Planung und Messung der Datenqualität in Data-Warehouse-Systemen

DISSERTATION
der Universität St. Gallen,
Hochschule für Wirtschafts-,
Rechts- und Sozialwissenschaften (HSG)
zur Erlangung der Würde eines
Doktors der Wirtschaftswissenschaften

vorgelegt von

Markus Helfert

aus

Deutschland

Genehmigt auf Antrag von

Herrn Prof. Dr. Robert Winter

und

Frau Prof. Dr. Andrea Back

Dissertation Nr. 2648

Difo-Druck GmbH, Bamberg 2002

Geleitwort

As information technology continues its relentless innovation, enabling organizations to solve ever new kinds of business problems, and enabling them to capture ever new kinds of information, such as signature, biometrics, video and others, the importance – and complexity – of information quality increases. The increasing dependence on automation is a two-edged sword, transforming work in ways unimaginable ten years ago, and yet increasing the potential for enterprise failure if information is not managed as a strategic resource and the processes that produce it are not managed as production processes with quality management principles applied.

As information quality and data quality become the new buzzwords many organizations are undertaking IQ or DQ projects in some shape or form, whether in the context of data warehousing or CRM or e-business. But there are many attempts to apply the words „data quality“ or „information quality“ to practices that are NOT „quality“ management practices. Some practices merely automate and *institutionalize* information scrap and rework. Other practices measure information „quality“ but fail to provide a true measure of *accuracy*. Still other practices seek to provide „data quality“, but actually create a brand new information quality problem of „inconsistency“ by correcting data in the data warehouse while leaving data uncorrected in the source data files, resulting in continued failure of processes accessing data from the source, inability to support drill down from aggregated data to source data, and leaving defective data that can potentially contaminate the data warehouse in subsequent changed data propagation.

The advent of the third millennium brings humankind closer to what I call the „realized“ Information Age, in which we understand the paradigm shift and implement the principles of the Information Age. To survive and thrive in the new

economy of the realized Information Age, an organization must implement an *effective* information quality management system. A true information or data quality management system has several essential – and non-optional – ingredients:

1. It understands information quality is a *business problem*, not just a systems problem, and solves it as a business process, not just as a systems process.
2. It focuses on the *information customers* and *information suppliers*, not just the data.
3. It focuses on all components of information, including definition, content and presentation.
4. It implements information quality management *processes*, not just information quality software.
5. It measures data *accuracy*, not just validity.
6. It *measures costs* – not just percent – of nonquality information, and business results of quality information.
7. It is proactive, emphasizing process improvement and preventive maintenance (Plan-Do-Check-Act), not just corrective maintenance (data cleansing) to eliminate the costs of information scrap and rework.
8. It improves processes *at the source*, not just in downstream business areas and results in reduced costs of process failure and „information scrap and rework“ as well as increased business effectiveness and opportunity gain.
9. It provides quality *training* to managers and information producers (who are my information customers and what do they need).
10. It actively *transforms* the *culture*, not just implements activities or slogans.

I fully believe that organizations that embark implementing a proactive information quality environment will produce the same kind of economic revolution that Japan did in the 1970s and 80s. The Japanese quality revolution changed the ground rules for competing in the manufacturing markets. Organizations that ignore the

IQ revolution will be at risk as their competitors who understand the imperatives of Information Quality Management implement it – and benefit from it.

Larry P. English
President and Principal, INFORMATION IMPACT International, Inc.
Author, *Improving Data Warehouse and Business Information Quality*.

Vorwort

Die vorliegende Arbeit entstand im wesentlichen im Rahmen meiner Tätigkeit in den Kompetenzzentren „Data Warehousing Strategie“ und „Data Warehousing 2“ am Institut für Wirtschaftsinformatik der Universität St. Gallen. In den Forschungsprojekten wurden und werden vielfältige Probleme im Bereich von Data-Warehouse-Systemen untersucht und intensiv mit verschiedenen Partnerunternehmen diskutiert. Ein erfolgskritischer Problembereich für Data-Warehouse-Systeme stellt die Sicherstellung einer angemessenen Datenqualität dar. Dieses, inzwischen zwar häufig als relevant eingeschätztes Thema, findet sowohl in der Wissenschaft als auch in der Praxis noch immer nicht die notwendige Bedeutung. Ausgehend von den Arbeiten im „Total Data Quality Management Program“ am Massachusetts Institute of Technology (MIT) und den Arbeiten von Larry English wurden insbesondere Managementaspekte zur Sicherstellung einer angemessenen Datenqualität untersucht sowie ein Ansatz zur Planung und Messung der Datenqualität entwickelt. Zentral ist hierbei die proaktive Verhinderung von Ursachen mangelnder Datenqualität.

Die Erstellung der Arbeit haben zahlreiche Personen ermöglicht und unterstützt, bei denen ich mich herzlich bedanke. Mein besonderer Dank gilt Herrn Prof. Dr. Robert Winter, der die Arbeit von der Themenfindung bis zur Veröffentlichung betreute und mir die entsprechenden Möglichkeiten am Institut für Wirtschaftsinformatik bot. Bedanken möchte ich mich auch sehr bei Frau Prof. Dr. Andrea Back für ihr Interesse am Thema, den inhaltlichen Diskussionen sowie der Übernahme des Korreferates. Herrn Prof. Dr. Reinhard Jung, der mich bei der inhaltlichen Fokussierung im Rahmen des Kompetenzzentrums „Data Warehousing Strategie“ durch eine freundschaftliche Zusammenarbeit unterstützt hat, gilt insbesondere mein Dank. Weiter möchte ich seinem Nachfolger, Herrn Dr. Eitel von Maur danken.

Die Zusammenarbeit mit Unternehmen und Praktikern bildet die Grundlage anwendungsorientierter Forschung. Durch vielzählige Diskussionen während Workshops, im Rahmen der Projektarbeit oder beim Data-Warehouse-Forum St. Gallen konnten viele Anregungen für meine Arbeit gewonnen werden. Insbesondere danke ich den Mitarbeitern der Credit Suisse Financial Services, die mir ein ideales Arbeitsumfeld zur Reflexion und Umsetzung meiner Forschungsergebnisse boten. Stellvertretend möchte ich mich namentlich bei Hans Wegener, Stefan Meissner, Engelbert Schollenberger und Marcel Winter besonders bedanken.

Der Erfolg einer Dissertation wird durch die Betreuung während der Promotion wesentlich beeinflusst. Allerdings gilt mein persönlicher Dank auch Prof. Dr. Helmut Merkel von der Universität Mannheim. Bereits während des Studiums und der Betreuung meiner Diplomarbeit hat er meine Vorgehensweise geprägt und mich zur Promotion ermutigt. Insbesondere hat er eine intensive Auseinandersetzung mit der Systemtheorie gefördert und mir die Möglichkeit zum selbständigen wissenschaftlichen Arbeiten geboten.

Ein ganz persönlicher und herzlicher Dank gilt meinen Kollegen und Freunden Dr. Bernd Heinrich, Gregor Zellner, Silvia Choinowski sowie Gunnar Auth, Thomas Stiffel, Clemens Herrmann und Alexander Schwinn für die schöne und witzige Zeit in St. Gallen. Bedanken möchte ich mich auch bei allen Kollegen am Lehrstuhl für die freundschaftliche Zusammenarbeit und die konstruktiven Anregungen. Insbesondere möchte ich Dr. Susanne Leist, Dr. Thorsten Frie, Dr. Bernhard Strauch und Dr. Stefan Schwarz sowie den Mitarbeitern des Kompetenzzentrums „Data Warehousing Strategie“ bzw. „Data Warehousing 2“ danken.

Für die Durchsicht meiner Arbeit möchte ich besonders Gregor Zellner, Silvia Choinowski, Gunnar Auth, Clemens Herrmann, Samuel von Siebenthal sowie Stephan Böhm herzlich danken. Bei zahlreichen Fragen bezüglich des Textverarbeitungssystems LaTeX unterstützten mich vor allem Dr. Joachim Schelp und Thomas Fugmann. Für ihre Hilfe möchte ich mich ebenfalls herzlich bedanken. Ein besonderer persönlicher Dank gilt meiner Partnerin Elfriede Krauth für die Geduld, Motivation und Unterstützung, die engagierte Durchsicht der Arbeit sowie die Hinweise stilistischer Natur.

Abschliessend möchte ich einen meiner treuesten Begleiter während der Disserta-

tion in St. Gallen erwähnen. Jeden morgen begrüsste er mich und begleitete mich durch den Tag. In schwierigen, einfallslosen Zeiten stand er zu mir und machte mich auf den Wechsel der Jahreszeiten aufmerksam: Der Baum vor meinem Fenster.

Mai 2002 Markus Helfert

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

API Application Programming Interface (Programmierschnittstelle)

ARIS Architektur integrierter Informationssysteme

BDB Bereichsdatenbank

bzgl. bezüglich

bzw. beziehungsweise

CC DW2 Kompetenzzentrum Data Warehousing 2

CLIQ Data Cleansing mit intelligentem Qualitätsmanagement

CORBA Common Object Request Broker Architecture

CWM Common Warehouse Metamodell

d. h. das heisst

DCE Distributed Computing Environment

DCOM Distributed Component Object Model

DIN Deutsches Institut für Normung

DQ Datenqualität

DQM Datenqualitätsmanagement

DV Datenverarbeitung

DWH Data-Warehouse

DWQ Forschungsprojekt Foundations of Data Warehouse Quality

Email elektronische Post

ER Entity-Relationship

et al. et alii

etc. et cetera

ETL Extraktion, Transformation und Laden

f. folgende

ff. fortfolgende

i. d. R. in der Regel

i. e. S. im engeren Sinne

IS Informationssystem

ISO International Organization for Standardization

KOI Kontrollinformation

MDX Multidimensional Expressions

MIT Massachusetts Institute of Technology

OLAP Online analytical processing

OMG Object Management Group

QuAsAR Quality Assessment using Association Rules

RW Realwelt

SIZ Informatikzentrum der Sparkassenorganisation

SMS Short message service

SOI Sollwertinformation

SPC statistische Prozesskontrolle

SQL Structured Query Language

TDQM Total Data Quality Management

TQM Total Quality Management

u. a. unter anderem

u. und

UML Unified Modeling Language

usw. und so weiter

vgl. vergleiche

XML Extensible Markup Language

z. B. zum Beispiel

zugl. zugleich

Kapitel 1

Einleitung

1.1 Motivation und Problemstellung

Der wirtschaftliche Rahmen für unternehmerisches Planen und Handeln ist in den letzten Jahren zunehmend komplexer geworden. Sowohl das Unternehmensumfeld als auch die unternehmensinternen Strukturen und Abläufe sind dabei äusserst vielschichtig und dynamischen Veränderungen unterworfen. Zur Begegnung dieser Herausforderungen hat sich sowohl aus unternehmensexterner wie auch aus unternehmensinterner Sicht die Fähigkeit zur Sicherstellung einer optimalen Informationsversorgung als entscheidende Voraussetzung gezeigt. Informationen haben sich so zunehmend zu einem wettbewerbsentscheidenden Produktionsfaktor entwickelt. Informationen werden in betrieblichen Systemen zur Durchführung der Planungs-, Steuerungs- und Kontrollaufgaben für Leistungsprozesse benötigt und verarbeitet. Die Qualität dieser Eingangsinformationen beeinflusst damit direkt das Ergebnis der Informationsverarbeitungsaufgaben. Daher ist eine optimale Informationsversorgung erforderlich, die sicherstellt, dass Informationen in angemessener Qualität für die entsprechenden Zwecke zur Verfügung stehen.

Wenngleich in den letzten Jahren durch technische Entwicklungen die Möglichkeiten einer umfassenden Informationsversorgung zunahmen, finden sich allerdings vermehrt Aussagen wie beispielsweise:

- „Ein Datenfriedhof ist keine Informationsquelle“.[1]

[1] Vgl. Gertz (1999), S. 49.

- „Operative Datenqualität nicht überschätzen“.[2]
- „Informationsarmut im Informationsüberfluss.“[3]

Diese oder ähnliche Aussagen deuten darauf hin, dass die verfügbaren Informationen die Informationsbedarfe nicht zufriedenstellend und optimal abdecken.

Zur Informationsversorgung betrieblicher Systeme haben sich seit geraumer Zeit Informationssysteme und insbesondere Anwendungssysteme für unterschiedliche Anwendungsgebiete und Personengruppen entwickelt. So ist es auch nicht verwunderlich, dass sich seit den Anfängen der elektronischen Datenverarbeitung Ansätze zur Informationsversorgung des Managements entwickelt haben.[4] Diese Systeme zielen auf die problemadäquate Bereitstellung von Informationen für Entscheidungsträger ab. Nach anfänglichen Misserfolgen gab es in den letzten Jahren verschiedene Versuche, solche Systeme in den Unternehmen zu etablieren. Allerdings hat ein unter dem Begriff des „Data Warehousing“ diskutiertes Konzept erst in den letzten Jahren breiten Einzug in die Praxis gehalten. Dieses Konzept wird in Bezug auf unterschiedliche Fragestellungen seit einiger Zeit in der Wissenschaft untersucht. Es hat sich weitgehend als Kern einer integrierten Informationslogistik etabliert. Zentraler Gedanke ist eine von den operativen Systemen separate Speicherung der entscheidungsorientierten Daten. Bislang standen bei der Diskussion technische Aspekte und Architekturkonzepte sowie Fragestellungen der Datenmodellierung im Vordergrund. Fragen der organisatorischen Gestaltung und der methodischen Unterstützung wurden in den letzten Jahren Gegenstand von Forschungsansätzen. Eine bislang weniger tiefgreifende Auseinandersetzung erfolgte mit Fragestellungen hinsichtlich der Sicherstellung von Daten in hochwertiger Qualität. Zwar wird die Wichtigkeit des Themas in vielzähligen Publikationen erwähnt,[5] jedoch fehlen bislang Konzepte für das Datenqualitätsmanagement, operationalisierte Qualitätskriterien und Möglichkeiten, diese in bisherige Data-Warehouse-Systemarchitekturen zu integrieren. Dies ist um so überraschen-

[2] Vgl. Soeffky (1999), S. 8.

[3] Vgl. Nieschlag, Dichtl und Hörschgen (1994), S. 1005.

[4] Vgl. z. B. Chamoni und Gluchowski (1998), S. 6.

[5] Vgl. z. B. Keppel, Müllenbach und Wölkhammer (2001), S. 103f.; Conrad (2000), S. 292ff.; Müller (2000), S. 14-17; Schwinn, Dippold, Ringgenberg und Schnider (1999), S. 209ff.; Holthuis (1999), S. 33ff.; Soeffky (1998), S. 49f.

der, als die Sicherstellung der Datenqualität häufig als ein wesentlicher Erfolgsfaktor für die Umsetzung von Data-Warehouse-Systemen genannt wird.[6] Ohne eine ausreichende Datenqualität kann das Data-Warehouse-System seine Nutzenpotentiale als Lieferant entscheidungsrelevanter Daten allerdings nicht ausschöpfen.

Aufgrund mangelnder Datenqualität können die an das Data-Warehouse-System gestellten Anforderungen, insbesondere an die Aussagekraft der Daten, häufig nicht oder nur teilweise erfüllt werden. Dies führt letztendlich nicht selten zum Scheitern gesamter Data-Warehouse-Projekte. Nach einer 1999 durchgeführten Umfrage wird die Sicherstellung der Datenqualität bei nahezu allen Unternehmen als problematisch eingeschätzt.[7] Insbesondere sind die Durchsetzung von Massnahmen zur Datenqualitätssicherung in den operativen Systemen sowie semantische Aspekte erfolgskritisch und schwer zu lösen. Mangelnde Datenqualität führt nicht nur zur aufwendigen Suche nach korrekten Werten und nachträglichem Aufwand bei der Datenbereinigung sondern hat darüber hinaus weitere Auswirkungen. So führt mangelnde Datenqualität nicht selten zur Verringerung der internen Akzeptanz des Data-Warehouse-Systems, zu nicht optimalen Entscheidungen sowie unzureichender Unterstützung operativer Geschäftsprozesse. In Tabelle 1.1 sind beispielhaft einige Auswirkungen mangelnder Datenqualität aufgeführt, die im Rahmen eines Workshops mit Partnerunternehmen diskutiert wurden.[8] Bereits diese wenigen Beispiele verdeutlichen die Relevanz des Themenkomplexes der Datenqualität, nicht nur für entscheidungsorientierte Systeme.

Die Daten eines Data-Warehouse-Systems werden aus verschiedenen Vorsystemen entnommen und durch Transformationsprozesse zu einer integrierten Datenbasis des Unternehmens zusammengeführt. Als Liefersysteme agieren unterschiedliche unternehmensinterne operative Systeme als auch teilweise unternehmensexterne Systeme, die sich in ihren jeweiligen Zielsetzungen meist unterscheiden. Oft überschneiden sich die betrachteten Themenfelder, wobei aufgrund der unterschiedlichen Zielsetzungen und der historischen Entwicklung in der Regel syntaktisch und semantisch heterogene Datenbestände vorherrschen. Im Trans-

[6] Vgl. Häussler (1998), S. 75; Watson und Haley (1998), S. 38.

[7] Vgl. Helfert (2000b), S. 13.

[8] Der Workshop fand am 28. und 29. Oktober 1999 in Irsee (Deutschland) im Rahmen des Kompetenzzentrums „Data Warehousing Strategie“ mit 29 Teilnehmern statt.

Kategorie	Beispiel
Zusatzaufwand	Aufwendige Suche nach den richtigen Werten
	Nachträglicher Aufwand beim Erstellen von Analysen und Berichten
	Doppelerfassungen
	Aufwendige Transformationslogik (Entwicklungs- und Betriebsaufwand)
Interne Akzeptanz	Unglaubwürdigkeit
	Interner Imageverlust
	Erwarteter Nutzen wird nicht erreicht
	Nur von Spezialisten nutzbar
Unterstützung operativer Prozesse	Kundenbeschwerden, Kundenabwanderungen
	Ansprache der falschen Zielgruppe
	Ungenutzte Cross-Selling-Möglichkeiten
	Falsche Provisions- und Prämienberechnungen
Entscheidungs-prozesse	Ansammlung unerwünschter Risiken
	Falsche Tarif- und Preiskalkulation
	Ungenaue Rentabilitätsberechnungen
	Falsche strategische Ausrichtung

Tabelle 1.1: Auswirkungen mangelnder Datenqualität

formationsprozess zwischen den Vorsystemen und der zentralen Data-Warehouse-Datenbasis wird im allgemeinen versucht, diese Heterogenitäten zu vereinheitlichen und zu einem konsistenten, integrierten Datenbestand zusammenzuführen. Heterogenitäten zwischen Datenmodellen lassen sich meist mit Hilfe einer aufwendigen, als Datenbereinigung bezeichneten Transformationslogik ausgleichen. Es zeigt sich jedoch, dass sich so nicht alle Probleme lösen lassen. Insbesondere lassen sich inkorrekte oder zeitlich inkonsistente Datenwerte nur bedingt durch Bereinigungsmassnahmen verbessern. Deren Ursachen liegen vielmehr in den operativen Vorsystemen und deren Erfassungsprozessen begründet. Eine mögliche Lösung des Problems stellt die Betrachtung des Gesamtprozesses der Datenentstehung bis hin zur Datenverwendung mit allen damit zusammenhängenden Aktivitäten hinsichtlich qualitativer Zielsetzungen dar. Es gilt Ursachen mangelnder Datenqualität zu identifizieren und adäquate Massnahmen zur Qualitätsver-

besserung zu ermitteln. Diese Sichtweise auf den gesamten Datenfluss und die primäre Ursachenbeseitigung wurde bislang in den Data-Warehouse-Konzepten nicht in dem Masse betrachtet, als es zur Sicherstellung qualitativ hochwertiger Daten notwendig wäre.

Bevor allerdings Verbesserungsmassnahmen zu identifizieren sind, ist die Frage der Art qualitativ hochwertiger Daten zu untersuchen. Leider besteht bislang wenig Einigkeit diesbezüglich und über den Begriff der Datenqualität an sich. Der Begriff wird über zahlreiche Datenqualitätsmerkmale beschrieben, dessen Ausprägungen im allgemeinen anwendungsspezifisch sind. Weiter zeigt sich, dass Qualitätsforderungen vom jeweiligen Aufgabenzweck als auch von Anspruchsgruppen abhängen. Daher sind im Vorfeld der Sicherstellung qualitativ hochwertiger Daten zunächst projektspezifische Anforderungen an diese zu erheben und festzulegen. Im allgemeinen Qualitätsmanagement wird dieser Aufgabenbereich als Qualitätsplanung bezeichnet. Nach Spezifikation qualitativ hochwertiger Daten, können diese im Rahmen der Qualitätslenkung durch Qualitätsmessungen überprüft und geeignete Massnahmen zur Qualitätssicherung abgeleitet werden. Hier sind Möglichkeiten zur Qualitätsspezifikation und deren Messung von zentralem Interesse, die zusammen ein Datenqualitätsmodell bilden. Diese beiden operativen Aufgabenbereiche bilden eine zentrale Komponente für das ganzheitliche Management von qualitativ hochwertigen Daten, welches im Zentrum dieser Arbeit als Konzept eines *proaktiven Datenqualitätsmanagements* für Data-Warehouse-Systeme betrachtet wird.

1.2 Zielsetzung und Zielgruppe der Arbeit

Aufgrund der hohen Relevanz des Problemkomplexes und der bislang vernachlässigten Berücksichtigung in Forschungsansätzen untersucht die Arbeit Fragestellungen im Bereich der Datenqualität für Data-Warehouse-Systeme. Ziel der Arbeit ist eine umfassende Problemanalyse zu den Fragen der Datenqualität, der Qualitätsplanung und Qualitätslenkung sowie die Entwicklung eines Konzeptes zum Management qualitativ hochwertiger Daten in Data-Warehouse-Systemen. Aufbauend auf diesem konzeptionellen Rahmen sollen die zentralen Aufgabenbe-

reiche der Qualitätsplanung und Qualitätslenkung in Data-Warehouse-Systemen untersucht und anhand eines Fallbeispiels konkretisiert werden. Hierzu sind insbesondere Datenqualitätsmerkmale auszuwählen, zu klassifizieren und zu gewichten sowie die Datenqualität durch Qualitätsprüfungen in quantitative Kennzahlen auszudrücken. Die Operationalisierung der Qualitätsaussagen ermöglicht es, sowohl die derzeitige als auch die sich entwickelnde Datenqualität abzuschätzen. Damit wird eine Planung und Kontrolle der Datenqualität im Sinne eines Datenqualitätsmanagements möglich. Mit Hilfe der Qualitätsaussagen können den Datenverwendern Angaben über die derzeitige Datenqualität zur Verfügung gestellt werden. Anhand der Angaben können diese dann ihre Entscheidungen besser abschätzen.

