Steven Helmis | Robert Hollmann

Webbasierte Datenintegration

VIEWEG+TEUBNER RESEARCH

Ausgezeichnete Arbeiten zur Informationsqualität

Herausgeber:

Dr. Marcus Gebauer

Bewertungskomission des
Information Quality Best Master Degree Award 2008:

Prof. Dr. Holger Hinrichs, FH Lübeck (Kommissionsvorsitz)
Dr. Marcus Gebauer, WestLB AG und Vorsitzender der DGIQ
Prof. Dr. Knut Hildebrand, HS Darmstadt
Bernhard Kurpicz, OrgaTech GmbH
Prof. Dr. Jens Lüssem, FH Kiel
Michael Mielke, Deutsche Bahn AG und Präsident der DGIQ
Prof. Dr. Felix Naumann, HPI, Uni Potsdam
Prof. Dr. Ines Rossak, FH Erfurt

Die Deutsche Gesellschaft für Informations- und Datenqualität e.V. (DGIQ) fördert und unterstützt alle Aktivitäten zur Verbesserung der Informationsqualität in Gesellschaft, Wirtschaft, Wissenschaft und Verwaltung. Zu diesem Zweck befasst sie sich mit den Voraussetzungen und Folgen der Daten- und Informationsqualität. Sie fördert zudem durch Innovation und Ausbildung die Wettbewerbsfähigkeit der Unternehmen sowie die des unternehmerischen und akademischen Nachwuchses in Deutschland.

Die vorliegende Schriftenreihe präsentiert ausgezeichnete studentische Abschlussarbeiten in der Daten- und Informationsqualität. Veröffentlicht werden hierin die Siegerarbeiten des jährlich stattfindenden „Information Quality Best Master Degree Award“.

Steven Helmis | Robert Hollmann

Webbasierte Datenintegration

Ansätze zur Messung und Sicherung der Informationsqualität in heterogenen Datenbeständen unter Verwendung eines vollständig webbasierten Werkzeuges

Mit einem Geleitwort von Dr. Marcus Gebauer

VIEWEG+TEUBNER RESEARCH

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Gedruckt mit freundlicher Unterstützung
der Information Quality Management Group

1. Auflage 2009

Lektorat: Christel A. Roß

Vieweg+Teubner ist Teil der Fachverlagsgruppe Springer Science+Business Media.
www.viewegteubner.de

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Druck und buchbinderische Verarbeitung: STRAUSS GMBH, Mörlenbach
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.
Printed in Germany

ISBN 978-3-8348-0723-6

Geleitwort

Als Vorsitzender der Deutschen Gesellschaft für Informations- und Datenqualität (DGIQ e.V.) bin ich glücklich darüber, dass Sie dieses Buch in Ihren Händen halten. Das vorliegende Buch ist Ausdruck unseres Bestrebens, dem wissenschaftlichen Nachwuchs die Möglichkeit zu eröffnen, ihre Arbeiten einem breiten Publikum darstellen zu können.

Dass Sie gerade diese Arbeit vorfinden, ist Ergebnis eines strengen Auswahlprozesses, den die DGIQ mit dem zum ersten Mal ausgeschriebenen „Information Quality Best Master Degree Award 2008“ durchgeführt hat. Studenten waren aufgefordert, ihre Abschlussarbeiten zum Thema Informationsqualität in diesem Wettbewerb durch ihre begutachtenden Professoren einreichen zu lassen. Vertreter aus Wissenschaft, Forschung und Industrie haben diese akademischen Abschlussarbeiten begutachtet.

Die vorliegende „Doppel“-Arbeit von Robert Hollmann und Steven Helmis zeichnet sich insbesondere durch das geschlossene Bild ihrer Forschung aus. Neben der Betrachtung der Datenqualität in der heutigen typischen heterogenen Systemlandschaft, bei gleichzeitig komplizierter werdenden Systemarchitekturen, stehen auch die semantischen Herausforderungen für die Datenqualität im Fokus. Die Darstellung, wie Daten zu konsolidieren, zu bereinigen, und der Kunde dabei auch noch konstruktiv mit Hilfe eines Werkzeuges in diesen Prozess einzubinden ist, ist den Kandidaten herausragend gelungen.

