

---

# Contents

<b>1</b>	<b>Introduction to Data Quality</b>	<b>1</b>
1.1	Why Data Quality is Relevant	1
1.2	Introduction to the Concept of Data Quality	4
1.3	Data Quality and Types of Data	6
1.4	Data Quality and Types of Information Systems	9
1.5	Main Research Issues and Application Domains in Data Quality	11
1.5.1	Research Issues in Data Quality	12
1.5.2	Application Domains in Data Quality	12
1.5.3	Research Areas Related to Data Quality	16
1.6	Summary	17
<b>2</b>	<b>Data Quality Dimensions</b>	<b>19</b>
2.1	Accuracy	20
2.2	Completeness	23
2.2.1	Completeness of Relational Data	24
2.2.2	Completeness of Web Data	27
2.3	Time-Related Dimensions: Currency, Timeliness, and Volatility	28
2.4	Consistency	30
2.4.1	Integrity Constraints	30
2.4.2	Data Edits	31
2.5	Other Data Quality Dimensions	32
2.5.1	Accessibility	34
2.5.2	Quality of Information Sources	35
2.6	Approaches to the Definition of Data Quality Dimensions	36
2.6.1	Theoretical Approach	36
2.6.2	Empirical Approach	38
2.6.3	Intuitive Approach	39
2.6.4	A Comparative Analysis of the Dimension Definitions	39
2.6.5	Trade-offs Between Dimensions	40
2.7	Schema Quality Dimensions	42
2.7.1	Readability	45

2.7.2	Normalization .....	45
2.8	Summary .....	48
<b>3</b>	<b>Models for Data Quality .....</b>	<b>51</b>
3.1	Introduction .....	51
3.2	Extensions of Structured Data Models .....	52
3.2.1	Conceptual Models .....	52
3.2.2	Logical Models for Data Description .....	54
3.2.3	The Polygen Model for Data Manipulation .....	55
3.2.4	Data Provenance .....	56
3.3	Extensions of Semistructured Data Models .....	59
3.4	Management Information System Models .....	61
3.4.1	Models for Process Description: the IP-MAP model ...	61
3.4.2	Extensions of IP-MAP .....	62
3.4.3	Data Models .....	64
3.5	Summary .....	68
<b>4</b>	<b>Activities and Techniques for Data Quality: Generalities ...</b>	<b>69</b>
4.1	Data Quality Activities .....	70
4.2	Quality Composition .....	71
4.2.1	Models and Assumptions .....	74
4.2.2	Dimensions .....	76
4.2.3	Accuracy .....	78
4.2.4	Completeness .....	79
4.3	Error Localization and Correction .....	82
4.3.1	Localize and Correct Inconsistencies .....	82
4.3.2	Incomplete Data .....	85
4.3.3	Discovering Outliers .....	86
4.4	Cost and Benefit Classifications .....	88
4.4.1	Cost Classifications .....	89
4.4.2	Benefits Classification .....	94
4.5	Summary .....	95
<b>5</b>	<b>Object Identification .....</b>	<b>97</b>
5.1	Historical Perspective .....	98
5.2	Object Identification for Different Data Types .....	99
5.3	The High-Level Process for Object Identification .....	101
5.4	Details on the Steps for Object Identification .....	103
5.4.1	Preprocessing .....	103
5.4.2	Search Space Reduction .....	104
5.4.3	Comparison Functions .....	104
5.5	Object Identification Techniques .....	106
5.6	Probabilistic Techniques .....	106
5.6.1	The Fellegi and Sunter Theory and Extensions .....	107
5.6.2	A Cost-Based Probabilistic Technique .....	112