Aufgrund der sowohl für die Wissenschaft als auch in praktischen Anwendungsfällen interessanten Fragestellung richtet sich die Arbeit an Wissenschaftler und Mitarbeiter in Unternehmen. Es sind Personen angesprochen, die sich im Bereich der Informationssysteme und des Informationsmanagements sowie des Qualitätsmanagements betätigen. Insbesondere ist die Arbeit für Forscher im Bereich von Data-Warehouse-Systemen und Datenqualität interessant, wobei sich die Forschungsergebnisse auch auf andere Forschungsbereiche übertragen lassen. Als weitere Zielgruppe ist die Arbeit für Mitarbeiter bei Unternehmen in den Bereichen des Informationsmanagements und der Data-Warehouse-Systeme relevant. Die Arbeit ist für Mitarbeiter mit und ohne Führungsverantwortung und externe Projektmitglieder in der Rolle des Projektverantwortlichen, des Technikspezialisten als auch des Analysten von Nutzen. Wenngleich sich die Arbeit vorwiegend an die Verantwortlichen zur Sicherung der Datenqualität richtet, sind alle Bereiche von Data-Warehouse-Systemen betroffen. Weiter sind die Ergebnisse der Arbeit für Mitarbeiter in Software- und Beratungsunternehmen relevant.

1.3 Forschungsmethodische Grundlagen

Bislang existiert keine Einigkeit über die in der Wirtschaftsinformatik anzuwendenden Forschungsansätze.[9] Welche Forschungsmethoden anzuwenden sind,

[9] Vgl. Frank, Klein, Krcmar und Teubner (1999), S. 71.

hängt in erster Linie von den verfolgten Forschungszielen, der Forschungsthematik sowie der Forschungsfrage ab. Forschungsziele der Wirtschaftsinformatik sind sowohl erkenntnis- als auch handlungsorientiert, wenngleich i. d. R. ein hoher Anwendungsbezug befürwortet wird.[10] Während beim theoretischen Wissenschaftsziel Erkenntnisse über den Gegenstand mit Hilfe wissenschaftlicher Theorien gewonnen werden, zeichnet sich die handlungsorientierte Wissenschaft durch einen hohen Pragmatismus aus. Es sollen Erkenntnisse zur Gestaltung einer technischen und sozialen Realität durch Entwicklung von Handlungsanleitungen und Instrumenten gewonnen werden. Die vorliegende Arbeit verfolgt in erster Linie Gestaltungsziele im Rahmen anwendungsbezogener Forschung, die sich durch folgende Eigenschaften auszeichnet:[11]

- Die untersuchten Probleme entstehen in der Praxis.
- Angewandte Forschung ist interdisziplinär.
- Das Forschungsziel besteht in der Gestaltung der betrieblichen Realität durch die Erarbeitung konkreter Handlungsanweisungen und Modelle.
- Die Aussagen sind normativ und wertend.
- Das Fortschrittskriterium ist die praktische Problemlösungskraft der Modelle und Handlungsanweisungen.[12]

Wissenschaftstheoretisch ist die Arbeit dem methodischen Konstruktivismus zuzuordnen, der im Gegensatz zum kritischen Rationalismus ein korrespondenztheoretisches Wahrheitsverständnis ablehnt.[13] Es wird vielmehr eine Konzeption vertreten, die Gegenstände der Wissenschaften als Konstruktion zweckgerichteten menschlichen Handelns versteht. Etwas gilt genau dann als wahr, wenn in einem unvoreingenommenen Diskurs (in idealer Sprechsituation) jeder Sachkundige und Gutwillige zustimmen kann. Eine wesentliche Forderung um Theorien im

[10] Vgl. Schütte (1998), S. 11ff.; Müller-Böling und Klandt (1996), S. 5.
[11] Vgl. Ulrich (1984), S. 202f.
[12] Vgl. Österle, Brenner und Hilbers (1991), S. 35.
[13] Vgl. hierzu im folgenden Schütte (1998), S. 16-25.

Konstruktivismus nachvollziehbar zu machen ist die Bildung einer Wortgemeinschaft auf die für die Formulierung der Basissätze einer Theorie relevanten Begriffe und der zulässigen Ableitungsregeln. Als methodisches Prinzip wird auch im Konstruktivismus eine kritische Prüfung akzeptiert. Allerdings entstammen die Massstäbe, im Gegensatz zum kritischen Rationalismus, nicht nur aus dem Untersuchungsgegenstand selbst, sondern auch aus Begründungsrekonstruktionen.

Wird ein angewandtes Forschungsziel angestrebt, so eignen sich besonders qualitative Ansätze, die sich durch einen hohen Grad an Interdisziplinarität, durch eine verstärkte Verbindung der Wissenschaft mit der Praxis, durch eine Problem- und Handlungsorientierung, durch zyklisches Vorgehen und durch die Partizipation seitens der Praxis auszeichnen.[14] Grundlage der Arbeit bilden daher Erfahrungen, die im Rahmen von Workshops und der Projektarbeit mit Partnerunternehmen des Kompetenzzentrums „Data Warehousing Strategie“ sowie dessen Nachfolgeprojekt „Data Warehousing 2“ durchgeführt wurden.[15] Im Sinne einer Exploration des Untersuchungsgegenstandes[16] und zur kritischen Reflexion der erarbeiteten Konzepte steht eine Fallstudie einer Schweizer Universalbank im Zentrum der Arbeit. Durch die intensive Projektarbeit konnten die Konzepte diskutiert und weiter verfeinert sowie deren Realisierungsmöglichkeiten untersucht werden.

Häufige Kritik, die sich Forschung anhand von Fallstudien (insbesondere Einzelfallstudien) aussetzt, ist die der mangelnden Eignung generalisierbare Ergebnisse zu liefern. Fallstudien stellen weder repräsentative Beispiele einer bestimmten Grundgesamtheit dar, noch sagen sie etwas über die Häufigkeitsverteilung der beschriebenen Phänomene in der Realität aus. Sie erscheinen ebenfalls zur Plausibilisierung bestimmter Theorien oder Modelle ungeeignet. Ihre Berechtigung haben sie aber bei der Neu- und Weiterentwicklung von Modellen und Theorien.[17] Deren Informationsgehalt ist im Hinblick auf pragmatische Aussagen extrem hoch,

[14] Vgl. Probst und Raub (1995), S. 8ff.

[15] Im folgenden wird die Bezeichnung „Kompetenzzentrum Data Warehousing 2“(CC DW2) verwendet, wobei allerdings jeweils beide Forschungsprojekte angesprochen werden sollen. Das Kompetenzzentrum „Data Warehousing Strategie“wurde 1999 als zweijähriges Forschungsprojekt am Institut für Wirtschaftsinformatik der Universität St. Gallen mit zentralen Fragestellungen im Data Warehousing gegründet. Seit dem Jahr 2001 führt das Kompetenzzentrum „Data Warehousing 2“ als Nachfolgeprojekt die Forschungsarbeiten fort.

[16] Vgl. Lamnek (1995), S. 10; Kromrey (1998), S. 520.

[17] Vgl. Yin (1994), S. 10.

wenngleich der Allgemeinheitsgrad sich i. d. R. geringer darstellt.[18] Daher erhebt die vorliegende Arbeit auch nicht den Anspruch auf „statistische Generalisierbarkeit“, sondern bezweckt die Entwicklung eines Konzeptes für das Datenqualitätsmanagement im Sinne einer eher „analytischen Generalisierbarkeit“.[19] Als Instrumentarien zur Erkenntnisgewinnung standen vor allem die Durchführung von Interviews und Workshops, die direkte und teilnehmende Beobachtung sowie die Untersuchung von Werkzeugen und Softwarelösungen im Rahmen der Arbeit innerhalb des Kompetenzzentrums und der Projektarbeit zur Verfügung.

1.4 Aufbau der Arbeit

Der grundsätzliche Aufbau der Arbeit ist in Abbildung 1.1 dargestellt. Zunächst werden in Kaptitel 2 die konzeptionellen Grundlagen der Arbeit beschrieben. Nachdem die begrifflichen Grundlagen von Daten und Informationen aufgezeigt werden, wird auf die Architektur betrieblicher Informationssysteme eingegangen. Aufgrund der zentralen Bedeutung in der Arbeit werden Datenhaltungssysteme in Abschnitt 2.3 näher betrachtet. Anschliessend befasst sich Abschnitt 2.4 mit dem zentralen Untersuchungsgegenstand der Arbeit, den Data-Warehouse-Systemen als analytische Informationssysteme.

Kapitel 3 betrachtet die Datenqualität und das Management qualitativ hochwertiger Daten. Zunächst werden der Qualitätsbegriff und die Qualitätssichten sowie ausgewählte Ansätze zum Begriff der Datenqualität erörtert. Anschliessend werden Aspekte der Datenqualität in Data-Warehouse-Systemen näher analysiert und anhand einer empirischen Untersuchung die Problematik und Lösungsansätze in der Praxis dargestellt. Aufbauend auf diesen Erkenntnissen wird in Abschnitt 3.4 ein Konzept für ein proaktives Datenqualitätsmanagement erarbeitet. Als Kern des Konzeptes wird anschliessend das operative Datenqualitätsmanagement untersucht und anhand der grundsätzlichen Anforderungen mit ausgewählten Ansätzen in der Literatur verglichen.

Ein Ansatz für ein operatives Datenqualitätsmanagement wird in Kapitel 4 be-

[18] Vgl. Müller-Böling und Klandt (1996), S. 87.
[19] Vgl. Yin (1994), S. 10.

schrieben, der im Rahmen der Projektarbeit mit einer Schweizer Universalbank erarbeitet wurde. Zunächst werden die Rahmenbedingungen und die zentralen Problemfelder im Bereich der Datenqualität sowie die wesentlichen Datenqualitätsforderungen dargestellt. Anschliessend wird ein in die Metadatenverwaltung integriertes Datenqualitätssystem beschrieben und insbesondere auf die Spezifikation und Prüfung der Ausführungsqualität eingegangen. Zusammenfassend werden dann in Abschnitt 4.2.3 die erarbeiten Möglichkeiten zur Messung der Ausführungsqualität aufgezeigt und exemplarisch Qualitätskennzahlen genannt. Abschliessend werden die Ergebnisse in Kapitel 5 kritisch beleuchtet und Ansätze für weitere Forschungsarbeiten aufgezeigt.

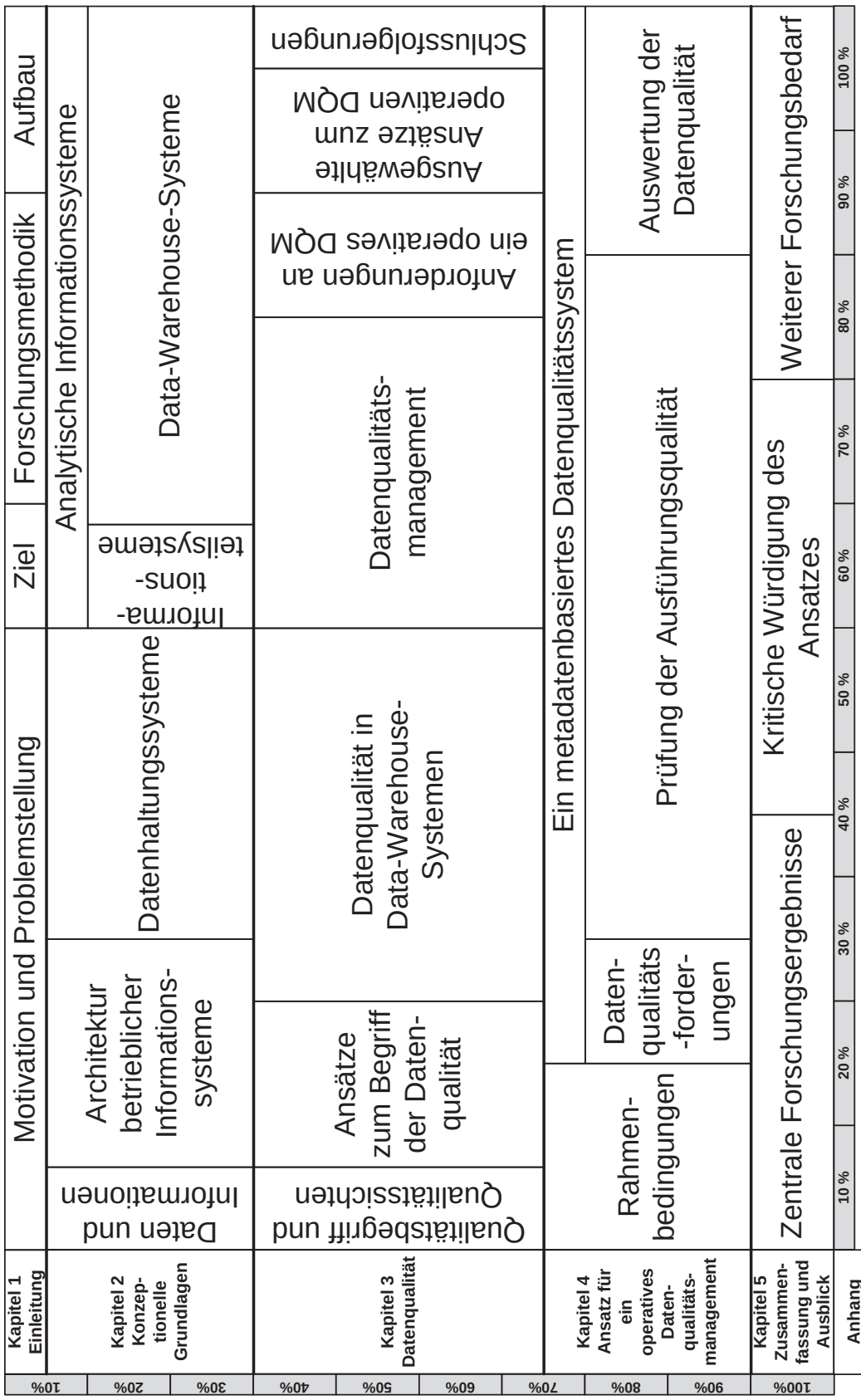


Abbildung 1.1: Grundsätzlicher Aufbau der Arbeit (Eigene Darstellung)

Kapitel 2

Konzeptionelle Grundlagen

Im folgenden Kapitel werden die wesentlichen konzeptionellen Grundlagen, die im Rahmen der Arbeit angewendet werden, dargestellt. Zunächst werden die Begriffe Daten und Information abgegrenzt und in Abschnitt 2.2 die Architektur betrieblicher Informationssysteme erläutert. Aufgrund der zentralen Bedeutung von Datenhaltungssystemen werden diese in Abschnitt 2.3 beschrieben. In Abschnitt 2.4 werden dann analytische Informationssysteme und insbesondere Data-Warehouse-Systeme untersucht. Aufbauend auf den zentralen Komponenten und der organisatorischen Gestaltung von Data-Warehouse-Systemen werden dann zusammenfassend Betrachtungsebenen eines Data-Warehouse-Systems dargestellt.

2.1 Daten und Information

Es ist weder in der Betriebswirtschaftslehre noch in der Informatik oder der Wirtschaftsinformatik bislang gelungen, die häufig verwendeten und grundlegenden Begriffe „Wissen“, „Information“ und „Daten“ klar abzugrenzen und zu definieren.[20] Bei der Verwendung dieser Begriffe bestehen Unklarheiten, Unsicherheiten und Missverständnisse. Es existieren Mehrdeutigkeiten, inkonsistente Verwendungen sowie unscharfe Begriffsdefinitionen. Daher soll im folgenden eine für die Arbeit geeignete Begriffsabgrenzung vorgenommen werden.

Im wesentlichen lassen sich die unterschiedlichen Begriffsdefinitionen durch die Ausprägungen der Dimensionen Semiotik, Träger der Information, Neuheitsgrad,

[20] Vgl. Maier und Lehner (1995), S. 165; Bode (1997), S. 451.

Wahrheitsgehalt und Zeitbezogenheit einordnen.[21] Die in der Betriebswirtschaftslehre häufig übernommene Begriffsdefinition von Information als zweckorientiertes Wissen, basiert auf dem Ansatz von WITTMANN.[22] Der Zweck orientiert sich dabei an der Vorbereitung von Handlungen und insbesondere von Entscheidungen, welche unmittelbar den subjektbezogenen Charakter dieses Informationsbegriffs verdeutlichten. Definitionen, die diesem Begriffsverständnis folgen, ordnen die Begriffe Daten, Information und Wissen häufig anhand der semiotischen Ebenen Pragmatik, Semantik und Syntaktik in ein hierarchisches System ein.[23] Wissen stellt den übergeordneten Rahmen dar, innerhalb dessen Informationen generiert und verwendet werden. Auf der pragmatischen Ebene wird zwischen Wissen *ohne* Zweckbezug und Informationen *mit* Zweckbezug unterschieden. Die Begriffe Nachrichten und Daten werden meist der semantischen Ebene zugeordnet, welche die Beziehungen zwischen Zeichen und ihrer Bedeutung behandelt. Die syntaktische Ebene beschäftigt sich mit den Darstellungsformen von Wissen, den Signalen, Zeichen und Symbolen.

Obwohl sich diese (oder ähnliche) Ansätze für zahlreiche Fragestellungen als zweckmässig erwiesen haben, sind sie in der Literatur nicht kritiklos.[24] Bei den Definitionen wird häufig zwischen Daten und Wissen anhand des konstitutiven Kriteriums der *Maschinenverarbeitbarkeit* eine Abgrenzung vorgenommen.[25] Problematisch ist insbesondere die Dynamik der technischen Möglichkeiten, die diese Abgrenzung fortlaufend verschieben. So hat beispielsweise die technische Entwicklung in der Vergangenheit dazu geführt, dass mittlerweile nahezu jegliches Wissen in irgendeiner Form maschinell zumindest erfassbar wird. Es ist zwischen der prinzipiellen und der tatsächlichen, im konkreten Einzelfall vorhandenen Maschinenverarbeitbarkeit zu unterscheiden. Untersuchungsgegenstand der Arbeit sind Data-Warehouse-Systeme, die im Kern aus Datenhaltungssystemen bestehen.[26] Daher erscheint es sinnvoll, eine Begriffsabgrenzung anhand der Da-

[21] Vgl. Bode (1997), S. 451; auf eine Darstellung der zahlreichen Definitionen soll hier verzichtet werden; vgl. z. B. Maier und Lehner (1995), S. 251ff.; Streubel (1996), S. 21.

[22] Vgl. Wittmann (1959), S. 14.

[23] Vgl. Streubel (1996), S. 22; Maier und Lehner (1995), S. 231; Picot und Reichwald (1991), S. 251f.; Rüttler (1991), S. 28; Augustin (1990), S. 16.

[24] Vgl. Bode (1997), S. 455-458.

[25] Vgl. Streubel (1996), S. 23.

[26] Vgl. Abschnitt 2.4.2.

tenhaltungssysteme vorzunehmen. Die Eigenschaft der Maschinenverarbeitbarkeit soll sich daher auf die derzeitige Möglichkeit einer, unter ökonomischen Gesichtspunkten realisierbaren, Verarbeitung in einem verfügbaren Datenhaltungssystem beziehen.

Eine Differenzierung zwischen Informationen und Daten anhand der *Zweckorientierung* ist ebenfalls problematisch.[27] In Data-Warehouse-Systemen, die auf die Erfüllung zukünftiger Informationsbedarfe ausgerichtet sind, finden sich neben zweckbezogenen Informationen auch Wissensbestandteile ohne direkten Zweckbezug, die auf Vorrat und unabhängig vom Endanwender produziert werden. Eine begriffliche Trennung wäre angebracht, wenn eine sachliche Unvereinbarkeit dies erfordern oder ein wesentlicher Nutzen hieraus entstehen würde. Da dies im Rahmen der Arbeit nicht ersichtlich ist, soll auch eine begriffliche Trennung anhand der Zweckorientierung nicht vorgenommen werden.

Ausgehend von Wissen als übergeordnetem Rahmen, in den sich die Begriffe Informationen und Daten einordnen lassen, ist eine möglichst genaue und geeignete Begriffsbestimmung notwendig. Wenngleich eine objektive und stabile Begriffsbestimmung hier nicht möglich erscheint, sollen dennoch die zentralen Begriffe beschrieben werden. Im Rahmen der Arbeit stellt *Wissen* „jede Form der Repräsentation von Teilen der realen oder gedachten (d. h. vorgestellten) Welt in einem materiellen Trägermedium“[28] dar. Kennzeichnend für diesen Wissensbegriff ist die Repräsentation von Ausschnitten der realen Welt. Wissen als Abbildung ist mit den Realitätsausschnitten nicht identisch, steht jedoch stets in Relation zu diesen und besitzt somit eine Bedeutung (Semantik). Aufbauend auf diesem Begriffsverständnis kann Information als echte Teilmenge von Wissen definiert werden. „*Informationen* sind Wissensbestandteile, die in Form menschlicher Sprache repräsentiert sind.“[29] Menschliche Sprache wird dahingehend eingeschränkt, dass diese die Übermittlung von Wissen zwischen Menschen erlaubt. Berücksichtigt werden so auch alle vom Menschen verfassten Sprachen, wie beispielsweise Computersprachen. Zur Unterscheidung zwischen maschinenverarbeitbaren und nicht maschinenverarbeitbaren Informationen soll der Begriff Daten als echte Teil-

[27] Vgl. hierzu Bode (1997), S. 455.
[28] Bode (1997), S. 458.
[29] Bode (1997), S. 459.

menge von Informationen definiert werden. *Daten* stellen maschinenverarbeitbare Informationen dar.[30]

Wenngleich die Zweckmässigkeit dieser Begriffsabgrenzung für andere Fragestellungen zu prüfen ist, erscheint sie für die Beschreibung und Einordnung der Begriffe im Rahmen der Arbeit ausreichend. Die hier angesprochenen Begriffe von Wissen, Information und Daten bilden eine begriffliche Grundlage der weiteren Arbeit und dienen so als Basis des in Abschnitt 3.2 erläuterten Begriffs der Datenqualität.

2.2 Architektur betrieblicher Informationssysteme

Im folgenden Abschnitt sollen betriebliche Systeme strukturiert und insbesondere das Teilsystem der Informationssysteme durch Informationssystemarchitekturen näher beschrieben werden. Basis dieser Ausführungen bilden insbesondere die in Abschnitt 2.2.2 dargestellte Systemtheorie und die in Abschnitt 2.2.3 beschriebene Abbildung von Systemen durch Modelle.

2.2.1 Betriebliche Informationssysteme

Untersucht man betriebliche Systeme, so können diese grundsätzlich in ein Basissystem und ein Informationssystem unterteilt werden.[31] Das betriebliche Basissystem realisiert die Sachziele der Unternehmung, indem es Einsatzgüter aus der Umwelt bezieht und diese in einem Leistungserstellungsprozess in Produkte transformiert.[32] Die hierfür notwendige Planung, Steuerung und Kontrolle des Basissystems wird durch ein Informationssystem wahrgenommen. Beide Teilsysteme des betrieblichen Systems sind durch Informationsbeziehungen miteinander verbunden.

Betriebliche Systeme können anhand verschiedener Abgrenzungskriterien in Teilsysteme gegliedert werden. So unterteilen beispielsweise FERSTL und SINZ diese anhand des Objektprinzips, des Phasenprinzips und der Aufgabenträger in die

[30] Vgl. hierzu auch Bode (1997), S. 460.
[31] Vgl. Ferstl und Sinz (2001), S. 28-30; Busch (1983), S. 34; Grochla (1975), S. 12f.
[32] Vgl. Ferstl und Sinz (2001), S. 28f.