Besonders freue ich mich, dass wir mit dem Verlag Vieweg+Teubner nun die Siegerarbeiten in einer Schriftenreihe jährlich veröffentlichen können. Für die Initiative des Verlages möchte ich mich recht herzlich bedanken.

Offenbach, 27. August 2008

Dr. Marcus Gebauer

Vorwort

Jeder Prozess ist mit der Erzeugung von Daten verbunden. So generieren unzählige heterogene IT-Systeme in immer globaler werdenden Wirtschaftsunternehmen Tag für Tag Millionen von Datensätze, die in richtigen Zusammenhang gebracht, wertvolle Informationen für das Unternehmen und die Entscheidungsfindung in Selbigem von existentieller Bedeutung sein können. Ganzheitliche Sichten auf die Gesamtinformationen sind von großem Wert für Entscheidungsprozesse. Hinzu kommt die überwältigende Informationsflut diverser Internetquellen, die eine Bereicherung für einen Datenbestand bedeuten kann, aber auch Probleme erzeugt und so die Daten(-qualität) in einem Bestand nachhaltig schädigen kann. Diese Heraus- und Anforderungen an Informationssysteme sind nur unter Beachtung von ausreichender Informationsqualität zu erreichen und zu bewältigen. Bei der Integration der heterogenen Datenquellen für eine ganzheitliche Sicht spielt die Qualität der Daten vor- und nach der Integration eine bedeutende Rolle. Datenqualitätsmängel, wie z.B. Duplikate, mangelnde Reputation oder Überalterung bzw. Inkonsistenz können den Informationsgewinn eines integrierten Systems empfindlich stören. Das führt zu falschen bzw. nicht zielgerichteten Entscheidungen deren Grundlage diese „schlechten“ Daten bildeten. Jedes Informationssystem kämpft mit solchen Problemen, die schon im jeweiligen System selbst entstehen, die ihre Manifestation oft jedoch bei der Integration zeigen. Nach und nach werden von Softwareherstellern und Unternehmen diese Probleme und die Vorteile von hoher Datenqualität erkannt. Jedoch besteht auf diesem Gebiet enormer Handlungsbedarf bei Benutzern wie auch Systemherstellern. Vorreiter hier sind einige Unternehmen und Experten, die in einem Zusammenschluss der Deutschen Gesellschaft für Informations- und Datenqualität (DGIQ) diese Probleme offensiv thematisieren, Lösungen vorstellen und für die Problematik sensibilisieren.

Auch die Autoren der vorliegenden Arbeiten konnten so, an einer immer lebendiger agierenden „IQ-Society“ partizipieren. Im Rahmen ihres Studiums und ihren Masterarbeiten befassten sich die Autoren eingehend mit dem Thema Datenqualität und der Analyse und Beseitigung von Datenqualitätsmängeln. Im Ergebnis konnten sie einen Überblick über IST-Stand, Vorgehensweisen und Chancen der Datenqualitässicherung erarbeiten. Gemeinsam entstanden Ansätze zur Umsetzung sowie die prototypische Implementierung eines Datenqualitätswerkzeu-

ges, das ausschließlich auf aktuellen Webtechnologien aufbaut. Mit diesen Ideen konnte der erste Platz des „DGIQ Best Master Degree Award“, der im Jahre 2007 ausgeschrieben wurde, erreicht werden. Mit dieser Veröffentlichung wollen die Autoren zur weiteren Sensibilisierung für gute bzw. schlechte Informationsqualität beitragen und konkrete Lösungen für die Sicherung einer solchen Qualität aufzeigen. Denn nur valide Information schaffen einen Vorteil, den jedes Unternehmen auf einem globalen Markt für sich in Anspruch nehmen möchte. Wir wollen Sie mit unseren Arbeiten dazu motivieren sich aktiv an der Diskussion zum Thema zu beteiligen. Vorteile für sich und Ihr Unternehmen zu erkennen und vielleicht auch Teil der „IQ-Society“ zu werden.

