Simple Data Rules.	2691
A.6	Complex Data Rules.	2701
A.7	Value Rules.	271?
A P P E N D I X B	Content of a Data Profiling Repository.	272
B.1	Schema Definition.	272
B.2	Business Objects.	27
B.3	Domains.	27
B.4	Data Source.	27
B.5	Table Definitions.	271
B.6	Synonyms.	27i
B.7	Data Rules.	27
B.8	Value Rules.	27
B.9	Issues.	27

References.	279
Books on Data Quality Issues.	279
Books on Data Quality Technologies.	279
Articles.	281
Index.	283
About the Author.	294