Aufgabenobjekt	**Aufgabenträger**		**Aufgabenphase**
	automatisiert	nicht automatisiert	
Informations-system (Objektart Information)	*Anwendungs-systeme*	Sachbearbeiter, Datenerfasser, Manager	Lenkungssystem (Planung, Steuerung, Kontrolle)
	Anwendungs-systeme	Sachbearbeiter, Datenerfasser	Leistungssystem (Durchführung)
Basissystem (Objektart Nicht-Information)	Bearbeitungs-, Transportsysteme, ...	Werker	

Tabelle 2.1: Abgrenzung von Teilsystemen des betrieblichen Systems (In Anlehnung an Ferstl und Sinz (2001), S. 4)

in Abbildung 2.1 dargestellten Teilsysteme.[33] Nach dem Objektprinzip werden anhand der Beziehungsart das Informationssystem und das Basissystem unterschieden. Als Beziehungsarten wird dabei zwischen Informationen und Nicht-Informationen unterschieden. Das Phasenprinzip untergliedert das System in Leistungs- und Lenkungssystem. Die Unterscheidung nach maschinellen und personellen Aufgabenträgern führt zur Differenzierung zwischen einem automatisierten und einem nicht automatisierten Teilsystem.

Wenngleich in der Wirtschaftsinformatik der Begriff des Informationssystems nicht einheitlich verwendet wird,[34] bilden die betrieblichen Informationssysteme (IS)[35] den Erkenntnisgegenstand der Wirtschaftsinformatik.[36] Es sind soziotechnische Systeme, die menschliche oder maschinelle Komponenten als Aufgabenträger umfassen, die voneinander abhängig sind, ineinandergreifen und / oder zusammenwirken.[37] Grundsätzlich lassen sich die auszuführenden Aufgaben und deren Aufgabenträger als eine Beschreibungsdimensionen von Informationssyste-

[33] Vgl. Ferstl und Sinz (2001), S. 4f.

[34] Vgl. Mertens und Holzner (1992), S. 21f.; Ferstl und Sinz (2001), S. 8f.

[35] Genauer als Informations- und Kommunikationssysteme bezeichnet.

[36] Vgl. Wissenschaftliche Kommission der Wirtschaftsinformatik (1994), S. 81; Ferstl und Sinz (2001), S. 1.

[37] Vgl. Wissenschaftliche Kommission der Wirtschaftsinformatik (1994), S. 81.

men unterscheiden. Die Aufgabenebene besteht aus Informationsverarbeitungsaufgaben, die durch den Austausch von Informationen in Beziehung stehen.[38] Menschen und maschinelle Systeme, wie beispielsweise Rechner- und maschinelle Kommunikationssysteme, kommunizieren und führen so die Informationsverarbeitungsaufgaben durch.[39] Der Teil des Informationssystems, der die maschinellen Aufgabenträger umfasst, wird als Anwendungssystem bezeichnet[40] und stellt aus Sicht der Wirtschaftsinformatik die zentrale Komponente von Informationssystemen dar.[41] Anwendungssysteme umfassen den datenverarbeitungstechnisch realisierten, automatisierbaren Teil eines Informationssystems. Informationssysteme, die auch nicht automatisierte Aufgabenträger berücksichtigen, sind umfassender als Anwendungssysteme.

2.2.2 Systemtheoretische Grundlagen

Zur Analyse und Beschreibung des Aufbaus und der Funktionsweise von Informationssystemen bedarf es eines allgemein verwendbaren methodischen Rahmens, welcher beispielsweise durch die allgemeine Systemtheorie zur Verfügung gestellt wird. Im folgenden sollen daher die für die Arbeit wesentlichen Grundlagen der allgemeinen Systemtheorie dargestellt werden.[42]

Ein System[43] bedeutet die Vorstellung einer gegliederten Ganzheit und ermöglicht als Denkweise interessierende Zusammenhänge zwischen irgendwelchen Objekten in einer grösseren Gesamtheit zu erfassen.[44] Ein System entspricht so einer Sicht eines Betrachters auf einen bestimmten und explizit von einer Umgebung abgegrenzten Gegenstand.[45] Wesentlich ist die Feststellung, dass ein System nicht

[38] Vgl. Ferstl und Sinz (2001), S. 2.

[39] Vgl. Ferstl und Sinz (2001), S. 2f.

[40] Vgl. Ferstl und Sinz (2001), S. 5; vgl. auch Anwendungsmodelle in Barkow, Hesse, Kittlaus, Scheschonk und von Braun (1997), S. 45.

[41] Vgl. Mertens, Bodendorf, König, Picot und Schumann (2000), S. 1.

[42] Siehe hierzu z. B. Ferstl und Sinz (2001), S. 11-18 oder Guntram (1985), S. 296ff.; vgl. auch Seiffert (1994), S. 329f.; Jantsch (1994), S. 331-338; Seiffert (1992), S. 95-97; Ulrich (1970), S. 102-104.

[43] Der Begriff System entstammt aus dem griechischen Wort „$\tau\grave{o}\ \sigma\acute{\upsilon}\sigma\tau\eta\mu\alpha$“ Zusammenstellung, Vereinigung, Ganzes, welches die Existenz von Teilen oder Elementen voraussetzt; vgl. Seiffert (1994), S. 329f.; Vetter (1998), S. 109; Ulrich (1970), S. 105.

[44] Vgl. Ulrich (1970), S. 107.

[45] Vgl. Holten (1999), S. 126.

per se existiert, sondern aufgrund der gedanklichen Leistung eines Betrachters entsteht. Es werden gedankliche und natürliche Systeme unterschieden.[46] Als Betrachtungsgegenstand der Informatik und Wirtschaftsinformatik sind insbesondere die natürlichen Systeme in Form von Informationen von hoher Bedeutung.[47]

Als formale Wissenschaft ist der Begriff der allgemeinen Systemtheorie rein formal.[48] Er enthält keine Aussagen über die Art der Elemente, deren Eigenschaften und Beziehungen innerhalb und zwischen Systemen sowie dem Systemzweck und den Systemgrenzen.[49] Aufbauend auf dem Mengenbegriff und in Anlehnung an HABERFELLNER kann ein System als eine Menge von Elementen, zwischen denen irgendwelche Beziehungen bestehen, definiert werden.[50] Ein allgemeines System S^G wird daher formal als Teilmenge der kartesischen Produkte über den Mengen V_i als

$$S^G \subseteq \times_{i \in I} V_i$$

definiert, wobei $I \neq \emptyset$ eine beliebige Indexmenge und $V = \{V_i : i \in I\}$ eine Menge von nicht-leeren Mengen ist.[51] Die in S^G auftretenden Mengen V_i werden als Systemelemente oder -komponenten bezeichnet. Die Menge der in S^G enthaltenen Tupel von Elementen aus V_i definiert das Verhalten des Systems. Die Menge

$$R^G \in \{(V_i, V_j) : i, j \in I \wedge i \neq j\}$$

heisst Struktur des Systems und beschreibt diese durch paarweise Beziehungen zwischen den Systemelementen.[52] Neben diesen formalen Systemeigenschaften sind Systeme durch weitere Eigenschaften gekennzeichnet.[53] Hierzu zählen beispielsweise:

- der Systemzweck und die Zielorientierung von Systemen (zweckorientiert; zielorientiert).

[46] Vgl. Schütte (1998), S. 37.
[47] Vgl. Schütte (1998), S. 37.
[48] Vgl. Ulrich (1970), S. 105.
[49] Vgl. Holten (1999), S. 126f.; Haberfellner (1975), S. 6.
[50] Vgl. Haberfellner (1975), S. 6; Ferstl und Sinz (2001), S. 12-16; Ulrich (1970), S. 105.
[51] Vgl. hierzu Ferstl und Sinz (2001), S. 12-16.
[52] Vgl. Ferstl und Sinz (2001), S. 12.
[53] Vgl. Ulrich (1970), S. 107-118.

- das Systemverhalten (statisch; dynamisch).
- die Vorhersagbarkeit (determiniert; stochastisch).
- die Offenheit von Systemen (offen; geschlossen).
- die Komplexität von Systemen (einfach; komplex).

Die hier im Überblick dargestellten Grundlagen der Systemtheorie bilden den methodischen Rahmen zur Analyse und Beschreibung von Informationssystemen und insbesondere von Data-Warehouse-Systemen. Im folgenden Abschnitt soll die Systemabbildung anhand von Modellen dargestellt werden.

2.2.3 Systemabbildungen durch Modelle

Modelle spielen in der Wirtschaftsinformatik bei der Untersuchung von Informationssystemen eine zentrale Rolle.[54] Zur Erfassung der Problemsituation bildet sich ein Modellersteller ein Wahrnehmungsmodell der ihn umgebenden Realität. Er strukturiert hierbei den Untersuchungsbereich in Objekte und Beziehungen zwischen diesen. Dabei verwendet er eine Modellierungs- oder Modellbeschreibungssprache. Diese legt die zur Verfügung stehenden Elemente sowie deren Semantik und bestimmte Konstruktionsregeln fest und ist in einem Metamodell hinterlegt.[55]

Ausgehend von der wissenschaftstheoretischen Grundposition kann die Modellbildung grundsätzlich in zwei Typen, einerseits das abbildungstheoretische und andererseits das konstruktivistische Modellverständnis, unterteilt werden.[56] Bislang dominiert in der Wirtschaftsinformatik das abbildungstheoretische Modellverständnis, das auf einer realistischen wissenschaftstheoretischen Grundposition basiert.[57] Im allgemeinen wird davon ausgegangen, dass ein Diskursbereich und

[54] Vgl. Heine (1999), S. 39; Ferstl und Sinz (2001), S. 119; Schütte (1998), S. 63.

[55] Vgl. Rauh und Stickel (1997), S. 14; in der Wirtschaftsinformatik werden üblicherweise formale (künstliche) Sprachen verwendet, deren Syntaxbeschreibung i. d. R. als Metamodell bezeichnet wird; vgl. Schütte (1998), S. 68 oder Strahringer (1996).

[56] Vgl. im folgenden Schütte (1998), S. 46ff.; eine ausführliche Diskussion des philosophischen Streits um den „richtigen“ Modellbegriff ist hier nicht beabsichtigt. Für eine Diskussion des abbildungs- und konstruktionsorientierten Modellbegriffs und zum Vergleich verschiedener Modellbegriffe vgl. z. B. Schütte (1998), S. 40ff.

[57] Vgl. Schütte (1998), S. 52ff.

ein Objektsystem S_O existieren. Eine Abbildungsfunktion f überführt dann das Objektsystem in ein Modellsystem S_M.[58] Die Abbildung 2.1 veranschaulicht diese Zusammenhänge.

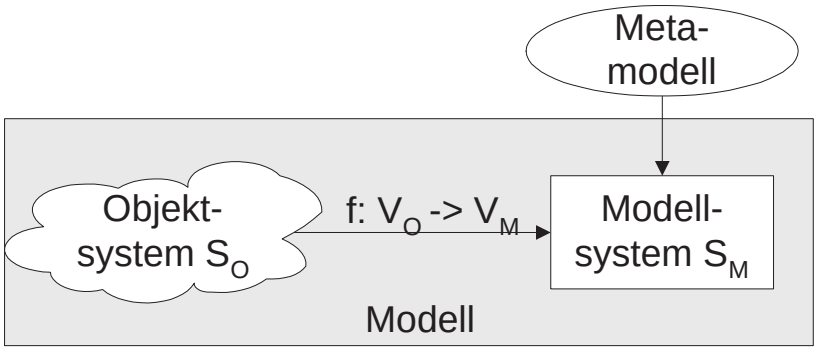


Abbildung 2.1: Abbildungstheoretischer Modellbegriff (Vgl. Ferstl und Sinz (2001), S. 121)

Das konstruktivistische Modellverständnis geht, im Gegensatz zum abbildungstheoretischen Modellverständnis, nicht von Strukturen in der Realität aus, die unabhängig vom Erkenntnisvermögen des Betrachters bestehen.[59] Die Wahrnehmung der Realität wird im Konstruktivismus als Interpretation verstanden. Im konstruktivistischen Modellverständnis wird ein Modell verstanden, als „das Ergebnis einer *Konstruktion eines Modellierers*, der für *Modellnutzer* eine Repräsentation eines Originals zu einer *Zeit* als relevant mit Hilfe einer *Sprache* deklariert“[60] (Vgl. Abbildung 2.2).

Interessant am konstruktivistischen Modellverständnis ist die Erkenntnisauffassung, die den Realitätsbezug immer auf den Erkenntnisinhalt eines Subjektes auf ein reales Objekt bezieht. Ausgangspunkt der Modellierung ist ein subjektives Problemempfinden.[61] Beim konstruktivistischen Modellverständnis wird keine Homomorphie zwischen Modell und Realwelt gefordert bzw. kann nicht gefordert werden. Modelle sind vielmehr subjektiv begründet und folglich wird ein

[58] Vgl. Ferstl und Sinz (2001), S. 120f.
[59] Vgl. Schütte (1998), S. 56ff.
[60] Schütte (1998), S. 59 (Hervorhebungen im Original).
[61] Vgl. Schütte (1998), S. 61.

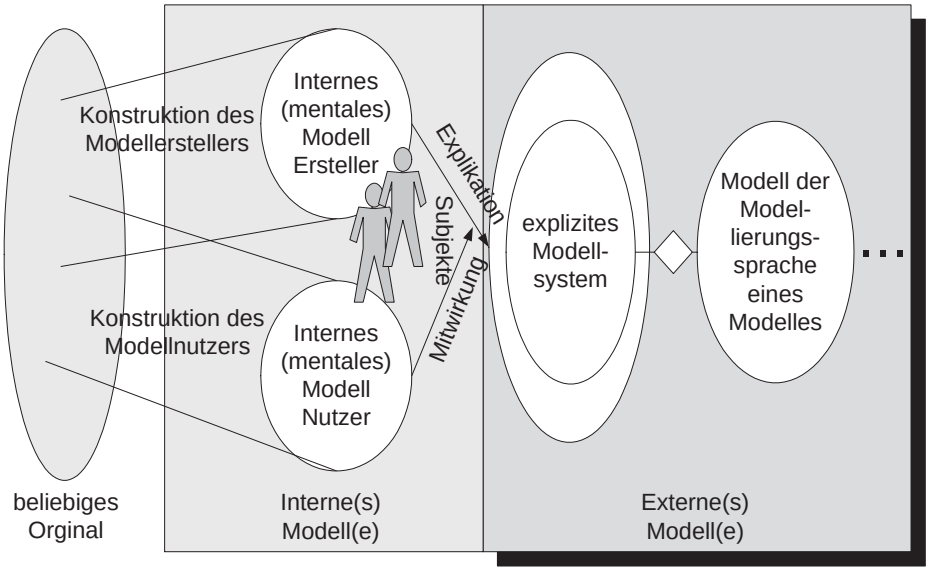


Abbildung 2.2: Konstruktivistischer Modellbegriff (In Anlehnung an Schütte (1998), S. 61)

objektiver Vergleich mit der Realität abgelehnt.[62] Eine Bewertung der Modelle erfolgt durch Beteiligung der an der Modellierung beteiligten Subjekte (Modellhersteller sowie Modellnutzer). Sie entscheiden schlussendlich über die Eignung eines Modells und somit dessen Qualität.[63]

2.2.4 Informationssystemarchitekturen

Informationssysteme können durch Modelle beschrieben werden. Für diese wird im Allgemeinen der Begriff der Informationssystemarchitektur[64] verwendet.[65] Informationssystemarchitekturen umfassen dabei sowohl die eigentliche Beschreibung im Modell als auch das ihnen zugehörige Metamodell. Zur Komplexitätsreduktion können Modelle zunächst in Modellebenen und zugehörige (Modell-) Sichten untergliedert werden.[66] Der in Abbildung 2.3 aufgezeigte generische Architekturrahmen beschreibt den Zusammenhang zwischen Modellebenen, Modell-

[62] Vgl. Schütte (1998), S. 62.
[63] Vgl. Schelp (2000), S. 19f.
[64] Auch kurz als IS-Architektur oder Architektur bezeichnet.
[65] Vgl. Sinz (1997), S. 876; Heinrich (1992), S. 76; Vetter (1994), S. 165.
[66] Vgl. Sinz (1997), S. 877.

sichten, dem Metamodell sowie den zugehörigen Konstruktionsregeln. Eine Modellebene $i = \{1, \ldots, n\}$ umfasst für eine bestimmte Zielsetzung eine vollständige Beschreibung des zu modellierenden Systems und basiert dabei auf einem Metamodell MM_i. Werden entsprechende Projektionen $\Pi_S(MM_i)$ auf das Metamodell einer Modellebene definiert, wird von einer Sicht S auf die Modellebene i gesprochen. Beziehungen zwischen den Modellebenen werden durch Beziehungen zwischen den Metamodellen festgelegt. Hierzu wird für jede paarweise Beziehung zwischen den Modellebenen i und j ein entsprechendes Beziehungs-Metamodell BMM_i definiert, das Elemente der Modellebene i mit Elementen der Modellebene j verknüpft.

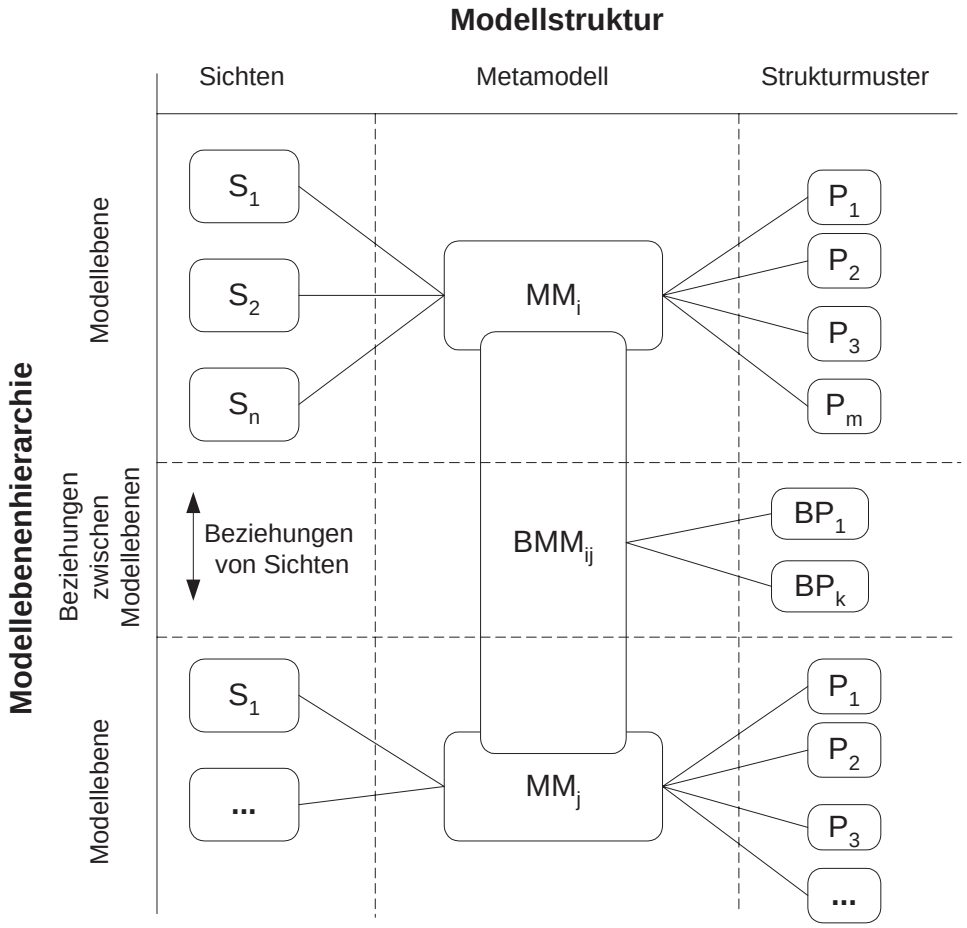


Abbildung 2.3: Generischer Architekturrahmen für Informationssysteme (Vgl. Sinz (1997), S. 876)

Zur Beschreibung von Informationssystemarchitekturen existieren zahlreiche An-

sätze,[67] welche im folgenden exemplarisch aufgeführt und anschliessend strukturiert werden sollen. Der von ZACHMAN entwickelte Gestaltungsrahmen zur Beschreibung von Informationssystemarchitekturen ist durch fünf Modellebenen und drei Sichten gekennzeichnet.[68] Die Modellebenen werden durch die jeweiligen Zielsetzungen der bei der Informationssystementwicklung beteiligten Personengruppen gebildet. Daten, Funktionen und das Verarbeitungsnetzwerk bilden die jeweiligen Sichten auf diese Ebenen. ÖSTERLE konzipiert eine Methode zur Prozessentwicklung und prozessorientierten Einführung von Standardsoftware,[69] welche als (geschäfts-)prozessorientiertes Architekturkonzept für Informationssysteme einzustufen ist.[70] Die Geschäftsstrategie, die Prozesse und das Informationssystem[71] bilden die jeweiligen Modellebenen. Hierauf lassen sich Sichten der Organisation, der Daten sowie der Funktionen bilden. Das Modellierungskonzept des semantischen Objektmodells unterscheidet drei Modellierungsebenen.[72] Der Unternehmensplan repräsentiert die Aussensicht eines betrieblichen Systems. Die Innensicht bezieht sich auf die Geschäftsprozessmodelle, die durch Inanspruchnahme von Ressourcen durchgeführt werden. Auf der Ressourcenebene wird zwischen Spezifikationen der Aufbauorganisation, der Anwendungssysteme sowie der Maschinen und Anlagen unterteilt. Als Sichten werden die Modellebenen in ein strukturorientiertes und ein verhaltensorientiertes Teilmodell gegliedert. Im Rahmen der Architektur integrierter Informationssysteme (ARIS) wird von SCHEER die Informationssystemarchitektur primär in Sichten und sekundär in Modellebenen unterteilt.[73] Als Sichten werden Daten, Funktionen und die Organisation genannt. Zeitliche Aspekte werden durch eine vierte Sicht, die Steuerungssicht, berücksichtigt. Die Steuerungsicht verdeutlicht das Zusammenwirken der anderen Sichten. In Abhängigkeit der Nähe zur Informationstechnik wird weiter in die Modellebenen des Fachkonzepts, des DV-Konzepts und der Implementierung unterschieden.

[67] Vgl. Sinz (1997), S. 878.
[68] Vgl. Zachman (1987), S. 462-469.
[69] Vgl. Österle (1995).
[70] Vgl. Sinz (1997), S. 880f.
[71] Im Sinne dieser Arbeit ist das Informationssystem als Anwendungssystem aufzufassen.
[72] Vgl. Sinz (1997), S. 881f.; Ferstl und Sinz (2001), S. 180ff.
[73] Vgl. Scheer (1998), S. 10ff.

Merkmal	Merkmalsausprägung			
Beschreibungssicht	Stuktursicht		Verhaltenssicht	
Abstraktionsebene	Ausprägungsebene	Typebene	Metaebene	Meta-Metaebene
Beschreibungsebene	Fachkonzept	DV-Konzept	Implementierung	
Konkretisierungsgrad	abstrakt		ausformuliert	
Geltungsanspruch	Istmodell	Sollmodell	Idealmodell	
Inhaltliche Individualität	(unternehmens-) spezifisches Modell	domänenabhängiges Referenzmodell	domänenunabhänges Referenzmodell	

Abbildung 2.4: Morphologischer Kasten der Informationssystemmodellierung (Vgl. Schwegmann (1999), S. 9)

Zwecks einer Systematisierung können die verschiedenen Informationssystemmodelle anhand verschiedener Merkmale differenziert werden. Diese sind in Abbildung 2.4 in Form eines morphologischen Kastens zusammengestellt.[74] Systeme können sowohl aus einer struktur- und einer verhaltensorientierten Sicht betrachtet werden. In der Struktursicht werden Systemelemente und Beziehungen zwischen diesen beschrieben. Die Verhaltenssicht fokussiert auf die Interaktion der in der Struktursicht beschriebenen Systemelemente. ROSEMANN sieht eine an ARIS orientierte Sichtenbildung vor und unterscheidet zwischen Daten-, Funktions-, Organisations- und Prozesssicht.[75]

Bei der Modellbildung lassen sich verschiedene sprachliche Abstraktionsebenen differenzieren. Es wird zwischen der Ausprägungsebene, der Typebene, der Metaebene und der Meta-Metaebene etc. unterschieden. Auf der Ausprägungsebene werden konkrete Gegenstände, Personen usw. sowie deren Struktur und Verhalten

[74] Vgl. zu den folgenden Ausführungen Schwegmann (1999), S. 9-12 sowie Rosemann (1996), S. 22-38.