Die Autoren

Steven Helmis, Robert Hollmann

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AHP	Analytic Hierarchy Process
Ajax	Asynchronous JavaScript and XML
ASCII	American Standard Code for Information Interchange
ASP	Application Service Provider
BI	Business Intelligence
BNF	Backus-Naur-Form
CRM	Customer Relationship Management
CSS	Cascading Style Sheets
CSV	Comma Separated Values
CWM	Common Warehouse Metamodel
DB	Datenbank
DBMS	Datenbank Management System
DBS	Datenbank System
DCT	Data Cleaning Toolkit
DD	Data Dictionary
DDL	Data Definition Language
DEA	Data Envelopment Analysis
DGIQ	Deutsche Gesellschaft für Informations- und Datenqualität
DIN	Deutsches Institut für Normung
DML	Data Manipulation Language
DQ	Datenqualität
ER	Entity-Relationship
ERP	Enterprise Resource Planning
ETL	Extraktion, Transformation, Laden

FDH	Free Disposable Hull
GD	gefundene Datensätze die als Duplikat markiert wurden
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IDs	Identifikationen
IQC	Information Quality Criterion
IQS	Information Quality Scores
IR	Information Retrival
IT	Informationstechnologie
JMS	Java Message Service
KMU	kleine und mittelständige Unternehmen
LDAP	Lightweight Directory Access Protocol
MADM	Multi-Attribute Decision-Making
MDBS	Multidatenbanksystem
MVC	Model-View-Controller
OCRA	Operational Competitiveness Rating
ODBC	Open Database Connectivity
OLEDB	Object Linking and Embedding Database
PDMS	Peer Daten Management System
PK	Primary Key
RAID	Redundant Array of Inexpensive Disks
RD	relevante Datensätze, die in der Realität Duplikate darstellen
ROI	Return on Investment

SAW	Simple Additive Weighting
SFA	Stochastic Frontier Analysis
SNMP	Simple Network Management Protocol
SQL	Structured Query Language
SSIS	SQL Server Integration Services
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
Web 2.0	populärer Sammelbegriff für Server- und Client-basierte Internettechnologie
WHIRL	Word-based Heterogenous Information Retrieval Logic
WS	Work-Space
WST	Work-Space-Table
XML	Extensible Markup Language
XQuery	XML Query Language

Teil I

Datenbereinigung und Konsolidierung von heterogenen Datenbeständen

– Steven Helmis –

1 Einleitung

Technischer Fortschritt und Globalisierung führten in den vergangenen Jahren zu einem expandierenden Datenaufkommen. Durch den Einsatz von modernen Datenbanktechnologien wird die effektive Verwaltung und Speicherung der Daten mit verschiedenartigen Strukturen im Giga- und Terabyte-Bereich jedoch beherrschbar. Durch das Kommunikationsmedium Internet beispielsweise erfolgt ein weltweiter Zugriff auf verteilte Informationen. Die Folge ist eine Informationsflut, die neben relevanten auch redundante und inkonsistente Fakten beinhaltet. Möchte man solche Daten in unternehmensinternen Informationssystemen einsetzen, stellen Angaben zur Aktualität oder der Vertrauenswürdigkeit einen relevanten Faktor zur Entscheidungsfindung dar.

