

Carlo Batini, Monica Scannapieco

Qualità dei Dati: Concetti, Metodi e Tecniche

SPIN Springer's internal project number, if known

– Monograph –

25 febbraio 2008

Springer

Berlin Heidelberg New York

Hong Kong London

Milan Paris Tokyo

Indice

1	Introduzione alla Qualità dei Dati	1
1.1	Perché la Qualità dei Dati è Importante	1
1.2	Introduzione alla Nozione di Qualità dei Dati	5
1.3	Qualità dei Dati e Tipi di Dati	7
1.4	Qualità dei Dati e Tipi di Sistema Informativo	9
1.5	Principali Problemi di Ricerca e Domini Applicativi della Qualità dei Dati	12
1.5.1	Problemi della Ricerca nel Campo della Qualità dei Dati	13
1.5.2	Domini Applicativi della Qualità dei Dati	14
1.5.3	Aree di Ricerca Legate alla Qualità dei Dati	17
1.6	Sommario	19
2	Dimensioni della Qualità dei Dati	21
2.1	Accuratezza	22
2.2	Completezza	26
2.2.1	Completezza dei Dati Relazionali	26
2.2.2	Completezza dei Dati Web	29
2.3	Dimensioni temporali: Aggiornamento, Tempestività e Volatilità	31
2.4	Consistenza	33
2.4.1	Vincoli di Integrità	33
2.4.2	Data Edit	35
2.5	Altre Dimensioni della Qualità dei Dati	36
2.5.1	Accessibilità	38
2.5.2	Qualità delle Sorgenti Informative	39
2.6	Approcci alla Definizione delle Dimensioni di Qualità dei Dati	40
2.6.1	Approccio Teorico	40
2.6.2	Approccio empirico	42
2.6.3	Approccio Intuitivo	43
2.6.4	Analisi Comparativa delle Definizioni delle Dimensioni	43
2.6.5	Trade-off Tra Dimensioni	46

2.7	Dimensioni di Qualità dello Schema	46
2.7.1	Leggibilità	49
2.7.2	Normalizzazione	51
2.8	Sommario	53
3	Modelli per la Qualità dei Dati	55
3.1	Introduzione	55
3.2	Estensioni dei Modelli dei Dati Strutturati	56
3.2.1	Modelli Concettuali	56
3.2.2	Modelli Logici per la Descrizione dei Dati	58
3.2.3	Il Modello Polygen per la Manipolazione dei Dati	58
3.2.4	Provenance dei Dati	60
3.3	Estensione dei Modelli per Dati Semistrutturati	63
3.4	Modelli per i Sistemi Informativi Gestionali	65
3.4.1	Modelli per la Descrizione dei Processi: il modello IP-MAP	65
3.4.2	Estensioni di IP-MAP	67
3.4.3	Modelli per i Dati	69
3.5	Sommario	73
4	Attività e Tecniche Inerenti la Qualità dei Dati: Generalità	75
4.1	Attività Inerenti la Qualità dei Dati	76
4.2	Composizione della Qualità	78
4.2.1	Modelli e Assunzioni	79
4.2.2	Dimensioni	82
4.2.3	Accuratezza	84
4.2.4	Completezza	86
4.3	Localizzazione e Correzione degli Errori	88
4.3.1	Localizzare e Correggere le Inconsistenze	89
4.3.2	Dati Incompleti	91
4.3.3	Scoperta dei Valori Anomali	93
4.4	Classificazioni dei Costi e dei Benefici	95
4.4.1	Classificazioni dei Costi	96
4.4.2	Classificazione dei Benefici	101
4.5	Sommario	102
5	Identificazione degli Oggetti	103
5.1	Cenni Storici	104
5.2	Identificazione degli Oggetti per le Diverse Tipologie di Dati ..	105
5.3	Il Processo di Identificazione degli Oggetti ad Alto Livello ...	107
5.4	Dettagli sui Passi dell'Identificazione degli Oggetti	109
5.4.1	Preprocessing	109
5.4.2	Riduzione dello Spazio di Ricerca	110
5.4.3	Funzioni di Confronto	111
5.5	Tecniche di Identificazione degli Oggetti	112