[75] Vgl. Rosemann (1996), S. 23.

beschrieben. Indem für gleichartige Elemente Klassen gebildet werden, kann von den Elementen und Beziehungen der Ausprägungsebene abstrahiert und generalisierte Aussagen über diese gemacht werden. Durch Klassifikation gleichartiger Elemente der Typebene erfolgt eine Abstraktion von der Typ- zur Metaebene. Diese beschreibt die Klassen und Beziehungen der Typebene.

Von der Abstraktionsebene ist die Beschreibungsebene von Modellen zu unterscheiden. Die Beschreibungsebene bezeichnet die Nähe eines Modells zur Problemdomäne bzw. zur datenverarbeitungstechnischen Implementierung. Bei der Erstellung eines Informationssystemmodells wird allgemein zwischen Modellen auf der Ebene des Fachkonzeptes, des DV-Konzeptes und der Implementierung unterschieden. Fachkonzeptuelle Modelle beinhalten die Anforderungen an ein Informationssystemmodell und werden zusammen mit dem Fachanwender erstellt. Sie sind unabhängig von den technischen und organisatorischen Rahmenbedingungen einer zukünftigen Implementierung. Modelle auf DV-Konzeptebene enthalten den für die Systementwicklung relevanten Teil der fachkonzeptuellen Modelle und beschreiben das „Wie“. Implementierungsmodelle sind die detaillierteste Stufe eines Informationssystemmodells. Implementierungsmodelle für die Anwendungssystementwicklung werden in Programmiersprachen ausgedrückt und können durch Anwendungssysteme ausgeführt werden. Implementierungsmodelle für die Organisationsgestaltung beinhalten detaillierte organisatorische Gestaltungsanweisungen, beispielsweise in Form von Organigrammen und Stellenbeschreibungen.

Nach dem Geltungsanspruch kann zwischen Ist-, Soll- und Idealmodellen unterschieden werden.[76] Istmodelle dokumentieren den aktuellen Zustand der betrachteten Problemdomäne. Idealmodelle beschreiben die optimale Gestaltung einer Problemdomäne, ohne vorhandene Restriktionen zu berücksichtigen. Ein Sollmodell kann als die realisierungsfähige Version eines Idealmodells angesehen werden und beschreibt den Zustand, der in einem bestimmten Zeitraum erreicht werden kann.

Die inhaltliche Individualität charakterisiert den Umfang des Adressatenkreises, für den ein Modell von Bedeutung ist. Zur Abstufung der inhaltlichen Individua-

[76] Vgl. Rosemann und Rotthowe (1995), S. 14.

lität von Modellen werden domänenunabhängige und domänenabhängige Referenzmodelle sowie (unternehmens-)spezifische Modelle differenziert.

2.3 Datenhaltungssysteme

Ein modernes Datenbanksystem besteht aus einer Datenbasis, einem Datenbankverwaltungssystem und den Datenbankkommunikationsschnittstellen.[77] Dieser grundsätzliche Aufbau bezweckt insbesondere die Integration von isoliert gehaltenen Datenbeständen in ein Datenhaltungssystem, was Vorteile in bezug auf die Reduktion redundanter und inkonsistenter Daten, den Zugriff mehrerer Anwender, die Zusammenfassung einzelner Datenbankoperationen zu Transaktionen, die Förderung der Datensicherheit sowie die Anwendung von Standards hat.[78] Eines der zentralen Ziele der Nutzung eines Datenbanksystems besteht darin, die enge Verflechtung und Abhängigkeit zwischen Daten und den auf ihnen operierenden Anwendungssystemen zu entkoppeln. Im wesentlichen wird dies durch Trennung von Daten und den auf sie zugreifenden Programmen erreicht.

Allgemein unterscheidet man in Datenbanksystemen drei Abstraktionsebenen:[79]

- Die physische Ebene, auf der die physische Speicherung der Daten festgelegt wird.
- Die logische Ebene, die anhand von Datenbankschemata die gespeicherten Daten beschreibt.
- Die Sichten, die für jeweilige Anwendungsbereiche aus der gesamten Datenmenge nutzerabhängige Teilmengen spezifiziert.

Dabei können zwei Stufen der Datenunabhängigkeit in Datenbanksystemen betrachtet werden.[80] Die erste Stufe, die physische Datenunabhängigkeit, bezieht sich auf die Unabhängigkeit zwischen physischer Speicherstruktur und logischer

[77] Vgl. z. B. Vossen (2000), S. 9f.; Wedekind (2001), S. 139f.; Rauh und Stickel (1997), S. 19; Kemper und Eickler (1996), S. 25ff.; Gabriel und Röhrs (1995), S. 189ff.

[78] Vgl. Date (2000), S. 16-19.

[79] Vgl. Kemper und Eickler (1996), S. 17f.

[80] Vgl. Date (2000), S. 19f.; Heuer und Saake (2000), S. 26f.; Kemper und Eickler (1996), S. 19.

Ebene der Datenmodellierung. Hierdurch wird beispielsweise der Einsatz eines effizienteren Suchverfahrens auf der physischen Ebene möglich, ohne die darauf zugreifenden Anwendungssysteme anpassen zu müssen. Die logische Datenunabhängigkeit, als zweite Stufe, bezieht sich auf die Unabhängigkeit zwischen der Datenbeschreibung in Datenbankschemata und den darauf zugreifenden Anwendungssystemen. Während die physische Datenunabhängigkeit zumeist von den heutigen Datenbanksystemen erfüllt wird, kann die logische Datenunabhängigkeit lediglich für beschränkte Modifikationen des Datenbankschemas gewährleistet werden.[81]

In Abschnitt 2.3.1 wird zunächst die Modellierung der Datenbasis als wichtige Komponente eines Datenhaltungssystems beschrieben. Im Anschluss daran wird in Abschnitt 2.3.2 die Abfragesprache SQL als Standardschnittstelle zur Kommunikation mit dem Datenbanksystem erläutert. Auf das eher technisch orientierte Datenbankverwaltungssystem, als dritte Komponente, soll hier nicht explizit eingegangen werden.

2.3.1 Datenmodelle

Ein Ziel der Verwendung von Datenbanksystemen ist die Trennung von Anwendungssystem und dem direkten, physischen Zugriff auf die Daten. Zur Kommunikation der in der Datenbank gehaltenen Daten und Datenstrukturen dienen Datenmodelle, die im allgemeinen Fall ein System von Konzepten zur Beschreibung relevanter Daten charakterisieren.[82]

Wenngleich in der Literatur nicht immer streng zwischen unterschiedlichen sprachlichen Abstraktionsebenen unterschieden wird,[83] sollen für den weiteren Verlauf der Arbeit die Begriffe Metadatenmodell, Datenmodellschema[84] und die

[81] Vgl. Kemper und Eickler (1996), S. 19.

[82] Vgl. Heuer und Saake (2000), S. 50.

[83] Vgl. z. B. Date (1992), S. 361; Winter (1998), S. 33.

[84] Ein Schema stellt die Beschreibung eines Datenbestandes dar; vgl. Zehnder (1998), S. 24; sofern es der Kontext ermöglicht, sollen auch die Bezeichnungen Datenmodell und Schema verwendet werden. Das Ergebnis der Modellierung kann auch im Sinne des Vorgangs der Spezifikation als „Spezifikationen“ bezeichnet sein; vgl. hierzu Winter (1998), S. 34. und Vossen (2000), S. 74-76; vgl. in diesem Zusammenhang auch die Qualitätssichten in Abschnitt 3.1.

Instanz bzw. Ausprägung eines Datenmodellschemas im folgenden Sinn unterschieden werden. Das Metadatenmodell beschreibt als Beschreibungssprache die zur Modellierung eines Datenmodellschemas zur Verfügung stehenden Elemente, deren Semantik sowie eventuelle Konstruktionsregeln. Die zentrale Betrachtungsebene stellt die Beschreibung eines konkreten Sachverhaltes (Realitätsausschnitt) in Form eines Datenmodellschemas auf Typebene dar. Die konkreten Einzeldaten finden sich in der Ausprägung bzw. Instanz eines Datenmodellschemas wieder. Die Instanz eines Datenmodellschemas wird als Datenbank(-zustand) bezeichnet.

Datenmodelle werden üblicherweise über deren Beschreibungsbestandteile beschrieben, wozu im allgemeinen ein Strukturteil, ein Operationenteil und Konsistenzregeln zählen.[85] Der Strukturteil beschreibt die Eigenschaften der erfassten Gegenstände und deren Beziehungen untereinander. Dynamische Eigenschaften werden über die möglichen Datenoperationen auf den Datenelementen beschrieben. Die Konsistenz des Modells[86] hinsichtlich Struktur und Operationen wird über Konsistenzbedingungen sichergestellt.

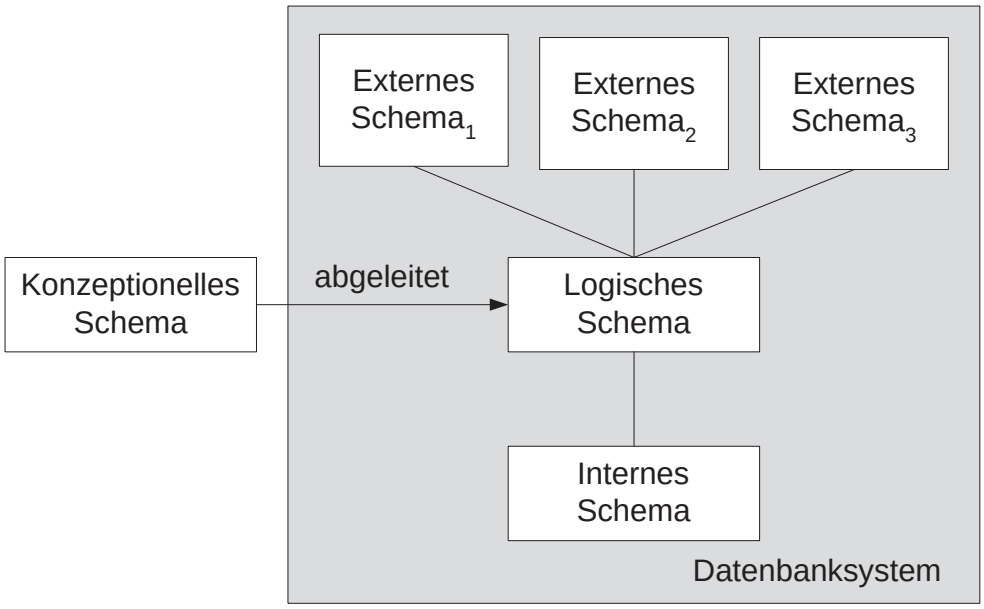


Abbildung 2.5: Schemata-Konzept (In Anlehnung an Zehnder (1998), S. 237)

[85] Vgl. Schelp (2000), S. 34; Winter (1998),S. 33.

[86] In diesem Zusammenhang bezeichnet Konsistenz (semantische Integrität, logische Korrektheit) einer Datenbank, die Eigenschaft, dass die in der Datenbank gespeicherten Inhalte die Semantik des zu modellierenden Realitätsausschnitts korrekt abbilden; vgl. Winter (1998), S. 36.

Anhand der unterschiedlichen Beschreibungsebenen werden verschiedene Datenmodelle unterschieden.[87] Meist wird, wie in Abbildung 2.5 skizziert, zwischen konzeptionellen, logischen, internen und externen Schemata differenziert.[88] Während sich die externen, logischen und internen Schemata auf ein Datenbanksystem beziehen, beschreibt das konzeptionelle Schema unabhängig von einem konkreten Datenbanksystem und den dadurch entstehenden Restriktionen den jeweilig zu modellierenden Realweltausschnitt.[89] Das konzeptionelle Schema betrachtet die über die konkrete Umsetzung in einem Datenbanksystem hinausgehenden Aspekte. Eines der bekanntesten Modelle in diesem Zusammenhang ist das Entity-Relationship-Modell (ER-Modell) von CHEN.[90] In diesem Modell werden Gegenstände zu Entitätstypen und Beziehungen zwischen den Entitäten zu Beziehungstypen abstrahiert. Weiterhin werden den Entitätstypen und Beziehungstypen Attribute zugeordnet. Das konzeptionelle Schema wird dann für ein konkret einzusetzendes Datenbanksystem in ein logisches Schema überführt, welches sich im allgemeinen aus dem konzeptionellen Schema nach festen Regeln generieren lässt.[91] Als Modelltypen auf der logischen Ebene sind das hierarchische, das Netzwerk und das relationale Datenmodell zu nennen. Das interne Schema beschreibt auf der physischen Ebene den Inhalt der Datenbasis sowie die notwendigen Datenbankfunktionen. Das externe Schema legt für bestimmte Benutzergruppen Sichten auf die Daten fest und beschreibt für bestimmte Benutzergruppen zugängliche Bereiche.

Aufgrund der zentralen Bedeutung des relationalen Datenmodells und dessen weiten Verbreitung soll dieses als Beispiel eines logischen Schemas im folgenden skizziert werden.[92] Es wurde von CODD 1970 eingeführt und besteht aus einem Struktur- und einem Regelteil.[93] Zur Modellierung von Objekten und Beziehungen steht im relationalen Datenmodell ein einziger Strukturtyp, der Relationentyp,

[87] Vgl. Zehnder (1998), S. 23.

[88] Vgl. Zehnder (1998), 279; vgl. auch ANSI/X3/SPARC Study Group on Data Base Management Systems (1975), S. 12ff.; Stock (2001), S. 6f.; Schelp (2000), S. 42-44; der Begriff "semantisches Modell" soll hier nicht verwendet werden.

[89] Vgl. auch Winter (1998), S. 33.

[90] Vgl. Chen (1976) und auch beispielsweise Kemper und Eickler (1996), S. 35ff.

[91] Beispielsweise lassen sich relationale Datenmodelle relativ einfach aus dem Entity-Relationship-Modell ableiten; vgl. hierzu u. a. Vossen (2000), S. 125ff.

[92] Vgl. hierzu beispielsweise Date (2000), S. 109ff.; Elmasri und Navathe (1994), S. 138ff.

[93] Erstmals eingeführt in Codd (1970) und weiter detailliert in Codd (1972a) und Codd (1972b).

zur Verfügung. Für dessen Ausprägung bietet sich die Vorstellung einer Datentabelle bestehend aus Spalten und Zeilen an.[94] Der Dateninhalt kann mathematisch durch das Konzept der Relation beschrieben werden.[95] Die Struktur der Relation wird im sogenannten Relationenschema festgehalten, indem es die zeitinvarianten Eigenschaftsbezeichnungen der im Modell erfassten Elemente, die sogenannten Attribute, enthält.[96] Deren Inhalt, die Attributwerte a_i mit $i = 1,\ldots,m$ einer Relation bestehen in Form einer Menge von Datentupeln $r = \{t_1, t_2, \ldots, t_n\}$ aus Ausprägungen der m Attribute.[97] Der Wertebereich eines Attributes $dom(A_i)$ legt die Menge der möglichen Ausprägungen fest.[98] Formal kann eine Relation r einer Folge von m Attributen $(A_1,\ldots,A_m)$ mit den zugehörigen Wertebereichen $dom(A_i),\ 1 \leq i \leq m$, als eine Teilmenge des kartesischen Produkts der Wertebereiche, d. h. $r \subseteq dom(A_1) \times \ldots \times dom(A_m)$, definiert werden.[99]

Im allgemeinen interessiert man sich allerdings nicht für alle möglichen Ausprägungen, sondern nur für solche, die gewissen, in der Aussenwelt zu beobachtenden semantischen Bedingungen genügen. Sie repräsentieren in diesem Sinne einen als gültig angesehenen „Ausschnitt der realen Welt“.[100] Allgemein bezeichnet man solche Beobachtungen semantischer Natur als Datenabhängigkeiten oder Konsistenzbedingungen.[101] Sie erlauben Aussagen darüber, welche Relationen aus einer Menge aller möglichen Relationen als „sinnvoll“ anzusehen sind. Die Integrität einer Datenbank bezeichnet inwieweit diese Bedingungen eingehalten werden.[102] Da diese Bedingungen aus der Semantik der Anwendung folgen, wird auch der Begriff „*semantische Integritätsbedingungen*“ verwendet.[103] In diesem

[94] Siehe hierzu beispielsweise Darstellungen in Heuer und Saake (2000), S. 107 oder Elmasri und Navathe (1994), S. 140.

[95] Vgl. Vossen (2000), S. 115-118; Date (1992), S. 41 ff.

[96] Vgl. Vossen (2000), S. 115; Heuer und Saake (2000), S. 107; Elmasri und Navathe (1994), S. 139.

[97] Ist die Ordnung der Attribute festgelegt, ist die Darstellung $< v_1, v_2, \ldots, v_m >$ für ein Tupel gebräuchlich; vgl. Elmasri und Navathe (1994), S. 139.

[98] Vgl. Vossen (2000), S. 115; Zehnder (1998), S. 70; Codd (1970), S. 379.

[99] Vgl. z. B. Elmasri und Navathe (1994), S. 139f. Wenngleich diese Definition unter formalen Gesichtspunkten problematisch sein kann (vgl. hierzu z. B. Heuer und Saake (2000), S. 107 und S. 109) ist sie im Rahmen der Arbeit ausreichend. Eine alternative Definition findet sich z. B. in Heuer und Saake (2000), S. 108f. oder Elmasri und Navathe (1994), S. 141.

[100] Vgl. Vossen (2000), S. 118.

[101] Vgl. Vossen (2000), S. 119; vgl. auch Winter (1998), S. 36.

[102] Vgl. Heuer und Saake (2000), S. 495; Date (2000), S. 249.

[103] Vgl. Heuer und Saake (2000), S. 495; Kemper und Eickler (1996), S. 119; Elmasri und Navathe (1994), S. 14f.; Kandzia und Klein (1993), S. 130; im Verlauf der Arbeit werden solche Bedingun-

Zusammenhang spricht man auch von einer konsistenten Datenbank, falls die auf dem Datenbankschema festgelegten Bedingungen erfüllt sind.[104] Eine Relation ist punktweise konsistent, falls die auf dem Relationenschema festgelegten Bedingungen erfüllt sind.[105] Die automatische Überprüfung von Integritätsbedingungen und deren Sicherung ist erst seit kurzem in kommerziellen relationalen Systemen enthalten und immer noch ein relativ junges Forschungsgebiet.[106]

Zunächst können Integritätsbedingungen anhand des zeitlichen Kontexts in drei Arten unterschieden werden:[107]

- Statische Bedingungen schränken die Menge der durch die Festlegung des betreffenden Schemas möglichen Datenbankzustände ein.
- Transitionale oder halb-dynamische Bedingungen schränken mögliche Zustandsübergänge ein. Während statische Bedingungen auf einen Zustand Bezug nehmen, betrachten transitionale Bedingungen ein Paar von aufeinanderfolgenden Zuständen.
- Als Verallgemeinerung der transitionalen Bedingungen, schränken dynamische Integritätsbedingungen mögliche Zustandsfolgen ein. Sie beziehen sich nicht auf zwei, sondern auf eine Folge von Datenbankzuständen, welche in zeitlicher Abfolge durchlaufen werden können.

Eine weitere Klassifikation der Integritätsbedingungen kann durch die Granularität beschrieben werden, indem nach der von der Bedingung betroffenen Dateneinheit unterschieden wird.[108] In relationalen Datenbanken können Integritätsbedingungen auf Ebene von Attributen, von Tupeln, einer Relation oder einer Datenbank festgelegt sein. Univariate Bedingungen beziehen sich auf ein Attribut, während multivariate Bedingungen mehrere Attribute berücksichtigen.[109] An-

gen sowohl als Konsistenzbedingung als auch Integritätsbedingung bezeichnet.

[104] Vgl. Vossen (2000), S. 125; Elmasri und Navathe (1994), S. 538.

[105] Vgl. Vossen (2000), S. 121.

[106] Vgl. Heuer und Saake (2000), S. 521; Kemper und Eickler (1996), S. 119.

[107] Vgl. hierzu Vossen (2000), S. 148f.; Heuer und Saake (2000), S. 496 u. S. 507; Kemper und Eickler (1996), S. 119.

[108] Vgl. Heuer und Saake (2000), S. 507f.

[109] Vgl. Milek, Reigrotzki, Bosch und Block (2001), S. 191.

Art	Beschreibung	Beispiele
Wertebereichs- und Attributbedingungen	Einschränkung eines Attributs	Ober- und Untergrenzen für Werte
		Menge möglicher Werte
		Pflichtfelder, bzw. Ausschluss der Verwendung von Nullwerten (Pflichtfelder)
Tupelbedingungen	Spezifikation von Zusammenhängen zwischen Attributen einzelner Tupel innerhalb einer Relation	
Relationenbedingungen	Menge aller Tupel einer Relation	Schlüsselbedingungen
		Aggregat-Bedingungen (z. B. Ober- und Untergrenzen für Summe der Guthaben)
		Rekursive-Bedingungen
Interrelationale Bedingungen	Zusammenhänge zwischen Relationen	Fremdschlüsselbeziehungen
		Aggregat-Bedingungen
		Rekursive-Bedingungen

Tabelle 2.2: Wichtige statische Integritätsbedingungen

hand der Anzahl betrachteter Datentupel kann zwischen Einzel- und Multitupelbedingungen unterschieden werden.[110] Beziehen sich die Bedingungen lediglich auf eine Relation, wird von intrarelationalen Bedingungen gesprochen, während interrelationale Bedingungen mehrere Relationen umfassen.[111] Als weitere Unterscheidungskriterien für Integritätsbedingungen können die Ausdrucksfähigkeit der für die Formulierung benötigten Sprache, der Überprüfungszeitpunkt sowie die Reaktion auf eine Integritätsverletzung genannt werden.[112] Einige wichtige statische Integritätsbedingungen sind in Tabelle 2.2 zusammengefasst und werden im folgenden erläutert.[113]

Der Wertebereich eines Attributes $dom(A_i)$ legt die Menge der unterschiedlichen Ausprägungen fest, die ein Attribut annehmen kann. Üblicherweise steht hierfür

[110] Vgl. Milek et al. (2001), S. 191.

[111] Vgl. Vossen (2000), S. 119.

[112] Vgl. Heuer und Saake (2000), S. 507f.

