

Inhaltsverzeichnis

Abbildungsverzeichnis	XIII
Tabellenverzeichnis	XVII
Abkürzungsverzeichnis	XIX

I Datenbereinigung und Konsolidierung von heterogenen Datenbeständen

– Steven Helmis – 1

1 Einleitung	3
1.1 Motivation	4
1.2 Zielsetzung der Arbeit	5
1.3 Aufbau der Arbeit	5
2 Datenqualität	7
2.1 Datenqualität definieren	7
2.2 Datenfehler	8
2.3 Qualitätskriterien	11
2.4 Methoden zur Einstufung der Qualität	14
3 Dimensionen und Architektur der Informationsintegration	25
3.1 Verteilung	25
3.2 Heterogenität	26
3.3 Autonomie	28
3.4 Integrationsarchitektur	29
4 Data Cleaning	35
4.1 Datenanalyse	36
4.2 Normalisierung und Validierung	39
4.3 Record Matching	40

4.4	Record Merging	42
5	Konzeption des Data Cleaning Toolkits	49
5.1	Bewertung und Analyse existierender Systeme	49
5.2	Anforderungsanalyse	52
5.3	Architektur Data Cleaning Toolkit	54
5.4	Funktionsumfang	55
6	Implementierung	63
6.1	Datenbankentwicklung	63
6.2	Webentwicklung	71
6.3	Probleme während der Implementierungsphase	77
7	Zusammenfassung und Ausblick	79
	Literaturverzeichnis	81
II	Auffinden und Bereinigen von Duplikaten in heterogenen Datenbeständen	
	– Robert Hollmann –	89
8	Einleitung	91
8.1	Motivation	92
8.2	Zielstellungen dieser Arbeit	93
8.3	Gliederung dieser Arbeit	94
9	Informationen, Daten und Wissen- ein Definitionsversuch	95
9.1	Begriffsdefinitionen	96
9.2	Herkunft von Daten und Informationen	98
9.3	Beschaffenheit von Daten und Zugriff auf Informationen	98
10	Informationsintegration im Fokus der Datenqualität	103
10.1	Ist-Stand in Unternehmen- Notwendigkeit der Integration	103
10.2	Informations- und Datenqualität	105
10.3	Sicherung der Datenqualität	114
10.4	Kosten der Datenqualität	115
11	Duplikate in Datenbeständen	117
11.1	Dubletten und deren Identifikation	117

11.2	Ein Framework zur Objektidentifikation	118
11.3	Das Dilemma der Dublettensuche	120
12	Konkrete Verfahren zur Dublettensuche und Klassifikation	125
12.1	Ähnlichkeitsmessungen und Klassifikation	125
12.2	Ähnlichkeitsbestimmung bei Tupeln in einem Datenbestand	126
12.3	Vorselektion für die Dublettensuche	142
13	Konzept der Datenqualitätsanwendung „DCT“	147
13.1	Zielstellung der Applikation	147
13.2	Anforderungsanalyse	148
13.3	Technologiemodell	157
13.4	Datenbankmodell	160
13.5	Applikationsarchitektur	164
13.6	Applikationsstruktur	166
13.7	Entwicklung einer Benutzeroberfläche	169
14	Implementierung, ausgewählte Algorithmen- und Datenstrukturen	173
14.1	„DCT“- Der Verbindungsmanager	173
14.2	„DCT“- Der Workspace-Table Manager	176
14.3	„DCT- Data Profiling“	177
14.4	„DCT“-Plausibilitätskontrolle	180
14.5	„DCT“- Auffinden von Duplikaten	182
15	Fazit und Ausblick	187
	Literaturverzeichnis	189
16	Anhang	195