5.7	Empirical Techniques .....	113
5.7.1	Sorted Neighborhood Method and Extensions .....	113
5.7.2	The Priority Queue Algorithm .....	116
5.7.3	A Technique for Complex Structured Data: Delphi ....	117
5.7.4	XML Duplicate Detection: DogmatiX .....	119
5.7.5	Other Empirical Methods .....	120
5.8	Knowledge-Based Techniques .....	121
5.8.1	A Rule-Based Approach: Intelliclean .....	122
5.8.2	Learning Methods for Decision Rules: Atlas .....	123
5.9	Comparison of Techniques .....	125
5.9.1	Metrics .....	125
5.9.2	Search Space Reduction Methods .....	127
5.9.3	Comparison Functions .....	127
5.9.4	Decision Methods .....	128
5.9.5	Results .....	130
5.10	Summary .....	131
<b>6</b>	<b>Data Quality Issues in Data Integration Systems .....</b>	<b>133</b>
6.1	Introduction .....	133
6.2	Generalities on Data Integration Systems .....	134
6.2.1	Query Processing .....	135
6.3	Techniques for Quality-Driven Query Processing .....	137
6.3.1	The QP-alg: Quality-Driven Query Planning .....	138
6.3.2	DaQuinCIS Query Processing .....	140
6.3.3	Fusionplex Query Processing .....	141
6.3.4	Comparison of Quality-Driven Query Processing Techniques .....	143
6.4	Instance-level Conflict Resolution .....	143
6.4.1	Classification of Instance-Level Conflicts .....	144
6.4.2	Overview of Techniques .....	146
6.4.3	Comparison of Instance-level Conflict Resolution Techniques .....	156
6.5	Inconsistencies in Data Integration: a Theoretical Perspective .	157
6.5.1	A Formal Framework for Data Integration .....	157
6.5.2	The Problem of Inconsistency .....	158
6.6	Summary .....	160
<b>7</b>	<b>Methodologies for Data Quality Measurement and Improvement .....</b>	<b>161</b>
7.1	Basics on Data Quality Methodologies .....	161
7.1.1	Inputs and Outputs .....	161
7.1.2	Classification of Methodologies .....	164
7.1.3	Comparison among Data-driven and Process-driven Strategies .....	164
7.2	Assessment Methodologies .....	167

- 7.3 Comparative Analysis of General-purpose Methodologies ..... 170
  - 7.3.1 Basic Common Phases Among Methodologies ..... 171
  - 7.3.2 The TDQM Methodology ..... 172
  - 7.3.3 The TQdM Methodology ..... 174
  - 7.3.4 The Istat Methodology ..... 177
  - 7.3.5 Comparisons of Methodologies ..... 180
- 7.4 The CDQM methodology ..... 181
  - 7.4.1 Reconstruct the State of Data ..... 182
  - 7.4.2 Reconstruct Business Processes ..... 183
  - 7.4.3 Reconstruct Macroprocesses and Rules ..... 183
  - 7.4.4 Check Problems with Users ..... 184
  - 7.4.5 Measure Data Quality ..... 184
  - 7.4.6 Set New Target DQ Levels ..... 185
  - 7.4.7 Choose Improvement Activities ..... 186
  - 7.4.8 Choose Techniques for Data Activities ..... 187
  - 7.4.9 Find Improvement Processes ..... 187
  - 7.4.10 Choose the Optimal Improvement Process ..... 188
- 7.5 A Case Study in the e-Government Area ..... 188
- 7.6 Summary ..... 199
- 8 Tools for Data Quality ..... 201**
  - 8.1 Introduction ..... 201
  - 8.2 Tools ..... 202
    - 8.2.1 Potter's Wheel ..... 203
    - 8.2.2 Telcordia's Tool ..... 205
    - 8.2.3 Ajax ..... 206
    - 8.2.4 Artkos ..... 208
    - 8.2.5 Choice Maker ..... 210
  - 8.3 Frameworks for Cooperative Information Systems ..... 212
    - 8.3.1 DaQuinCIS Framework ..... 212
    - 8.3.2 FusionPlex Framework ..... 215
  - 8.4 Toolboxes to Compare Tools ..... 216
    - 8.4.1 Theoretical Approach ..... 216
    - 8.4.2 Tailor ..... 217
  - 8.5 Summary ..... 218
- 9 Open Problems ..... 221**
  - 9.1 Dimensions and Metrics ..... 221
  - 9.2 Object Identification ..... 222
    - 9.2.1 XML Object Identification ..... 223
    - 9.2.2 Object Identification of Personal Information ..... 224
    - 9.2.3 Record Linkage and Privacy ..... 225
  - 9.3 Data Integration ..... 227
    - 9.3.1 Trust-Aware Query Processing in P2P Contexts ..... 227
    - 9.3.2 Cost-Driven Query Processing ..... 228

9.4 Methodologies ..... 230

9.5 Conclusions ..... 235

**References** ..... 237

**Index** ..... 249

Data Quality

Concepts, Methodologies and Techniques

Batini, C.; Scannapieco, M.

2006, XIX, 262 p., Hardcover

ISBN: 978-3-540-33172-8