[113] Vgl. hierzu Vossen (2000), S. 149f.; Elmasri und Navathe (1994), S. 143-149; Kemper und Eickler (1996), S. 119f.

eine Menge von elementaren Wertebereichen und auf ihnen definierten Operationen zur Verfügung. In der Informatik ist dieses Konzept unter dem Begriff des abstrakten Datentyps bekannt.[114] Für relationale Datenbanken werden meist drei fundamentale Datentypen als Wertebereiche zur Verfügung gestellt:[115]

- Zahlen
- Zeichenketten
- Datumstyp

Häufig, insbesondere in kommerziellen Systemen, wird der Wertebereich zur Erfassung unbekannter Werte um einen Nullwert erweitert. Für unterschiedliche Interpretationen kann es allerdings zweckmässig sein, mehrere Nullwerte zu definieren.[116] Beispielsweise wäre folgende Alternative denkbar:

- Es sind keine Informationen über den realen Wert bekannt.
- Realer Wert existiert, ist aber nicht bekannt.
- Realer Wert existiert nicht.
- Falscher oder nicht sinnvoller Wert existiert.

Aus der Definition einer Relation als eine Menge von Datentupeln wird bereits deren Unterscheidung gefordert. Datentupel müssen paarweise unterschiedlich sein. Im allgemeinen wird zur eindeutigen Identifikation einzelner Tupel eine Teilmenge der Attribute als Schlüssel festgelegt. Eine besondere Integritätsbedingung in diesem Zusammenhang wird als Entity-Integrität bezeichnet und besagt, dass kein Attributwert des (Primär-)Schlüssels einen Nullwert aufweisen darf.[117] Verwendet man die Schlüsselattribute einer Relation als Attribute in einer anderen Relation, so spricht man von einem Fremdschlüssel. Die Menge der Attribute K sind ein Fremdschlüssel einer Relation s, falls für alle Tupel von s gilt, dass deren

[114] Vgl. Heuer und Saake (2000), S. 507f.
[115] Vgl. Kemper und Eickler (1996), S. 88; Elmasri und Navathe (1994), S. 143.
[116] Vgl. Vossen (2000), S. 117f.
[117] Vgl. Elmasri und Navathe (1994), S. 147.

Attributwerte in K Null sind oder ein Tupel in einer Relation r existiert, dessen Primärschlüssel die gleichen Attributwerte aufweist.[118] Die Erfüllung dieser Eigenschaft wird allgemein als referentielle Integrität bezeichnet.[119]

Die bisher beschrieben statischen Integritätsbedingungen stellen lediglich einen Teil aller möglichen Bedingungen dar.[120] Weitere Möglichkeiten für Integritätsbedingungen werden im Rahmen der Fallstudie in Kapitel 4 erläutert. Einige Integritätsbedingungen können bereits direkt bei der Definition des Relationsschemas in SQL beispielsweise als `not null`-Zusatz und in Form von sogenannten `check`-Klauseln spezifiziert werden.[121] Darüber hinaus besteht die Möglichkeit weitere Integritätsbedingungen explizit durch Integritätsregeln festzulegen.[122] Mit Hilfe dieser können dann relativ einfach zahlreiche Bedingungen, insbesondere interrelationale Bedingungen, implementiert werden. Eine Integritätsregel $IR = [B, O, A, R]$ besteht grundsätzlich aus den vier Komponenten:

B Integritätsbedingung.

O Menge von Datenbankobjekten, auf die sich die Bedingung bezieht.

A Auslöser, wann die Bedingung zu überprüfen ist.

R Die Reaktion bei Verletzung der Integritätsbedingung.

Im Rahmen einer Erweiterung von SQL wurde bereits sehr früh ein Vorschlag entwickelt, um Integritätsbedingungen abzubilden.[123] Dieser konnte sich allerdings bislang noch nicht durchsetzen. Die grundsätzliche Form der Regeln beträgt:

```
assert <Regelname>
       [immediate | deferred] [on <Operation> ]
       [for <Relation>]:
       <Bedingung analog where-Klausel>
       [else ( <Folge von SQL-Anweisungen>)]
```

[118] Vgl. Kemper und Eickler (1996), S. 120; Elmasri und Navathe (1994), S. 147.

[119] Vgl. Kemper und Eickler (1996), S. 120.

[120] Vgl. Elmasri und Navathe (1994), S. 149; es lassen sich über 90 Typen von Integritätsbedingungen beschreiben; vgl. Winter (1998), S. 24; Thalheim (1991), S. 4.

[121] Vgl. Vossen (2000), S. 151; Heuer und Saake (2000), S. 503-505.

[122] Vgl. im folgenden Heuer und Saake (2000), S. 508-513; zur formalen Beschreibung von Konsistenzbedingungen vgl. u. a. Winter (1998), S. 24-26 und die dort angegeben Literatur.

[123] Vgl. Heuer und Saake (2000), S. 509ff.; siehe auch Date (2000), S. 269f.

Diese Integritätsbedingungen können , wie in Abschnitt 4.2.3 erläutert, in einer leicht modifizierten Form prinzipiell zur Prüfung der Datenqualität eingesetzt werden. Im folgenden Abschnitt wird als zweite wichtige Komponente eines Datenhaltungssystems die Kommunikationsschnittstelle besprochen.

2.3.2 Kommunikationsschnittstelle

Über die Kommunikationsschnittstelle greifen Benutzer auf die Daten im Datenbanksystem zu, wobei je nach Aufgaben und Erfahrungen der jeweiligen Benutzergruppen unterschiedliche Schnittstellen zur Verfügung stehen.[124] Während routinemässige Aufgaben durch speziell optimierte Anwendungsprogramme ermöglicht werden, können komplexe Anforderungen über integrierte Programmroutinen durchgeführt werden. Zur Verwaltung des Datenbanksystems wird eine besondere Datenbankschnittstelle zur Verfügung gestellt. Für interaktive, flexible Anfragen steht meist eine Abfragesprache,[125] wie beispielsweise SQL, zur Verfügung. Diese ist in heutigen relationalen Datenbanksystemen als Standardabfragesprache verfügbar und soll aufgrund der zentralen Bedeutung durch einige Beispiele verdeutlichen werden.[126]

Eine Relation wird mit der `create`-Anweisung erzeugt:

```
create table Produkt (
       ProdNr integer not null,
       ProdKategorie varchar(3),
       ProdBezeichnung varchar(25) not null,
       ProdFarbe varchar(6),
       primary key (ProdNr),
       foreign key (ProdKategorie) reference Produktkategorien
                                   on delete set null);
```

[124] Vgl. Kemper und Eickler (1996), S. 25.

[125] Der Begriff Abfragesprache ist historisch geprägt und umfasst im allgemeinen auch Befehle zur Datendefinition und Datenmanipulation; vgl. Kemper und Eickler (1996), S. 87 und Date (2000), S. 84.

[126] Eine vollständige Beschreibung der Abfragesprache soll hier nicht erfolgen. Einen Überblick über SQL gibt beispielsweise Date (2000), S. 83ff.

Nach der Relationenbezeichnung folgt in Klammern eine Liste der Attribute und ihrer Datentypen. Die Einschränkung `not null` erzwingt, dass alle Tupel der Relation an dieser Stelle einen definierten Wert aufweisen. Neben den Attributbeschreibungen können auch Integritätsbedingungen festgelegt werden, wie beispielsweise Fremdschlüsselbeziehungen. Schemamodifikationen sind durch den `alter table`-Befehl möglich. Nicht mehr benötigte Relationen können mit dem Befehl `drop table` entfernt werden. Nachdem die Relation angelegt wurde, können konkrete Werte über den Befehl `insert into` tupelweise hinzugefügt werden. Durch die Befehle `update` und `delete` sind weitere Datenoperationen möglich. Verletzen Operationen zuvor festgelegte Integritätsbedingungen, werden sie abgebrochen.

Abfragen sind in SQL relativ einfach aufgebaut und bestehen aus drei Teilen:

```
select ProdNr, ProdKategorie
from   Produkt
where  Farbe = 'blau';
```

Der `select`-Teil bestimmt die im Ergebnis auszugebenden Attribute. Alle in der Abfrage benötigten Relationen sind im `from`-Teil anzugeben, wobei Bedingungen schliesslich im `where`-Teil genannt werden. Abfragen, die mehrere Relationen umfassen sind ebenfalls möglich. Dabei muss die aus dem Kreuzprodukt der Relationen entstehende Ergebnismenge auf solche beschränkt werden, die über gleiche Schlüsselwerte verfügen. Eine Anfrage über die beiden Relationen Produkt und Produktkategorien könnte wie folgt lauten:

```
select ProdNr
from   Produkt, Produktkategorien
where  Produkt.ProdKategorie = Produktkategorien.ProdKategorie
       and Produktkategorien.Bezeichnung = 'Fahrzeuge';
```

Zur Abfrage von multidimensionalen Datenbanken bzw. multidimensionalen Strukturen wird eine Erweiterung von SQL, die Multidimensional Expressions (MDX) vorgeschlagen.[127] Sie besitzt folgende Struktur:

[127] Es wird allgemein erwartet, dass sich der von Microsoft vorgeschlagene Standard OLE DB FOR OLAP mit der Abfragesprache MDX durchsetzt; vgl. Microsoft Corp. (1998); Albrecht et al. (2001), S. 170; Clausen (1998), S. 106-109.

```
[<Berechnungsformel>]
select   [<Dimensionsbeschreibung> [, <Dimensionsbeschreibung>...]]
from     <Multidimensionales Datenobjekt>
where    <Restriktionen>
<Zelleigenschaften>
```

Ausgehend von einem oder mehreren multidimensionalen Datenobjekten, wird im `select`-Teil eine multidimensionale Ergebnismenge beschrieben. Die Dimensionsbeschreibungen definieren die als Achsen bezeichneten Dimensionen und die Granularität. Formelausdrücke erlauben die Angabe von Berechnungsvorschriften zur Transformation der ursprünglichen Ergebniswerte. Im `from`-Teil werden die benötigten multidimensionalen Datenobjekte angegeben. Die `where`-Klausel spezifiziert die Restriktionen auf den Dimensionen der Datenobjekte. Eigenschaften einzelner Zellen können durch die Angabe von Zelleigenschaften festgelegt werden.

2.4 Analytische Informationssysteme

2.4.1 Vielfalt der Informations(teil)systeme

Für die Untersuchung betrieblicher Informationssysteme wird eine Strukturierung dieser benötigt. Grundsätzlich bietet sich hierfür eine horizontale oder eine vertikale Strukturierung von Informationssystemen an.[128] Horizontal findet eine Unterteilung entlang der Leistungsprozesse (Wertschöpfungskette) durch die betriebswirtschaftlichen Grundfunktionen statt. Üblicherweise werden u. a. die Funktionen Beschaffung, Produktion und Vertrieb sowie die Querschnittsfunktionen Produktionstechnik, Logistik und Personal betrachtet. Eine vertikale Differenzierung in verschiedene Ebenen wird anhand der Art der informationsverarbeitenden Aufgabe in Transformations- und Entscheidungsaufgaben vorgenommen.[129] Da bei Transformationsaufgaben die Beziehung zwischen Informationsinput und -output funktional ist, erzeugen sie eine Informationsausgabe aussch-

[128] Vgl. Heine (1999), S. 29.
[129] Vgl. Ferstl und Sinz (2001), S. 30-32.

liesslich in Abhängigkeit des Informationsinputs. Entscheidungsaufgaben sind, im Gegensatz zu den Transformationsaufgaben, durch die Wahl einer Alternative hinsichtlich einer Zielgrösse gekennzeichnet.

Wenngleich diese Strukturierung theoretisch als geeignet erscheint, ist sie zur Untersuchung der in der Praxis vorherrschenden Informationssysteme nicht hinreichend detailliert. Zahlreiche Autoren verfeinern daher diese vertikale Differenzierung von Informationssystemen. STAHLKNECHT und HASENKAMP unterteilen Informationssysteme vertikal in Administrations- und Dispositionssysteme für die operative Ebene, Führungssysteme für die Führungsebene und Querschnittssysteme zur Unterstützung beider Aufgabenebenen.[130] MERTENS unterscheidet Informationssysteme nach der Art der betriebswirtschaftlichen Aufgabe in Administrations-, Dispositions-, Planungs- und Kontrollsysteme.[131] FERSTL und SINZ unterscheiden Informationssysteme in operative Informationssysteme, Steuerungs- und Kontrollsysteme sowie Planungssysteme.[132] SCHEER unterteilt vertikal fünf Informationssystemebenen.[133] Mengenorientierte operative Systeme umfassen mengenorientierte Prozesse. Auf diesen setzen wertorientierte Abrechnungssysteme auf und umfassen eine betriebswirtschaftliche, monetäre Sicht. Auf diesen Systemen basieren Berichts- und Kontrollsysteme für die Berichterstattung und operative Steuerung. Auf einer weiteren Stufe befinden sich Analysesysteme, die neben den verdichteten Daten der operativen Systeme und der wertorientierten Abrechnungssysteme auch Daten externer Quellen einbeziehen. Die höchste Verdichtungsstufe bilden dann die Planungs- und Entscheidungssysteme, die vor allem die strategischen Planungs- und Entscheidungsprozesse unterstützen sollen. Die horizontale Gliederung nach betriebswirtschaftlichen Grundfunktionen und die vertikale nach Art der informationsverarbeitenden Aufgaben ergeben die in Abbildung 2.6 dargestellte Pyramide der Informationssysteme.

Da betriebliche Anwendungssysteme in der Praxis meist mehrere vertikale Ebenen bzw. horizontale betriebswirtschaftliche Funktionen abdecken,[134] ist sowohl

[130] Vgl. Stahlknecht und Hasenkamp (1999), S. 344ff.
[131] Vgl. Mertens (1995), S. 11ff.
[132] Vgl. Ferstl und Sinz (2001), S. 32-36.
[133] Vgl. Scheer (1998), S. 4ff.
[134] Vgl. Heine (1999), S. 30.

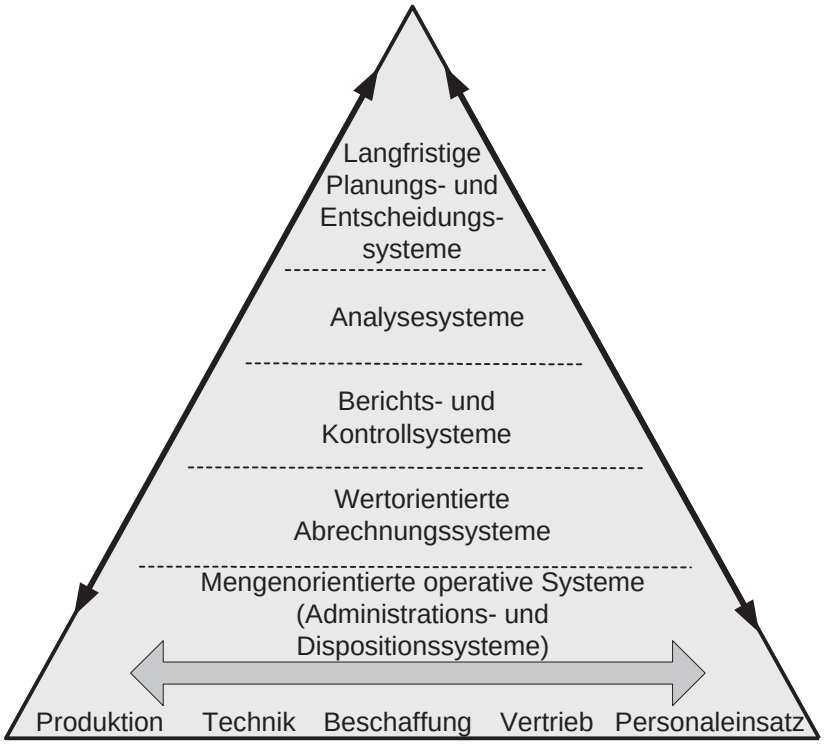


Abbildung 2.6: Pyramide der Informationssysteme (In Anlehnung an Scheer (1998), S. 5)

die horizontale als auch die vertikale Abgrenzung eher als Strukturierungshilfe zu verstehen. Aus diesem Grund sollen Informationssysteme begrifflich in die eher leistungsunterstützenden *„operativen Systeme“* und die vorwiegend *„analytischen Systeme“* unterteilt werden,[135] wenngleich eine strikte Trennung nicht möglich erscheint.[136] Während operative Systeme durch die operative Unterstützung der betrieblichen Leistungsprozesse eines Unternehmens, auch Geschäftsprozesse genannt,[137] gekennzeichnet sind, orientieren sich die analytischen Systeme an der Informationsversorgung betrieblicher Fach- und Führungskräfte zu Analysezwecken.[138]

[135] Vgl. Chamoni und Gluchowski (1998), S. 10f.
[136] Vgl. Devlin (1997), S. 15.
[137] Vgl. Becker und Vossen (1996), S. 18-19.
[138] Vgl. Chamoni und Gluchowski (1998), S. 11.

2.4.2 Data-Warehouse-Systeme

Operative Systeme sind heute meist ausgereift und äusserst stabil sowie in Form betriebswirtschaftlicher Standardsoftware für nahezu jeden Anwendungsbereich erhältlich.[139] So ist auch die methodische Unterstützung zur Entwicklung und Einführung operativer Anwendungssysteme weitgehend berücksichtigt.[140] Zwar lassen sich erste Ansätze für eine Unterstützung betrieblicher Entscheidungsträger bereits Anfang der 60er Jahre erkennen,[141] allerdings ist es bislang nicht gelungen die Erwartungen an analytische Informationssysteme zufriedenstellend zu lösen.[142] Im Laufe der letzten Jahrzehnte haben sich zahlreiche Ansätze analytischer Systeme entwickelt,[143] wobei im Gegensatz zu den operativen Systemen bislang meist noch durchgängige und praktisch umsetzbare Lösungsvorschläge fehlen.[144]

Ein in den letzten Jahren häufig diskutiertes und in zahlreichen Unternehmen eingeführtes Konzept[145] analytischer Informationssysteme kann allgemein unter dem Begriff des „Data Warehousing“ zusammengefasst werden.[146] Zentrale Komponente des Konzeptes ist eine als Data-Warehouse-Datenbank[147] bezeichnete Datenbasis.[148] Sie stellt eine von den operativen Systemen losgelöste, logisch zentralisierte, einheitliche und konsistente Datenbasis für analytische Informationssysteme dar,[149] in der periodisch relevante Daten zusammengetragen, bereinigt und für einen schnellen Zugriff archiviert werden.[150] Umfassend wird die Gesamtheit der Anwendungen und Datenbanken, die diese Datenbasis nutzbar macht, als

[139] Vgl. Chamoni und Gluchowski (1998), S. 10.

[140] Vgl. beispielsweise die Methoden in Scheer (1998), S. 10ff., Ferstl und Sinz (2001), S. 180ff. oder Österle (1995), S. 31ff.

[141] Vgl. Chamoni und Gluchowski (1998), S. 6.

[142] Vgl. Chamoni und Gluchowski (1998), S. 10.

[143] Vgl. Holthuis (1999), S. 36; Chamoni und Gluchowski (1998), S. 6-9.

[144] Vgl. Jung und Winter (2000), S. 19.

[145] Vgl. hierzu u. a. Devlin (1997), S. 8; Holten (1999), S. 39; Inmon (1996), S. 33; Holten, Rotthowe und Schütte (2001b), S. 3.

[146] Vgl. Jung und Winter (2000), S. 5; Flade-Ruf (1996), S. 25; Devlin (1997), S. 129.

[147] Teilweise auch als Data Warehouse i. e. S. bezeichnet; vgl. Winter (2000), S. 128; Holten (1999), S. 42.

[148] Vgl. Holten et al. (2001b), S. 5; Inmon, Imhoff und Sousa (1998), S. 51; Inmon, Zachman und Geiger (1997), S. 30-39.

[149] Vgl. Devlin (1997), S. 20.

[150] Vgl. Rautenstrauch (1997), S. 104.

Data-Warehouse-System bezeichnet.[151]

2.4.2.1 Komponenten eines Data-Warehouse-Systems

Abbildung 2.7 gibt einen groben Überblick über die wesentlichen Komponenten eines Data-Warehouse-Systems, die im folgenden kurz beleuchtet werden sollen.[152] Ausgangspunkt sind operative und unternehmensexterne Anwendungssysteme, welche als Datenlieferanten agieren. Mit Hilfe einer Transformationskomponente werden diese Daten extrahiert, transformiert und in die zentrale Data-Warehouse-Datenbank überführt. Aufgrund des Umfangs der Data-Warehouse-Datenbank, werden häufig kleinere, redundante Teilausschnitte aus der unternehmensweiten Datenbasis vorgehalten. Diese, allgemein als Data Mart bezeichneten, Datenhaltungskomponenten dienen zur fachspezifischen Datenbereitstellung eines auf die jeweiligen Informationsbedürfnisse zugeschnittenen Datenausschnitts und sind auf einen bestimmten Analysezweck optimiert. Die Datenbestände in Form der zentralen Data-Warehouse-Datenbank und der Data Marts stellen die Datenbasis für analyseorientierte Systeme dar. Diese werden durch Endbenutzerwerkzeuge den Anwendern zugänglich gemacht. Bezüglich der Datendarstellung und der Analysemöglichkeiten existiert bei den Endbenutzerwerkzeugen eine vielfältige Produktpalette.[153] Neben Eigenentwicklungen stehen heute funktionsreiche Tabellenkalkulationsprogramme sowie Werkzeuge zur Erstellung von Berichten, Ad-Hoc-Abfragen und vielschichtigen, mehrdimensionalen Auswertungen[154] als auch Möglichkeiten zur Erkennung von Datenbeziehungen zur Verfügung.[155] Zunehmend setzt sich auch die Datendarstellung auf Browser-basierten Abfragewerkzeugen durch.[156]

[151] Vgl. Winter (2000), S. 128.

[152] Vgl. Bange et al. (2001), S. 34ff.; Jung und Winter (2000), S. 10-13; Chamoni und Gluchowski (1998), S. 12; Müller (2000), S. 104ff.; Schelp (2000), S. 115ff.

[153] Vgl. Schinzer, Bange und Mertens (1999), S. 62.

[154] Diese Werkzeuge werden häufig als Online-Analytical Processing-Werkzeuge (kurz OLAP-Werkzeuge) bezeichnet. Sie bieten in der Regel umfassende Funktionen für Navigation und Analyse an, die auch als Drill Down, Roll Up und Drill Across Funktionalitäten bezeichnet werden; vgl. Schinzer et al. (1999), 65 u. 67-71.

[155] Diese Werkzeuge basieren auf den Methoden des Data Mining und werden daher allgemein als Data-Mining-Werkzeuge bezeichnet; vgl. Schinzer et al. (1999), S. 99.

[156] Vgl. Schinzer et al. (1999), S. 62.