5.6	Tecniche Probabilistiche	113
5.6.1	La Teoria di Fellegi e Sunter e sue Estensioni	113
5.6.2	Una Tecnica Probabilistica Basata sui Costi	118
5.7	Tecniche empiriche	119
5.7.1	Metodo del Sorted Neighborhood e sue Estensioni	120
5.7.2	L'Algoritmo a Coda di Priorità	122
5.7.3	Una Tecnica per Dati Strutturati Complessi: Delphi ...	123
5.7.4	Scoperta dei Duplicati XML: DogmatiX	126
5.7.5	Altri Metodi Empirici	127
5.8	Tecniche Basate sulla Conoscenza	128
5.8.1	Un Approccio Basato su Regole: Intelliclean	129
5.8.2	Metodi di Apprendimento per le Regole di Decisione: Atlas	130
5.9	Confronto delle Tecniche	133
5.9.1	Metriche	133
5.9.2	Metodi di Riduzione dello Spazio di Ricerca	134
5.9.3	Funzioni di Confronto	135
5.9.4	Metodi Decisionali	135
5.9.5	Risultati	137
5.10	Sommario	138
6	Problemi Inerenti la Qualità dei Dati nei Sistemi di Integrazione dei Dati	141
6.1	Introduzione	141
6.2	Generalità sui Sistemi di Integrazione dei Dati	142
6.2.1	Elaborazione delle Interrogazioni	144
6.3	Tecniche per l'Elaborazione delle Interrogazioni Guidata dalla Qualità	146
6.3.1	Il QP-alg: Pianificazione delle Interrogazioni Guidata dalla Qualità	146
6.3.2	Elaborazione delle Interrogazioni in DaQuinCIS	148
6.3.3	Elaborazione dell'Interrogazione con Fusionplex	150
6.3.4	Confronto tra le Tecniche di Elaborazione dell'Interrogazione Guidata dalla Qualità	152
6.4	Risoluzione dei Conflitti a Livello di Istanza	152
6.4.1	Classificazione dei Conflitti a Livello di Istanza	153
6.4.2	Panoramica delle Tecniche	155
6.4.3	Confronto tra le Tecniche di Risoluzione dei Conflitti a Livello di Istanza	166
6.5	Gestione delle Inconsistenze nell'Integrazione dei Dati: una Prospettiva Teorica	166
6.5.1	Un Framework Formale per l'Integrazione dei Dati	167
6.5.2	Il Problema dell'Inconsistenza	168
6.6	Sommario	170

7	Metodologie per la Misurazione e il Miglioramento della Qualità dei Dati	173
7.1	Fondamenti delle Metodologie per la Qualità dei Dati	173
7.1.1	Input e output	174
7.1.2	Classificazione delle Metodologie	176
7.1.3	Confronto tra Strategie Guidate dai Dati e Strategie Guidate dai Processi	177
7.2	Metodologie per la Valutazione	179
7.3	Analisi Comparativa Delle Metodologie per Scopi Generali	182
7.3.1	Fasi Fondamentali Comuni tra le Metodologie	183
7.3.2	La Metodologia TDQM	185
7.3.3	La Metodologia TQdM	188
7.3.4	La Metodologia Istat	190
7.3.5	Confronto delle Metodologie	193
7.4	La Metodologia CDQM	194
7.4.1	Ricostruire lo Stato dei Dati	195
7.4.2	Ricostruire i Processi Aziendali	195
7.4.3	Ricostruire Macroprocessi e Regole	196
7.4.4	Verificare i Problemi con gli Utenti	197
7.4.5	Misurare la Qualità dei Dati	198
7.4.6	Fissare Nuovi Livelli Target della QD	198
7.4.7	Scegliere le Attività di Miglioramento	199
7.4.8	Scegliere le Tecniche per le Attività dei Dati	200
7.4.9	Individuare i Processi di Miglioramento	201
7.4.10	Scegliere il Processo di Miglioramento Ottimale	202
7.5	Lo Studio di un Caso per l'Area e-Government	202
7.6	Sommario	214
8	Strumenti per la Qualità dei Dati	217
8.1	Introduzione	217
8.2	Strumenti	218
8.2.1	Potter's Wheel	220
8.2.2	Telcordia	221
8.2.3	Ajax	223
8.2.4	Arktos	225
8.2.5	Choice Maker	227
8.3	Framework per Sistemi Informativi Cooperativi	228
8.3.1	Framework DaQuinCIS	230
8.3.2	Framework FusionPlex	232
8.4	Toolbox per il Confronto degli Strumenti	233
8.4.1	Approccio Teorico	233
8.4.2	Tailor	234
8.5	Sommario	236

9 Problemi Aperti	237
9.1 Dimensioni e Metriche	237
9.2 Identificazione degli oggetti	238
9.2.1 Identificazione degli Oggetti XML	239
9.2.2 Identificazione degli Oggetti nel Personal Information Management	240
9.2.3 Record Linkage e Privacy	241
9.3 Integrazione dei Dati	244
9.3.1 Elaborazione delle Interrogazioni Trust-Aware nei Contesti P2P	244
9.3.2 Elaborazione delle Interrogazioni Guidata dai Costi	245
9.4 Metodologie	247
9.5 Conclusioni	252
Riferimenti bibliografici	253
Indice analitico	265



<http://www.springer.com/978-88-470-0733-8>

Qualità dei Dati

Concetti, Metodi e Tecniche

Batini, C.; Scannapieco, M.

2008, XXI, 279 pagg., Softcover

ISBN: 978-88-470-0733-8