Dabei ist die mangelnde Datenqualität der immer wiederkehrende Auslöser, der Customer Relationship Management (CRM) Projekte oder Enterprise Resource Planning (ERP) Systeme versagen lässt oder nicht den zu erwartenden Vorteil erbringt. Falsche, fehlende oder veraltete postalische Informationen sind in vielen Fällen die Ursache von Beeinträchtigungen, die schwerwiegende Konsequenzen nach sich ziehen. Inkorrekte Adressen sind verantwortlich für das Scheitern von internen und externen Kommunikationsprozessen in Unternehmen [UNI03]. Aber nicht nur personengebundene Daten können Mängel aufweisen. Auch bei der Präsentation naturwissenschaftlicher Daten sind Konflikte, die auf Grund von differenzierenden Erfassungsmethoden und diversen syntaktischen und semantischen Heterogenitäten zu einer verminderten Qualität des Datenbestandes führen nicht auszuschließen [MaJBL05].

Betrachtet man die zunehmende Globalisierung in mittelständischen Unternehmen, wächst die Nachfrage an Business Intelligence (BI) Produkten zum Sammeln und Aufbereiten von Daten. Das zeitnahe Darstellen von geschäftsrelevanten Informationen, wie z.B. den Ablauf oder die Ergebnisse laufender bzw. abgeschlossener Geschäftsprozesse fördert strategische Unternehmensentscheidungen. Liegen hier fehlerhafte Informationen vor, sind die Konsequenzen oft mit erhöhten Kosten verbunden. Dieser Aspekt erfordert eine stringente Datenqualitätssicherung während des *ETL-Prozesses*.

1.1 Motivation

Die durch ein BI-Werkzeug erzeugten Informationen bilden den Schlüssel für geschäftliche Innovationen. Kei Shen behauptet in [She06]: „*Unternehmen, die Informationsintegration mit maximaler Effizienz vorantreiben, erzielen mit fünf Mal höherer Wahrscheinlichkeit mehr Wertschöpfung.*“ Diese Aussage wird bekräftigt durch die Studie „Business Intelligence im Mittelstand“ von Dirk Fridrich und Dr. Carsten Bange vom Business Application Research Center [FB07]. In dieser Untersuchung wurden mittelständische deutsche Unternehmen mit einem Jahresumsatz zwischen 50 Millionen und einer Milliarde Euro und einer Mitarbeiteranzahl zwischen 100 und 10000 befragt. 279 ausgefüllte Fragebögen konnten zur Auswertung herangezogen werden. Es zeigte sich, dass bereits 48 % der Befragten Firmen Software zur Unternehmenssteuereung einsetzen und 40 % eine Anschaffung planen. Weiterhin stellte sich heraus, dass Datenqualität die wichtigste Eigenschaft als auch der bedeutendste Kritikpunkt an der Business-Intelligence-Software ist. Durch diese Studie wird ersichtlich, dass Lösungen in den Bereichen Berichtswesen, Planung, Datenanalyse, Budgetierung oder Konsolidierung in Unternehmen betreut bzw. noch benötigt werden. Für eine erfolgreiche Realisierung ist jedoch ein aktueller und qualitativ hochwertiger Datenbestand Voraussetzung.

2006	**gesamt**	**pro Stunde**
Geborene:	672.724	78
Gestorbene:	821.627	95
Eheschließung:	373.681	43
Ehescheidungen (2005):	201.693	23
Zuzüge über die Grenzen Deutschlands:	707.352	82
Fortzüge über die Grenzen Deutschlands:	628.399	73
Gewerbeanmeldungen:	895.144	104

Tabelle 1.1: Statistische Kennzahlen [SÄD07, Bund06]

Anhand der Tabelle 1.1 mit statistischen Kennzahlen der Bundesrepublik Deutschland (vgl. [SÄD07], [Bun06]) wird die Änderungsgeschwindigkeit von adressbezogenen Daten dargestellt. Betrachtet man die rapiden Veränderungen in den verschiedenen Sektoren, ist davon auszugehen, dass Adressdatenbanken keine 100-prozentigen aktuellen Anschriften enthalten. Der momentane Zustand der Daten kann nur über diverse Software, wie sie in Abschnitt 5.1 vorgestellt wird ermittelt und gegebenenfalls modifiziert werden.