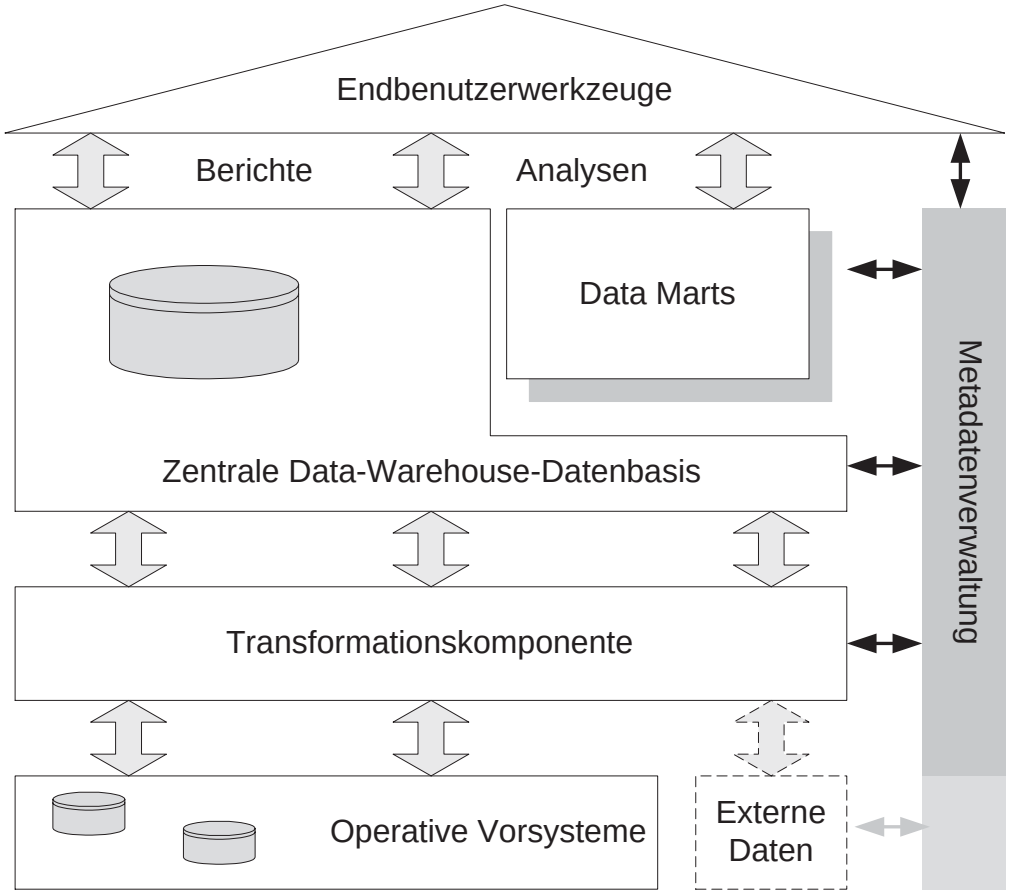


Abbildung 2.7: Data-Warehouse-System (In Anlehnung an Müller (2000), S. 104)

Auf der physischen Ebene besteht ein Data-Warehouse-System im wesentlichen aus Datenhaltungssystemen und Datentransformationskomponenten sowie Softwarekomponenten zur Datenanalyse und -darstellung. Datenhaltungssysteme werden zur Datenhaltung in den Quellsystemen, der zentralen Data-Warehouse-Datenbasis und den Data Marts eingesetzt. Zur Zeit sind im wesentlichen zwei Datenhaltungstechniken vorherrschend.[157] Einerseits ist die multidimensionale und andererseits die relationale Datenhaltung zu nennen. Wenngleich sich in den operativen Systemen die relationale Datenhaltung weitgehend durchgesetzt hat, finden sich noch eine Vielzahl anderer, häufig historisch gewachsener, Datenhaltungstechniken. Solche Datenverwaltungssysteme basieren beispielsweise auf dem Netzwerkmodell oder dem hierarchischen Datenmodell. Die objektorientier-

[157] Vgl. Eicker (2001), S. 70; Bange et al. (2001), S. 118f.

te Datenhaltung hat sich bislang für Data-Warehouse-Systeme noch nicht durchsetzen können. Bei der multidimensionalen Datenhaltung werden die Daten entsprechend einer mehrdimensionalen Würfelstruktur, wie sie sich für die Sicht des Endbenutzers auf entscheidungsrelevante Daten vielfach bewährt hat, auch physikalisch in multidimensionalen Zellstrukturen gespeichert. Eine Alternative zur multidimensionalen Speicherung bildet die Nutzung relationaler Datenbanksysteme und deren Aufbereitung in multidimensionale Datenstrukturen. Eine solche Softwarekomponente wird häufig von Datenbanksystemanbietern zur Verfügung gestellt, da viele Endbenutzerwerkzeuge nicht direkt auf relationale Datenstrukturen zugreifen können.

Meist greifen die eingesetzten Endbenutzerwerkzeuge und Softwarekomponenten nicht direkt auf die Datenhaltungssysteme zu, sondern über eine Client/Server-Architektur mit einer unterschiedlichen Anzahl von Schichten.[158] Häufig erfolgt der Datenzugriff der Endbenutzerwerkzeuge, aufgrund fehlender Standards, über eine multidimensionale Programmierschnittstelle (API).[159] Da diese bisher herstellerspezifisch sind und deren Entwicklung sich aufwendig zeigt, sind viele Endbenutzerwerkzeuge nur für spezielle Datenbanken einsetzbar. Derzeit lassen sich mehrere Standardisierungsansätze in diesem Bereich erkennen.[160] Neben dem direkten Zugriff auf Datenbanken durch Programmierschnittstellen bietet sich der Einsatz von Middleware-Systemen an.[161] Diese Systeme legen eine anwendungsunabhängige Zwischenschicht zwischen die verteilten, heterogenen Plattformen und Anwendungen. Beispiele für Middleware Architekturen sind die Common Object Request Broker Architecture (CORBA) der Object Management Group (OMG) und die Distributed Computing Environment (DCE), die unter anderem von Microsoft durch das Distributed Component Object Model (DCOM) implementiert wird.

[158] Vgl. Bange et al. (2001), S. 119ff.; Schinzer et al. (1999), S. 58.
[159] Vgl. Schinzer et al. (1999), S. 60.
[160] Vgl. Schinzer et al. (1999), S. 60.
[161] Vgl. Bange et al. (2001), S. 130ff.

2.4.2.2 Organisatorische Gestaltung und Anspruchsgruppen

Für den Aufbau und den Betrieb eines Data-Warehouse-Systems sind zahlreiche Aufgaben durch Aufgabenträger durchzuführen. Im folgenden sollen daher einige wesentliche Aufgabenbereiche genannt und die zentralen Aufgabenträger sowie deren organisatorische Gestaltung dargestellt werden. Auf diesen Ausführungen aufbauend, werden abschliessend idealtypische Anspruchsgruppen eines Data-Warehouse-Systems abgeleitet.

Als wesentliche Aufgabenbereiche für Data-Warehouse-Systeme lassen sich

- die Ermittlung der Informationsbedürfnisse der Endanwender,
- deren Umsetzung in geeignete Datenmodelle und in ein geeignetes Datenbereitstellungskonzept sowie
- die regelmässige Produktion und Bereitstellung der Daten

identifizieren.[162] Der Informationsbedarf lässt sich beschreiben als die Menge, Art und Qualität der Informationen, die ein Informationssubjekt (Entscheidungsträger) in einem gegebenen Informationskontext (Entscheidungssituation) zur Bewältigung einer Aufgabe zu einer bestimmten Zeit benötigt.[163] Dieser kann weiter in den subjektiven und den objektiven Informationsbedarf unterschieden werden.[164] Während der subjektive Informationsbedarf vom jeweiligen Informationssubjekt festgelegt wird, ist der objektive Informationsbedarf durch den Bedarf zur Bewältigung einer Aufgabe bestimmt. Der objektive Informationsbedarf ist dabei aber nicht notwendigerweise deckungsgleich mit dem subjektiven.[165] Im allgemeinen kann davon ausgegangen werden, dass das Informationssubjekt lediglich eine Teilmenge seines subjektiven Informationsbedarfes in Form einer Informationsnachfrage äussert. Somit bildet die tatsächliche Informationsnachfrage eine Teilmenge des subjektiven Informationsbedarfes. Dem Informationsbedarf

[162] Vgl. Meyer (2000), S. 84.
[163] Vgl. Henneböle (1995), S. 64; Schelp (2000), S. 98; Wolf (1999), S. 41; Augustin (1990), S. 118.
[164] Vgl. Holthuis (1999), S. 18.
[165] Vgl. Streubel (1996), S. 29; Wolf (1999), S. 41.

steht das Informationsangebot gegenüber.[166] Dieses stellt die Gesamtheit der Informationen dar, die dem Informationssubjekt zu einem bestimmten Zeitpunkt zur Verfügung stehen.[167] Die sich hieraus ergebenden Mengenbeziehungen zwischen objektivem und subjektivem Informationsbedarf, der Informationsnachfrage und dem Informationsangebot sind in Abbildung 2.8 dargestellt.

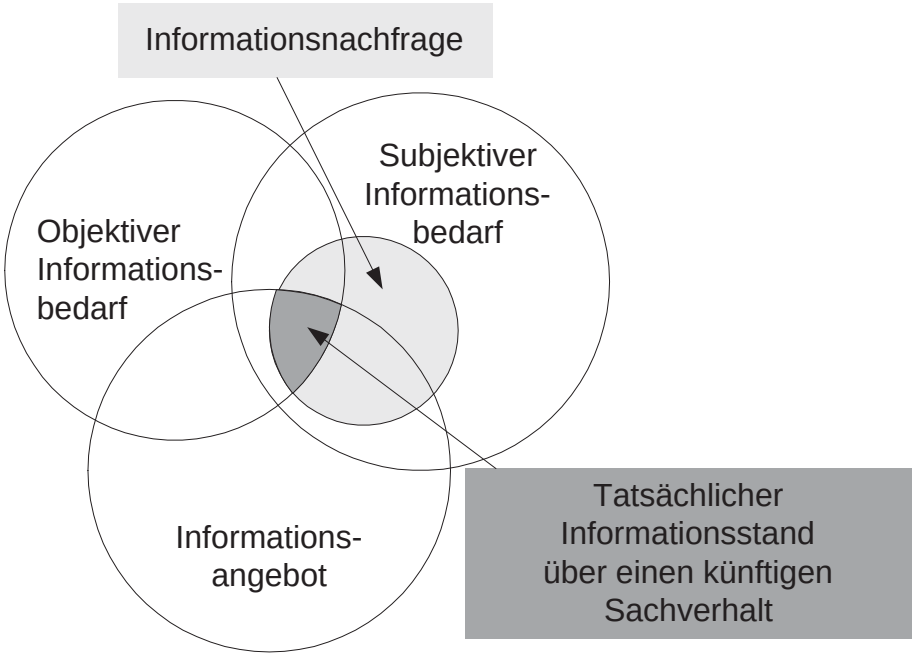


Abbildung 2.8: Informationsangebot, -nachfrage und -bedarf (Vgl. Holthuis (1999), S. 19)

Obwohl bereits zahlreiche Techniken und Verfahren existieren,[168] stellt sich die Ermittlung des Informationsbedarfes bislang als äusserst schwierig dar.[169] In der Literatur werden deduktive und induktive Verfahren der Informationsbedarfsanalyse diskutiert.[170] Bezogen auf die jeweiligen Informationsverarbeitungsaufgaben, versuchen deduktive Verfahren den aufgabenbezogenen richtigen Informa-

[166] In diesem Zusammenhang sei darauf verwiesen, dass im Rahmen eines Anwendungssystems, wie beispielsweise eines Data-Warehouse-Systems, eigentlich von Datenangebot gesprochen werden sollte. Allerdings soll dennoch die allgemein übliche Begriffsverwendung beibehalten werden. Zum hier verwendeten Daten- und Informationsbegriff siehe Abschnitt 2.1.

[167] Vgl. Picot und Frank (1988), S. 609; Wolf (1999), S. 42.

[168] Vgl. Meyer (2000), S. 118.

[169] Vgl. Holten (1999), S. 125.

[170] Vgl. Berthel (1992), S. 876 ff.; Holten (1999), S. 120-125.

tionsbedarf zu ermitteln. Induktive Verfahren zielen auf den personenbezogenen, subjektiven Informationsbedarf ab. Nach den verwendeten Verfahren der Informationsbedarfsanalyse können entweder isolierte oder kombinierte Verfahren genannt werden.[171] Zu den isolierten Verfahren zählen beispielsweise Interviews, Fragebögen, Beobachtung, Brain-Storming-Sitzungen, Aufgabenanalysen und Dokumentenanalysen. Aufgrund der Nachteile einzelner Verfahren wird bei kombinierten Verfahren versucht, die Vorteile der jeweiligen Methoden auszuschöpfen. Ein weit verbreiteter Ansatz zur Ermittlung des Informationsbedarfs basiert auf der Methode der kritischen Erfolgsfaktoren.[172] Kritische Erfolgsfaktoren sind solche Grössen, die für den Erfolg des Unternehmens von entscheidender Bedeutung sind und so die entscheidenden Steuergrössen des Unternehmens darstellen. In Interviews mit Führungskräften werden diese Faktoren, die quantitativer oder qualitativer Natur sein können, identifiziert. Anschliessend werden Lücken zwischen dem objektiven Informationsbedarf zur Verfolgung allgemein akzeptierter Erfolgsfaktoren und dem Informationsangebot ermittelt. Basierend auf den jeweiligen Rollen einzelner Führungskräfte bezüglich der Beherrschung der Erfolgsfaktoren wird dann der subjektive Informationsbedarf abgeleitet und irrelevante Informationen aus dem Informationsangebot eliminiert. Ergebnis ist ein System relevanter Informationen auf Fachkonzeptebene, das schliesslich dynamisch weiterentwickelt wird.

Basierend auf der Informationsbedarfsanalyse sind die Systemarchitektur und die Datenbereitstellungsprozesse zu konzeptionieren.[173] Zunächst sind potentielle Datenquellen zu identifizieren und in die Sichtweisen der Endbenutzersysteme zu transformieren. Hierzu sind die Schemata der Datenquellen mit dem der Data-Warehouse-Datenbank und den Data Marts auf konzeptioneller Ebene zu verbinden. Nach der Zuordnung können auf logischer Ebene die Beziehungen zwischen den Datenquellen und den Datensenken festgelegt werden (logisches Mapping).[174] Diese Zuordnung gilt es, in entsprechende Transferprozesse und Programmroutinen auf physischer Ebene umzusetzen und syntaktische und seman-

[171] Vgl. Holten (1999), S. 120.
[172] Vgl. Rockart (1979), S. 84ff.; Holten (1999), S. 123; Picot und Reichwald (1991), S. 278-282.
[173] Vgl. Meyer (2000), S. 122-128; Holthuis (1999), S. 227-230.
[174] Vgl. Eicker (2001), S. 72; Müller (2000), S. 183ff.

tische Heterogenitäten zu überwinden.[175] Hierbei sind häufig Transformationen von verschiedenen Datenhaltungssystemen in relationale Tabellenschemata und multidimensionale Datenmodelle vorzunehmen. Die Transfer- und Transformationsprozesse, im folgenden vereinfachend als Transferprozesse bezeichnet, sind in einem nächsten Schritt operativ zu steuern. Insbesondere ist deren zeitlicher Ablauf zu planen und schliesslich der Datentransfer zu initiieren.

Die Aufgaben im Data-Warehouse-System werden durch Aufgabenträger wahrgenommen. Neben der Unterscheidung zwischen menschlichen und maschinellen Aufgabenträgern lassen sie sich anhand der Verantwortungsbereiche grob den Komponenten eines Data-Warehouse-Systems zuordnen. Zunächst sollen die menschlichen Aufgabenträger anhand einer generellen Organisationsstruktur in Form idealtypischer Rollen betrachtet werden.[176] Ausgehend von diesen Rollen können dann allgemeine Anspruchsgruppen im Rahmen eines Data-Warehouse-Systems abgeleitet werden.

JARKE et al. nennen den Entscheidungsträger (Endanwender), den Data-Warehouse-Administrator, den Data-Warehouse-Designer sowie den Programmierer als grundsätzliche Personengruppen eines Data-Warehouse-Systems.[177] ENGLISH identifiziert eine Reihe von an der Planung und Datenbereitstellung beteiligten Personengruppen.[178] Er nennt die Datenempfänger, die Verantwortlichen für die Datentransformationen, die Datenlieferanten sowie Analysten, Entwickler und Programmierer der Datenmodelle, Datenbanksysteme und Anwendungen. BAUER et al. zählen zum Aufbau eines Data-Warehouse-Systems die Rollen des Projektmanagements, der Architektur, der Datenanalysten, der Spezialisten der Fachabteilung, der Systementwickler sowie der Endanwender auf.[179] Für den Betrieb schlagen sie die organisatorischen Einheiten des Rechenzentrums und des Data-Warehouse-Kompetenzzentrums vor. Neben der Bereitstellung entsprechender Hardwaresysteme stellt das Rechenzentrum den technischen Betrieb sicher, während das Data-Warehouse-Kompetenzzentrum für die Konzeptionierung des

[175] Vgl. Abschnitt 2.4.2.3.
[176] Vgl. hierzu u. a. Meyer (2000), S. 83ff.
[177] Vgl. Jarke, Lenzerini, Vassiliou und Vassiliadis (2000), S. 138.
[178] Vgl. English (1999), S. 54ff.
[179] Vgl. Bauer et al. (2001), S. 369.

Datenflusses, für die Erstellung und Pflege der Datenschemata sowie die Anwenderbetreuung zuständig ist. MEYER und WINTER schlagen eine organisatorische Grundstruktur anhand prozessorientierter Aufgaben vor.[180] Die Aufgaben umfassen die Verantwortung zur Bereitstellung bereichsspezifischer Daten für die Analyse, den Entwurf, die Implementierung sowie den Betrieb eines Data-Warehouse-Systems. Organisatorisch werden die Aufgaben durch eine aus Fachbereichsvertretern (Auftraggeber) und Informatikern (Auftragnehmer) bestehenden Einheit wahrgenommen. Übergreifende organisatorische Einheiten nehmen die Koordination der verschiedenen Datenbereitstellungsprozesse wahr. Rein technische Aufgaben rund um die Entwicklung und den Betrieb werden durch Infrastruktureinheiten übernommen.

Zusammenfassend berücksichtigen obige Vorschläge neben den Endanwendern in Form von Fachabteilungen technische Organisationseinheiten zur Sicherstellung der technischen Infrastruktur. Eine Koordinationseinheit zur Konzeptionierung und Koordination der Datenbereitstellungsprozesse verbindet die technische Infrastruktur mit den fachlichen Anforderungen. Die Endanwender repräsentieren in Form von Vorgaben und Anforderungen die externe Sicht auf das Data-Warehouse-System. Diese Anforderungen werden durch Entwickler in Architekturkonzepten, Datenschematas, Objektdatenflüssen und Transformationsprozessen erfasst und strukturiert. Schlussendlich werden die Konzepte auf der internen Ebene durch Softwareentwickler, Datenbank- und Data-Warehouse-Administratoren in Datenbereitstellungsprozesse umgesetzt. Daher können grob die Anspruchsgruppen eines Data-Warehouse-Systems in fachliche, konzeptionelle und technische Personengruppen untergliedert werden. Diese Personengruppen können die Verantwortung für eine oder mehrere Systemkomponenten übernehmen. Sie beeinflussen die Gestaltung und den Betrieb des Data-Warehouse-Systems wesentlich, so dass sich die in Tabelle 2.3 dargestellte Einordnung der zentralen Anspruchsgruppen ergibt. Unterstützt werden die menschlichen Aufgabenträger durch maschinelle Aufgabenträger beispielsweise in Form von Metadatenverwaltungs- und Datenbankentwicklungswerkzeugen sowie zahlreichen Programmen und Programmschnittstellen, Protokollen und Treibern zur

[180] Vgl. Meyer (2000), S. 83-106; Meyer und Winter (2000), 322-324.

Fachlicher Schwerpunkt / Systemkomponente	fachlich	konzeptionell	technisch
Endbenutzerwerkzeuge			
Data Marts / Multidimensionale Modelle			
Zentrale Data-Warehouse-Datenbasis			
Transformationskomponente			
Operative Vorsysteme			

Tabelle 2.3: Einordnungsrahmen für Anspruchsgruppen

Initiierung und Steuerung der physischen Datentransformation.

2.4.2.3 Datentransformation

Zwischen den operativen Systemen und der Data-Warehouse-Datenbasis stellt die Transformationskomponente die Funktionalitäten bereit, um Daten aus den operativen Vorsystemen zu entnehmen, zu einem konsistenten Gesamtdatenbestand zu transformieren und diesen in die Data-Warehouse-Datenbank einzufügen.[181] Die Funktionalität der Transformationskomponente lässt sich grob über folgende Aufgaben definieren:[182]

- Übernahme der relevanten Datenbestände aus den Vorsystemen.
- Beseitigung syntaktischer Heterogenitäten.
- Beseitigung semantischer Heterogenitäten.
- Verteilung der Quelldaten auf die Modellobjekte des Zielsystems (Mapping).
- Aggregation, Konsolidierung und Umwandlung der Datenbestände in die für die Zwecke des Data-Warehouse-Systems geeignete Form.

[181] Vgl. Müller (2000), S. 143; Eicker (2001), S. 72-76.
[182] Vgl. Müller (2000), S. 144; Bange et al. (2001), S. 49.

Die Datentransformationen[183] leisten einen entscheidenden Beitrag zur Integration von Daten aus heterogenen Datenquellen. Die Datenintegration zielt dabei auf die Beherrschung von Datenredundanz und fokussiert auf die logische Integrität der Datenbestände.[184] Ein Datenbestand soll als „integriert“ bezeichnet werden, wenn ein Zustand vorhanden ist, in dem eine zumindest logisch einheitliche Datenbasis vorliegt, deren Datenbestände in einem gemeinsamen Modell definiert und über eine Schnittstelle zugänglich sind.[185] Die in den Quellsystemen häufig vorherrschenden Qualitätsmängel in Form von fehlerhaften, redundanten, nicht aktuellen oder unvollständigen Daten werden meist durch Datenbereinigungsmassnahmen korrigiert.[186] Die Datenbereinigung kann dabei grundsätzlich in sieben Aufgabenschritte untergliedert werden.[187] Zunächst werden die Daten in atomare Einheiten zerlegt, was allgemein als Parsing bezeichnet wird. Anschliessend werden die Einheiten nach üblichen Regeln syntaktisch vereinheitlicht. Diese können dann mit anderen Referenzdaten, wie beispielsweise Adress- und Produktverzeichnissen, verglichen und eventuell korrigiert werden. Eine wichtige Aufgabe stellt das Eliminieren von Duplikaten und das Zusammenführen gleicher Objekte dar. Datentransformationen passen dann semantische Heterogenitäten an. Eine weitere Aufgabe wird durch das sogenannte „Householding“ durchgeführt, indem Beziehungen zwischen Objekten festgestellt werden. Abschliessend sollten die vorgenommenen Datentransformationen in den Metadaten dokumentiert werden.

Es lassen sich zwei Arten der Datenübernahme unterscheiden.[188] In einem ersten Schritt werden Daten für das *initiale Füllen* aus operativen Systemen und eventueller Archivsysteme entnommen und einmalig in die Data-Warehouse-Datenbank geladen. Während des Betriebs werden, in einem zweiten Schritt, diese Daten dann permanent oder zyklisch *aktualisiert* und ergänzt. Aufgrund der Häufigkeit der Aktionen haben diese Fälle unterschiedliche Bedeutung. So hat das initiale

[183] Auch teilweise unter dem Begriff „data migration“ (Datenmigration) zusammengefasst; vgl. Bange et al. (2001), S. 49).

[184] Vgl. Heine (1999), S. 65.

[185] Vgl. Müller (2000), S. 146.

[186] Vgl. Bange et al. (2001), S. 50; Müller (2000), S. 176ff.; eine Liste von Datenbereinigungswerkzeugen findet sich beispielsweise in English (1999), S. 318ff.

[187] Vgl. Helfert (2000a), S. 74.

[188] Vgl. Edelstein (1997), S. 42; Inmon (1996), S. 76.

Füllen den Charakter einer einmaligen Aktion, während die zyklische Aktualisierung häufig stattfindet und weitaus problematischer ist. Für das permanente Aktualisieren der Datenbestände sind Mechanismen zu implementieren, die laufend die Veränderungen der operativen Datenbestände erkennen und diese in die Data-Warehouse-Datenbank überführen.

Neben der Datenqualität in den operativen Systemen, hängt von diesen Importfunktionalitäten und Transferprozessen die Qualität der Daten des Data-Warehouse-Systems ab. So stellt die Transformationskomponente durch Datenbereinigungsmassnahmen einen zentralen Aspekt zur Sicherung der Datenqualität dar.[189] Da die Datenbereinigung auf die Verbesserung der Datenbestände im Nachhinein abzielt („Symptombekämpfung“), sollen diese Massnahmen als reaktive Qualitätsverbesserungen bezeichnet werden.[190] Während für das initiale Laden meist einmalige Datenbereinigungsaktionen geeignet sind, sind für häufig stattfindende Prozesse die Ursachen mangelnder Datenqualität zu identifizieren und daraus Massnahmen zur Sicherstellung und kontinuierlichen Verbesserung der Datenqualität abzuleiten und umzusetzen.[191] Aufgrund der Bedeutung der permanenten Datenaktualisierung soll diese im Zentrum der weiteren Arbeit stehen und wird in Abschnitt 3.4 unter dem Konzept eines proaktiven Datenqualitätsmanagements betrachtet.

Die Festlegung von Zeitpunkten, an denen die Daten in die Data-Warehouse-Datenbasis übernommen werden, hängt entscheidend von den Daten bzw. von den Anforderungen an die Datenauswertung ab. Prinzipiell kann die Extraktion periodisch, auf Anfrage, aufgrund spezieller Ereignisse oder unmittelbar bei Änderungen vorgenommen werden.[192] Für die technische Ausgestaltung der Datengewinnung aus den Vorsystemen und deren Aufbereitung sind verschiedene Konzepte denkbar, für die zahlreiche Werkzeuge angeboten werden.[193] Grundsätzlich können die Konzepte durch die Art der Aktivitäten zur Datengewinnung charakterisiert werden:[194]

[189] Vgl. Müller (2000), S. 169.
[190] Vgl. English (1999), S. 238.
[191] Vgl. English (1999), S. 285-289.
[192] Vgl. Bange et al. (2001), S. 49.
[193] Vgl. Eicker (2001), S. 75; Müller (2000), S. 166f.
[194] Vgl. Müller (2000), S. 152.

- Techniken, bei denen die operativen Systeme Daten unmittelbar oder über einen Puffer in die Data Warehouse-Datenbank schieben („Push-Techniken“).
- Techniken, bei denen Funktionen des Data-Warehouse-Datenbanksystems die Daten aus den operativen Systemen extrahieren und dabei entweder unmittelbar auf die Quelldaten oder auf dafür bereitgestellte Exportdaten zugreifen („Pull-Techniken“).
- Techniken, bei denen eine separate Komponente die Datenextraktion steuert und dabei Transformationsfunktionen durchführt.

Die Datenübernahme aus den operativen Systemen kann prinzipiell mittels zwei Vorgehensweisen durchgeführt werden.[195] Zunächst können bei jedem Übernahmeprozess alle Datensätze des operativen Systems kopiert („bulk copy“) und in die Data Warehouse-Datenbank integriert werden. Eine Alternative hierzu bietet das gezielte Suchen nach neuen oder aktualisierten Datensätzen („Deltas“) in den operativen Systemen und deren inkrementelle Übernahme in die Data-Warehouse-Datenbank. Während sich die erste Variante vor allem für das initiale Füllen eignet, bietet sich die zweite Variante für den laufenden Export der Daten aus den operativen Vorsystemen an.[196] Das Kernproblem dabei besteht insbesondere in der Ermittlung der relevanten Daten, den sogenannten Delta-Daten. Hierbei können fünf Techniken unterschieden werden:[197]

- Zeitstempel-gesteuerte Verfahren.
- Modifikationen der Anwendungsprogramme zur direkten Übergabe der Extraktionsdaten.
- Protokollierung der relevanten Datenbank-Transaktionen.
- Auswertung von systemeigenen Protokoll-Dateien.
- Vergleich von Schnappschüssen.

[195] Vgl. Müller (2000), S. 154.
[196] Vgl. Müller (2000), S. 154
[197] Vgl. hierzu Müller (2000), S. 155ff.; Bange et al. (2001), S. 47; Inmon (1996), S. 77.

Das Zeitstempel-gesteuerte Verfahren bietet sich für Datenbestände an, in denen die Zeitbezüge bereits erfasst sind.[198] Die Ermittlung der relevanten Delta-Daten kann dann durch Auswertung dieser Zeitbezüge erfolgen, indem genau die ab dem letzten Aktualisierungslauf geänderten Daten verwendet werden. Eine weitere Möglichkeit zur Ermittlung der Delta-Daten besteht, wenngleich in der Praxis mit Problemen behaftet,[199] in der Modifikation der auf die operativen Datenbasis zugreifenden Anwendungssysteme, die dann eine Protokolldatei mit relevanten Delta-Daten erzeugen. Mit Hilfe dieses Verfahrens können die Daten auch direkt in die Data-Warehouse-Datenbank geschrieben werden, was eventuell für einige Ausnahmefälle notwendig sein kann. Durch Erweiterung von Datentabellen oder der Definition eigenständiger Relationen, können in relationalen Datenbanksystemen Protokolle angelegt werden. Diese Protokolltabellen können entweder einen Verweis auf die Datensätze der Originaltabellen oder aber redundant zur Originaltabelle alle relevanten Werte enthalten. Im Rahmen der zyklischen Aktualisierungsläufe sind die Protokolldateien auszuwerten und die relevanten Datensätze in die Data-Warehouse-Datenbasis zu übertragen. Vorteilhaft ist dabei, dass der Datenbestand nicht erneut durchsucht werden muss. Auf den ersten Blick bieten sich anstelle spezieller Protokolldaten die bereits in Datenbankverwaltungssystemen angelegten Systemprotokolle an.[200] Allerdings ist dieses Vorgehen aufgrund technischer Restriktionen und auch aus Aspekten der Datensicherheit mit Vorsicht anzuwenden.[201] Eine weitere Möglichkeit ist der Vergleich von zwei aufeinanderfolgenden Zustandskopien, sogenannter Schnappschüsse. Auf diese Weise können Datenveränderungen erkannt und die Delta-Daten ermittelt werden.[202]

2.4.2.4 Metadatenverwaltung

Die Datenflüsse zwischen den Datenquellen und den Datensenken, die auszuführenden Transformationsschritte sowie deren zeitliche Reihenfolge und Aus-

[198] Alternativ können diese Bezüge auch aus den Beziehungen in den operativen Systemen hergeleitet werden; vgl. Müller (2000), S. 156.
[199] Vgl. Müller (2000), S. 158-160.
[200] Vgl. hierzu Elmasri und Navathe (1994), S. 535f. u. S. 598.
[201] Vgl. Müller (2000), S. 164.
[202] Vgl. Welch (1997), S. 175.

führungszeitpunkte sind zu koordinieren. Hierfür ist eine, im allgemein als Metadatenverwaltung bezeichnete Komponente notwendig.[203] Häufig werden die verwendeten Daten der Metadatenverwaltung vereinfachend als Daten über Daten bezeichnet,[204] was jedoch zur genauen Begriffsbestimmung noch wenig aussagekräftig ist. Der adjektivische Zusatz „Meta“ drückt bereits eine Reflexion und Abstraktion sowie den Wechsel der Bezugsebene aus.[205] HANSEN bezeichnet Metadaten als Daten, welche die Definitionen des internen und externen Schemas betreffen und einzelne Datenobjekte beschreiben und klassifizieren.[206] In der Literatur zu Data-Warehouse-Systemen ist das Begriffsverständnis im allgemeinen sehr umfassend.[207] Dieses Begriffsverständnis umschliesst alle Informationen, die den Aufbau, die Wartung und den Betrieb sowie die Nutzung der Daten in Data-Warehouse-Systemen ermöglichen. Im Gegensatz zur Nutzung von Metadaten in den operativen Systemen ist das Aufgabenspektrum von und somit die Anforderungen an Metadaten in analytischen Systemen und insbesondere in Data-Warehouse-Systemen wesentlich umfassender.[208]

Metadaten lassen sich nach unterschiedlichen Sichten und in verschiedene Ebenen klassifizieren, wobei zwischen den Sichtweisen und Ebenen Interdependenzen bestehen.[209] Eine grobe, aber wichtige Unterscheidung der Metadaten ergibt sich nach der Art des Verwendungszwecks in technische und fachliche Metadaten.[210] Technische Metadaten beinhalten Daten über die technische Beschreibung der Elemente des Data-Warehouse-Systems, über die Datenquellen und die Datensenken, über deren Strukturen sowie über die operativen Transformationsprozesse. Hierzu gehören die Beschreibungen der logischen und physischen Datenschemata, Integritätsbedingungen sowie Implementierungsinformationen der verschiedenen Extraktions- und Transformationsprozesse. Im wesentlichen umfassen

[203] Vgl. Mucksch (1998), S. 134.

[204] Vgl. Devlin (1997), S. 52; Holthuis (1999), S. 95.

[205] Vgl. Lehmann und Ortner (2000), S. 370.

[206] Vgl. Hansen (1996), S. 955f.

[207] Vgl. Bange et al. (2001), S. 68; Müller (2000), S. 72; Schwarz (2000), S. 103f.; Jung und Winter (2000), S. 12.; Holten (1999), S. 46.; Holthuis (1999), S. 98ff.; Holten (1999), S. 46f.; Devlin (1997), S. 52; Brackett (1996), S. 185ff.

[208] Vgl. Müller (2000), S. 120f.; Wieken (1997), S. 272.

[209] Vgl. Mucksch und Behme (2000), S. 24.

[210] Vgl. Bange et al. (2001), S. 68f.; Tozer (1999), S. 120-125; Poe (1998), S. 170f.; Lehmann und Ortner (2000), S. 370; Wieken (1997), S. 275.

technischen Metadaten

- Quelldateninformationen (Feld- und Tabelleninformationen der Anwendungssysteme bzw. Datenhaltungssysteme),
- Zieldateninformationen (Feld- und Tabelleninformationen des Datenhaltungssystems),
- Transferinformationen (Beziehung zwischen Datenquelle und Datensenke),
- Transferschritte (Teilschritte des Tranferprozesses, deren Bedingungen und Abhängigkeiten sowie zeitliche Reihenfolge) und
- Datentransformationen (Transferschritte, die die Datenstruktur und Datenwerte verändern) sowie
- Integritätsbedingungen zur Sicherstellung der Konsistenz in einem Data-Warehouse-System

soweit diese zur Datenbereitstellung analytischer Daten notwendig sind.[211]

Neben den technischen Metadaten werden zur Datennutzung durch den Endanwender fachliche Metadaten benötigt. Diese unterstützen die Endbenutzer des Data-Warehouse-Systems beim Auffinden und der Interpretation der Daten sowie beim Datenzugriff.[212] In diese Kategorie fallen anwendungsspezifische Informationen über Anfragen und Berichte sowie verwendete Begriffsterminologien und domänenspezifisches Wissen.[213] Sie umfassen strukturelle und semantische Informationen über die Daten des Data-Warehouse-Systems und sind so insbesondere für die Interpretation der Daten ausschlaggebend.

Eine weitere Einteilung gliedert Metadaten anhand der Funktionen eines Data-Warehouse-Systems in Datenquell-, Transformations-, Speicher- sowie Abfrage- und Auswertungsinformationen.[214] Eine ähnliche, architekturbasierte Einteilung orientiert sich an den Entstehungs- und Verwendungsorten von Metadaten.[215] Als

[211] Vgl. Tozer (1999), S. 121f.
[212] Vgl. Lehmann und Ortner (2000), S. 371.
[213] Vgl. Bange et al. (2001), S. 68.
[214] Vgl. Holthuis (1999), S. 99f.; Mucksch (1998), S. 135; Mucksch und Behme (2000), S. 24f.
[215] Vgl. Müller (2000), S. 121-133; Schelp (2000), S. 121-124.

wesentliche Komponenten sind die Schnittstelle zwischen Vorsystemen und Data-Warehouse-Datenbank, die Data-Warehouse-Datenbank als auch die Schnittstelle zu den Endbenutzerwerkzeugen zu nennen. MÜLLER betrachtet diese Einteilung differenzierter, indem er zwischen struktur- und funktionsorientierten Metadaten unterscheidet.[216] Während sich strukturorientierte Metadaten auf die Beschreibung der Data-Warehouse-Komponenten und Objektdaten beziehen, beschreiben funktionsorientierte Metadaten die funktionalen und ereignisorientierten Aspekte des Data-Warehouse-Systems. Strukturorientierte Metadaten beschreiben das „Was“, während das „Wie“ und „Wann“ durch funktionsorientierte Metadaten erfasst wird.[217] Die zentralen strukturorientierten Metadaten sind in Tabelle 2.4 zusammengefasst.

Metadaten werden im allgemeinen in einem Datenhaltungssystem gespeichert und verwaltet.[218] Diese Systeme reichen von einfacher Datenhaltung in Tabellen des Datenbanksystems über umfangreiche Datensammlungen in Datenbanken bis hin zu eigenständigen Systemen.[219] Die Begriffsverwendung für diese Systeme ist uneinheitlich.[220] So werden einfachere Datenbeschreibungssysteme häufig als „data dictionary“ bezeichnet, während für umfangreiche Systeme auch die Bezeichnung „repository“ gebräuchlich ist.[221] Neben der Bezeichnung Metadatenverwaltung soll hier in Anlehnung an Datenhaltungssysteme auch die Bezeichnung Metadatenbanksystem oder auch kurz Metadatenbank verwendet werden.[222]

Metadaten basieren auf einem Datenbankschema.[223] Wenngleich die Bezeichnungen Metamodell, Metaschema und Metadatenschema teilweise uneinheitlich verwendet werden,[224] soll hier die in Abschnitt 2.3.1 für Datenmodelle getroffene

[216] Vgl. Müller (2000), S. 121ff.

[217] Vgl. Müller (2000), S. 128.

[218] Vgl. Mucksch (1998), S. 134.; Albrecht et al. (2001), S. 328; Holten (1999), S. 46f.

[219] Vgl. Schreier (2001), S. 129f.; Schwarz (2000), S. 105.

[220] Vgl. Devlin (1997), S. 140; Holten (1999), S. 47; Schelp (2000), S. 121.

[221] Vgl. Stahlknecht und Hasenkamp (1999), S. 210 u. 302; Müller (2000), S. 137; Schelp (2000), S. 121.

[222] Vgl. Holthuis (1999), S. 95.

[223] Vgl. Albrecht et al. (2001), S. 328.

[224] Vgl. beispielsweise Albrecht et al. (2001), S. 328 u. S. 333. Die OMG untergliedert die Modelle bei der Spezifikation des CWM auf 4 Sprachebenen; vgl. Object Management Group, Inc. (2001), S. 4/34f. und Jung und Rowohl (2000), S. 120; vgl. auch Tozer (1999), S. 48; Lehmann und Ortner (2000), S. 377f.

Ebene	Komponente	Schnittstelle zu den Endbenutzersystemen	DWH-Datenbank	Schnittstelle zu den Vorsystemen
Extern		Views einzelner Enbenutzer-Anwendungen	Identifikation, Datenbank-API	Views einzelner Transformations- und Extraktionsmodule
Konzeptionell	Semantisch	Semantische, natürlichsprachliche Beschreibungen der DWH-Inhalte und Verknüpfungen	ER-Modelle, semantische multidimensionale Datenmodelle	Herkunftsbeschreibungen
	Logisch	Schematransformation	Schema-Beschreibung, Tabellenbeziehungsdiagramme, Benutzerprofile	Logisches Mapping auf Tabellen-, Feld- und/oder Dateiebene
Intern		Inter-Programm-Kommunikation, Protokolle, Treiber	Indizes, Indextypen (Bit-Indizierung), Zugriffspfade, Belastungsstatistiken	Physikalisches Mapping

Tabelle 2.4: Strukturorientierte Metadaten (In Anlehnung an Müller (2000), S. 122)

Festlegung übertragen werden. So wird das Schema der Metadaten als Metadatenmodell oder Metadatenschema bezeichnet, während die „Sprache“, welche die Objekt- und Beziehungstypen zur Erstellung des Modells festlegt, als Metamodell des Metadatenschemas bezeichnet wird. Die eigentlichen Metadaten finden sich dann als Ausprägungen bzw. Instanzen des Metadatenmodells in Form konkreter Werte im Metadatenbanksystem.

Für dezentrale Architekturen existiert eine Vielzahl werkzeugspezifischer Schemata, die sich im allgemeinen überschneiden.[225] Leider sind die jeweiligen proprietären Metadatenmodelle weder semantisch noch syntaktisch angepasst, was den Datenaustausch zwischen den einzelnen Metadatenbanken erschwert.[226] Zukünftig erscheint in diesem Zusammenhang die Standardisierung entsprechender Metadatenmodelle und der Zugriffsformen auf diese wichtig, um so die Bindung an einzelne, herstellerbezogene Werkzeuge zu lösen und einen übergreifenden Austausch von Metadaten zu ermöglichen.[227]

Als bedeutendste Bestrebung zur Erstellung standardisierter Metadatenmodelle ist das Industriekonsortium Object Management Group (OMG) zu nennen, indem inzwischen alle bedeutenden Hersteller von Data-Warehouse-Systemkomponenten vertreten sind. Seit September 1998 entwickelt die OMG ein Metamodell zur Spezifizierung von Metadatenmodellen.[228] Das Metamodell stellt eine Sprache dar, die es ermöglicht, Metadatenmodelle zu spezifizieren und so den Datenaustausch zwischen herstellerspezifischen Metadatenmodellen zu erleichtern. Das als Common Warehouse Metamodel (CWM) bezeichnete Metamodell ist eine Erweiterung der grafischen Beschreibungssprache Unified Modeling Language (UML) für die objektorientierte Modellierung von Anwendungssystemen. Es ist, wie in Abbildung 2.9 dargestellt, in 18 Teilmodelle untergliedert. Die Teilmodelle, die auf dem *Object Model* aufbauen, sind in vier Schichten strukturiert. Die vier Schichten sind auf Modellebene als UML-Pakete (Package) modelliert und enthalten wiederum einzelne Teilmodelle.

Grundlegende Strukturelemente und Konzepte, die von mehreren Teilmodellen

[225] Vgl. Albrecht et al. (2001), S. 334; Schwarz (2000), S. 113.
[226] Vgl. Jung und Rowohl (2000), S. 115.
[227] Vgl. Müller (2000), S. 137.
[228] Vgl. hierzu Object Management Group, Inc. (2001).

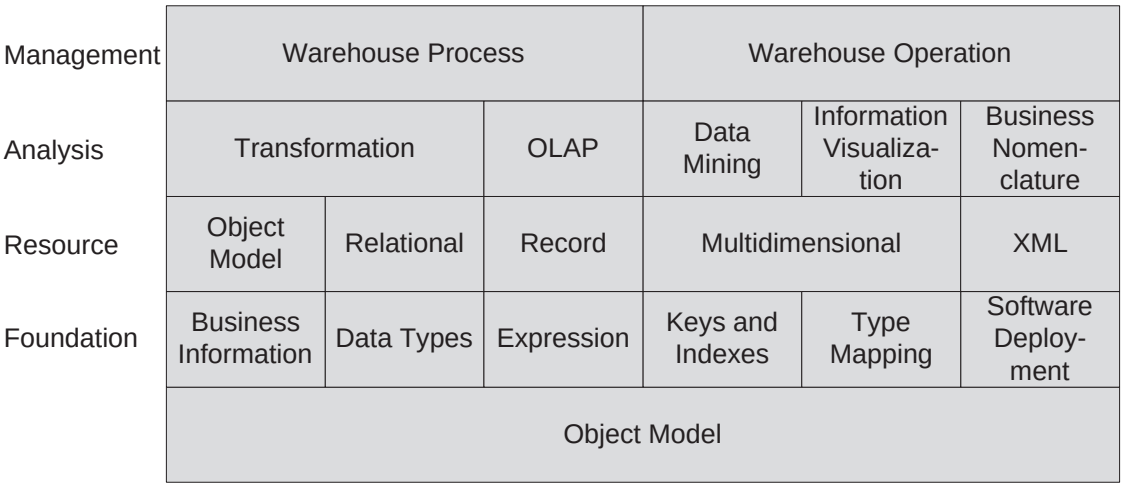


Abbildung 2.9: Das CWM Metamodell (In Anlehung an Object Management Group, Inc. (2001), S. 6/52)

verwendet werden, sind der Schicht *Foundation* zugeordnet. Hier finden sich Konstrukte zur Modellierung von Datentypen, Ausdrücken, Schlüsseln und Indizes sowie zur Beschreibung der eingesetzten Anwendungssysteme. Organisatorische Elemente, wie beispielsweise Personen, Kontaktmöglichkeiten und Dokumentationen können durch das *Business-Information-Paket* erfasst werden. Das *Resource-Paket* umfasst Teilmodelle zur Definition von Datenquellen und Datensenken. Diese können objektorientiert, relational, satzorientiert, multidimensional oder auch XML-basiert sein. Im *Analysis-Paket* sind Teilmodelle zur Beschreibung von Metadaten zur Transformation und zur Analyse der Objektdaten erfasst. Für die Datenanalyse stehen Modelle für OLAP, Data Mining und der Datenvisualisierung zur Verfügung. Fachliche Metadaten werden im *Business-Nomenclature-Paket* erfasst. Hier gibt es Konstrukte für die Modellierung von Begriffssystemen wie Taxonomien und Glossare. Die Modellierung des Betriebs eines Data-Warehouse-Systems wird durch das *Management-Paket* beschrieben. Zur Beschreibung des Ablaufs von Transformationsprozessen mit Hilfe von Ereignissen und Triggern enthält das *Warehouse-Process-Teilmodell* geeignete Konstruktoren. Die Dokumentation des täglichen Warehouse Betriebes wird durch das *Warehouse-Operation-Paket* erfasst.

Zwar existieren zahlreiche Werkzeuge für die Metadatenverwaltung,[229] jedoch sind derzeit noch keine zufriedenstellenden Lösungen verfügbar.[230] So zeigen Untersuchungen, dass Systeme zur Metadatenverwaltung in der Praxis nicht die in der Literatur genannten theoretischen Möglichkeiten aufweisen und so häufig die Erwartungen nicht erfüllen.[231] Wenngleich Metadaten zu den kritischen Erfolgsfaktoren in Data-Warehouse-Systemen gezählt werden,[232] werden bislang existierende Metadaten nicht effizient genutzt. Da die effiziente Verwaltung der Metadaten die Datenqualität beeinflusst, hat dies insbesondere Auswirkungen auf die Datenqualität in Data-Warehouse-Systemen.

2.4.2.5 Betrachtungsebenen für Data-Warehouse-Systeme

Bislang wurden die wesentlichen Komponenten eines Data-Warehouse-Systems als ein Konzept für ein analytisches Informationssystem beschrieben. Diese Betrachtungsebene fokussiert auf die einzelnen Komponenten und betrachtet so die innere Struktur und das Verhalten des Systems. Untersucht man dagegen das Data-Warehouse-System als Teil bzw. Element des betrieblichen Informationssystems, eröffnet sich eine weitere Betrachtungsebene. Zweck des Data-Warehouse-Systems ist die Bereitstellung analytischer Daten im Rahmen der betrieblichen Informationssysteme. Wie in Abschnitt 2.2.1 erläutert, umfassen Informationssysteme in dieser Arbeit das gesamte informationsverarbeitende Teilsystem eines betrieblichen Systems. Hier sind Entscheidungsaufgaben in bezug auf die Planung, Steuerung und Kontrolle des betrieblichen Basissystems zu lösen. Analytische Informationssysteme sind, wie in Abschnitt 2.4 ausgeführt, auf die Informationsversorgung betrieblicher Fach- und Führungskräfte zu Analysezwecken gerichtet. Entscheidungsträger sind hierbei mit entscheidungsrelevanten Informationen zu versorgen. Das Data-Warehouse-System ist dann als ein Teilsystem des analytischen Informationssystems aufzufassen, wobei es einen Teil der analytischen Informationen, die Daten, abdeckt. Neben den Daten des Data-Warehouse-Systems

[229] Vgl. Albrecht et al. (2001), S. 333.
[230] Vgl. Mucksch und Behme (2000), S. 27.; Jung und Winter (2000), S. 13.
[231] Vgl. Holthuis (1999), S. 96.
[232] Vgl. Jung und Rowohl (2000), S. 114.

fliessen weitere Informationen, wie beispielsweise unstrukturierte Daten und informelle Informationen in die Entscheidungsbildung ein.

Wird das Data-Warehouse-System durch Deduktion als eigenständiges System untersucht, ergibt sich die zweite Betrachtungsebene. Es kann dabei in einzelne Komponenten zerlegt werden, die durch ausführen bestimmter Funktionen den Systemzweck erfüllen. Der Systemzweck, das Bereitstellen analytischer Daten, wird durch einzelne Systemkomponenten realisiert. Wesentliche Komponenten stellen Datenhaltungssysteme, Transferprozesse und Endbenutzerwerkzeuge sowie die Metadatenverwaltung dar. Aufgabe der Komponenten und ihrer Funktionen ist es, die Daten aus den operativen Vorsystemen zu extrahieren, in entscheidungsorientierte Datenschemata zu transferieren und Analysemöglichkeiten hierfür anzubieten. Wie in Abschnitt 2.4.2 gezeigt, lassen sich grob die Funktionen

- der Datenerfassung in den operativen Systemen durch *Datenerfassungsprozesse*,
- der Datenextraktion, Datentransformation und des Datentransfers (kurz als *Transferprozesse* bezeichnet) sowie
- der *Datenhaltung* und
- der Datenabfrage, -auswertung und -präsentation als *Datenanalyse*

unterscheiden. Diese Daten, die durch das Data-Warehouse-System „fliessen“ werden als Objektdaten bezeichnet, während die Daten der Metadatenverwaltung als Metadaten zu unterscheiden sind. Metadaten stellen eine Abstraktion betrieblicher Objektdaten dar.

Eine der zentralen Erkenntnisse dieser Betrachtungsweise sind die in einem Data-Warehouse-System zu berücksichtigenden Datenmodelle auf den unterschiedlichen Beschreibungsebenen und Architekturebenen (Vgl. Abbildung 2.10).[233] Auf der Ebene der operativen Systeme, der zentralen Data-Warehouse-Datenbasis als auch auf Ebene der Fachbereiche existieren jeweils konzeptionelle, logische und

[233] Vgl. hierzu Jarke et al. (2000), S. 21-25.

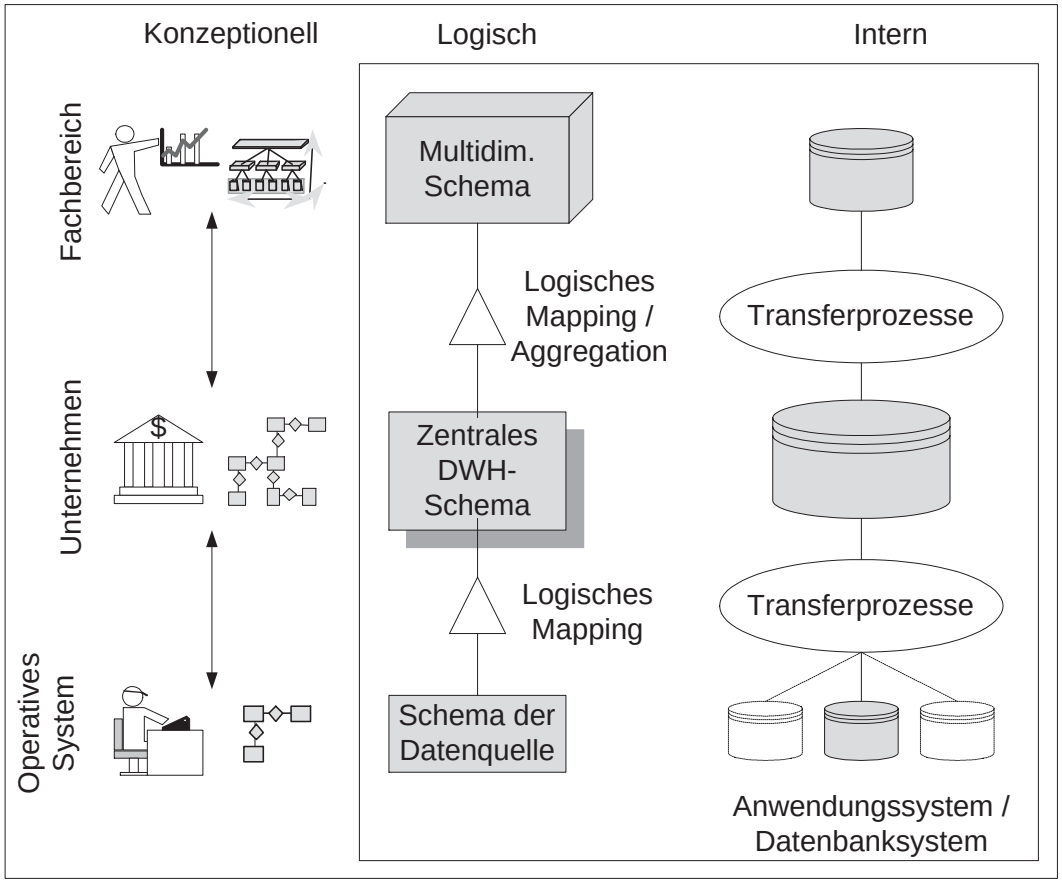


Abbildung 2.10: Betrachtungsweise eines Data-Warehouse-Systems (In Anlehnung an Jarke et al. (2000), S. 24)

interne Schemata der jeweiligen Datenhaltung. Diese stehen durch das Mapping und die Transferprozesse miteinander in Beziehung.

Wie in Abschnitt 1.1 bereits angedeutet, sind die Daten auf Ebene der Fachbereiche häufig in bezug auf die Anforderungen nicht zufriedenstellend. So stellt die Sicherstellung der Datenqualität eines der Hauptprobleme in Data-Warehouse-Systemen dar. Während die Datenqualität in den operativen Einheiten zur Unterstützung der Geschäftsprozesse weitgehend ausreichend ist, ist diese für analytische Fragestellungen oft ungenügend. Die Ursachen für diese Probleme sind vielschichtig und können beispielsweise in nicht berücksichtigten Attributen in Datenmodellen, fehlerhaften Datenerfassungs- und Transferprozessen, inkorrekten Datenaggregationen und -berechnungen durch Softwarekomponenten oder im Ausfall von Hardwarekomponenten liegen.

Zur Analyse dieser Probleme auf den Architekturebenen und -komponenten, ist eine Vergleichsbasis zwischen den verschiedenen Ebenen notwendig. Insbesondere sind die verschiedenen Datenmodelle in bezug auf das Schema und auf die konkreten Datenwerte innerhalb des Gesamtsystems konsistent zu halten. Als Vergleichsbasis bietet sich das zentrale, konzeptionelle Modell auf Unternehmensebene an. Sowohl das Modell auf Fachbereichsebene als auch das Modell auf operativer Ebene sind dann als Teilsichten des zentralen Modells aufzufassen.[234] Einen mächtigen Abbildungs- und Auswertungsmechanismus vorausgesetzt, können dann sowohl die Datenmodelle als auch Datenwerte auf Interpretierbarkeit, Konsistenz und Vollständigkeit in bezug auf das zentrale, konzeptionelle Modell geprüft werden.[235] Wenngleich ein solcher Abbildungs- und Auswertungsmechanismus in praktischen Anwendungen bislang nicht zur Verfügung steht, ist die Betrachtungsweise zentral. Als ein Gegenstandsbereich wird das Data-Warehouse-System so ganzheitlich als Datenhaltungs- und Datenbereitstellungssystem für analytische Daten verstanden.

[234] Vgl. Jarke et al. (2000), S. 23.
[235] Vgl. Jarke et al. (2000), S. 23.

Kapitel 3

Datenqualität

3.1 Qualitätsbegriff und Qualitätssichten

Aufgrund der zentralen Bedeutung des Begriffs der „Qualität“ wäre ein einheitliches Begriffsverständnis wünschenswert. Allerdings ist dieser sehr komplex und schwer zu beschreiben. Als Folge der Entwicklung des Qualitätsbegriffs und der damit verbundenen Diskussion existieren bereits eine Vielzahl von Definitions- und Interpretationsversuchen.[236] Der folgende Abschnitt dient daher einer Beschreibung des allgemeinen Qualitätsbegriffs und der daraus abgeleiteten Qualitätssichten. Ziel ist es, das Qualitätsphänomen zu erfassen und die Komplexität zu reduzieren, um operationale Aussagen abzuleiten.

Bemühungen nationaler und internationaler Standardisierungsorganisationen haben zu einer breit akzeptierten Begriffsbeschreibung geführt. Diese ist beispielsweise in Form von DIN- oder ISO-Normen festgelegt und wird als

> „die Gesamtheit von Eigenschaften und Merkmalen eines Produktes oder einer Dienstleistung, die sich auf deren Eignung zur Erfüllung festgelegter oder vorausgesetzter Erfordernisse beziehen“[237]

beschrieben.

In der Literatur lassen sich nach dem Systematisierungsansatz von GARVIN fünf Qualitätssichten unterscheiden.[238] Der *transzendente Ansatz* kennzeichnet Qua-

[236] Vgl. Seghezzi (1996), S. 16; Rinne und Mittag (1995), S. 9; Garvin (1984), S. 26; zur Entwicklung des Qualitätsbegriffs vgl. z. B. Geiger (2001), S. 806ff.

[237] Vgl. DIN, Deutsches Institut für Normung e. V. (1995), S. 212.

[238] Vgl. im folgenden Garvin (1984), S. 25-28.

lität als angeborene Vortrefflichkeit, Einzigartigkeit oder Superlative. Qualität sei ein Synonym für hohe Standards und Ansprüche. Der transzendente Ansatz folgt dem eher abstrakt philosophischen Verständnis, dass Qualität nicht exakt definiert werden kann, sondern nur erfahrbar sei. Diese Qualitätsauffassung ist für die weitere Betrachtung unzulänglich und soll daher nicht weiter verfolgt werden. Bei der *produktbezogenen Qualitätsauffassung* bestimmen materielle Produkteigenschaften die Qualität eines Produktes. Qualität ist nach diesem Verständnis präzise messbar und eine inhärente Eigenschaft des Produktes selbst. Beim *anwenderbezogenen Ansatz* liegt die Auffassung vor, dass Qualität durch den Produktnutzer und weniger durch das Produkt selbst bestimmt wird. Ein Produkt wird dann als qualitativ hochstehend angesehen, wenn es dem Zweck der Benutzung durch den Kunden während des Gebrauchs dient. Die individuellen Bedürfnisse des Kunden sind dabei bestimmend. Die Einhaltung von Spezifikationen und die Abwesenheit von Fehlern stehen beim *herstellungsbezogenen Ansatz* im Mittelpunkt. Ziel ist die Einhaltung der Produktspezifikation durch kontrollierte Produktionsprozesse. Schlussendlich betrachtet der *wertbezogene Ansatz* Qualität unter Kostengesichtspunkten. Ein Produkt ist dann von hoher Qualität, wenn die Kosten und die empfangene Leistung in einem akzeptablen Verhältnis stehen.

Mit Ausnahme des transzendenten Begriffsverständnisses stellen die Ansätze unterschiedliche externe und interne Aspekte in den Vordergrund.[239] Jede dieser Sichtweisen besitzt für unterschiedliche Zwecke ihre Berechtigung. Der anwenderbezogene Ansatz bezieht sich auf eine externe Sicht und stellt den Endnutzer mit seinen Anforderungen in den Vordergrund. Von diesen Qualitätsforderungen wird eine Produktspezifikation und ein Produktionsplan abgeleitet. Die konzeptionelle Spezifikation bildet die Grundlage für die Gestaltung der Produktionsprozesse. Der wertorientierte Ansatz betrachtet Kostengesichtspunkte und lässt sich vertikal zu den drei Ebenen einordnen. Wenngleich Kostengesichtspunkte im Rahmen des Qualitätsmanagements eine wichtige Rolle spielen, wird die Betrachtung auf die drei Qualitätsebenen eingeschränkt:

- Die anwenderbezogene, externe Ebene.

[239] Vgl. Wolf (1999), S. 68.

- Die produktbezogene, konzeptionelle Ebene.
- Die herstellungsbezogene, prozessorientierte Ebene.

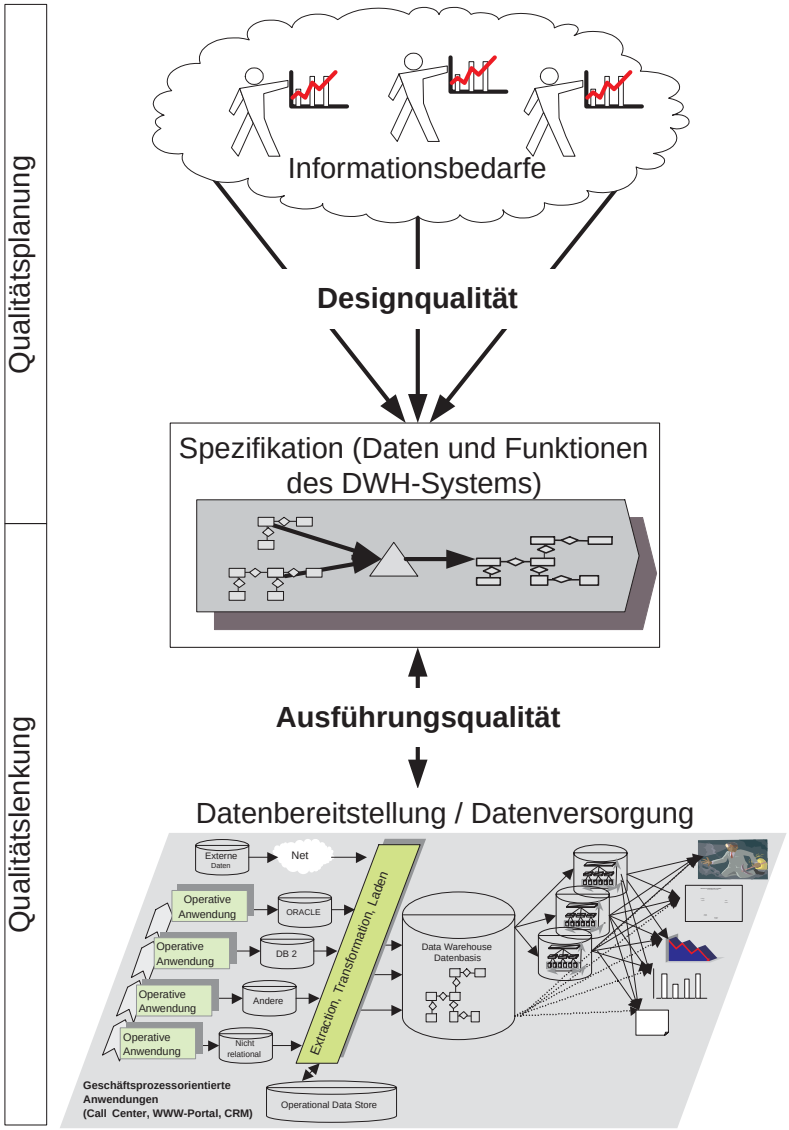


Abbildung 3.1: Qualitätssichten (Eigene Darstellung)

Auf Grundlage der Qualitätsebenen lässt sich Qualität, wie in Abbildung 3.1 dargestellt, grundsätzlich in zwei Faktoren untergliedern:[240]

[240] Vgl. Seghezzi (1996), S. 12 und S. 26; SEGHEZZI verwendet für den Begriff der Ausführungsqualität den an Dienstleistungen orientierten Begriff der Verrichtungsqualität. Mit dem Begriff der Ausführungsqualität soll hier die Ausführung der Erfassungs- und Transferprozesse zum Ausdruck gebracht werden.

- Designqualität.
- Ausführungsqualität.

Zunächst werden die Qualitätsforderungen der Endnutzer erfasst und durch eine Spezifikation konkretisiert. Es ist die Frage nach den geeigneten Produkteigenschaften zu beantworten. Es sind die Eigenschaften auszuwählen, welche die Bedürfnisse der Anwender am Besten erfüllen und so Kundenzufriedenheit erzeugen.[241] Spezifikationen sind beispielsweise technische Zeichnungen für die Produktion mechanischer Teile, Schaltpläne für die Montage elektrischer Komponenten oder auch Pflichten- und Lastenhefte für die Produktentwicklung und die Implementation von Softwarekomponenten. In einer Datenbank werden durch Datenschemata Entitäten und Eigenschaften der zu erfassenden Datenobjekte festgelegt und können so als Spezifikationen eingestuft werden.[242]

Sind die Anforderungen erfasst und in einer Spezifikation festgelegt, ändert sich die Zielsetzung des Qualitätsmanagements auf die Einhaltung der in der Spezifikation festgelegten Qualitätsforderungen. Nicht die Bedürfnisse der Anspruchsgruppen, sondern Konformität und fehlerfreie Erfüllung der in Spezifikationen niedergelegten Anforderungen ist das Ziel.[243] Die Produktionsprozesse sind dahingehend zu kontrollieren. *Designqualität* bezieht sich auf die Erfassung von Qualitätsforderungen aus Anwendersicht in eine Spezifikation, während *Ausführungsqualität* die Einhaltung der Spezifikation umfasst. Eine unzureichende Gesamtqualität kann sowohl in einer mangelhaften Design- als auch in einer nicht ausreichenden Ausführungsqualität begründet sein. Diese Unterscheidung ist zur Analyse und Identifikation von Ursachen mangelnder Qualität entscheidend.

In den folgenden Abschnitten soll, ausgehend von Ansätzen in der Literatur und der Übertragung der hier dargestellten Qualitätsauffassung, ein für die Arbeit zweckmässiges Begriffsverständnis entwickelt und empirisch reflektiert werden.

---

[241] Vgl. Juran (1999), S. 1.

[242] Vgl. auch Abschnitt 2.3.1; im Rahmen der Arbeit soll nicht zwischen dem Prozess der Spezifikation und dem Prozessergebnis der Spezifikationen unterschieden werden, wenn die Bedeutung aus dem jeweiligen Kontext hervorgeht.

[243] Vgl. Juran (1999), S. 2.

3.2 Ausgewählte Ansätze zum Begriff der Datenqualität

Ähnlich dem allgemeinen Qualitätsbegriff, wird Daten- und Informationsqualität in der Literatur vielfach beschrieben.[244] Allerdings hat sich bislang kein allgemeines Begriffsverständnis gebildet. Eine Auswahl von Ansätzen findet sich in den Tabellen 3.1, 3.2 und 3.3. Ergebnis dieser Arbeiten sind eine Vielzahl von Kriterienlisten und Einordnungsrahmen für unterschiedliche Anwendungsgebiete. Eine grobe Übereinstimmung lässt sich dahin erkennen, dass Qualität von Daten bzw. Informationen meist hinsichtlich des Beitrags zur Erreichung der Ziele des Datenempfängers bestimmt wird. Damit folgen die Ansätze weitgehend einer anwenderbezogenen Sicht. Der Begriff wird durch eine Fülle von Qualitätsmerkmalen konkretisiert, die in ihrer Bedeutung und Intensität erheblich vom Anwendungskontext abhängen. In den Ansätzen werden die Qualitätskriterien intuitiv anhand von Erfahrungen und Expertenwissen aufgestellt, auf Basis der in der Literatur genannten Kriterien zusammengestellt oder im Rahmen empirischer Untersuchungen erfasst. Allgemein werden Korrektheit, Vollständigkeit, Konsistenz und Aktualität genannt. Allerdings lässt sich auch hier keine Übereinstimmung bezüglich relevanter Qualitätskriterien, deren Begriffsbestimmung und Beziehungsstruktur erkennen.

Aus diesem Grund soll auf eine umfängliche Darstellung einzelner Merkmale der Datenqualität verzichtet und im folgenden lediglich einige ausgewählte Ansätze erläutert werden.[245] Diese bilden die Grundlage zur Aufstellung einer Liste von Datenqualitätskriterien für Data-Warehouse-Systeme.

Die Forschungsarbeiten können grob nach dem Forschungsansatz in theoriebasierte und auf Literatur bezogene Arbeiten, intuitive und empirische Ansätze ein-

[244] Eng mit dem Begriff der Datenqualität verbunden ist die Beurteilung der Anwenderzufriedenheit von Informationssystemen, die Qualität von Datenschemata sowie die Qualität von Anwendungssystemen und der Qualitätsbegriff in der Softwareentwicklung; vgl. DeLone und McLean (1992), S. 62 und S. 83; Wisom und Watson (2001), S. 19ff. Auf diese Ansätze soll im folgenden nicht direkt eingegangen werden. Zur Beurteilung der Anwenderzufriedenheit vgl. z. B. Ives, Olsen und Baroudi (1983), S. 785ff. und DeLone und McLean (1992), S. 65ff. sowie die in Wang, Storey und Firth (1995b) angegebene Literatur. Die Qualität von Datenschemata wird z. B. in Schelp (2000), S. 48ff. betrachtet. Zur Qualität von Softwaresystemen und der Softwareentwicklung vgl. z. B. Wallmüller (1990), S. 7; Schmitz (1990), S. 309ff.; Gillies (1992), S. 7.

[245] Eine Übersicht über verschiedene Ansätze findet sich beispielsweise in Wang et al. (1995b) oder Eppler und Wittig (2000).

Autor(en)	Anwendungskontext	Definitionsansatz	Ableitung und Strukturierung der Datenqualitätskriterien
WANG et al.[a]	Operative Datenhaltungssysteme, Wissensmanagement	Fitness for use by information consumers	Empirische Erfassung genereller Merkmale und Strukturierung in vier Kategorien: Innere Datenqualität, Kontextabhängige Datenqualität, Darstellung, Zugriff.
MÜLLER[b]	Data-Warehouse-Systeme	Eignung für Aufgabenträger in Entscheidungsprozessen	Semiotik als Strukturierungshilfe.
HOLTHUIS[c]	Data-Warehouse-Systeme	Nutzen von Daten	Auflistung von häufig in der Literatur genannten Merkmalen.
NAUMANN und ROLKER[d]	Internet	Anwenderorientiert	Literaturüberblick; Merkmale basieren weitgehend auf den von WANG et al. erfassten Kriterien. Untergliederung in anwenderorientierte, prozessorientierte und objektive Merkmale.
WEIKUM[e]	Informationsanbieter	–	Auflistung von systemorientierten, prozessorientierten und informationsorientierten Merkmalen.

[a] Vgl. Wang, Ziad und Lee (2001), S. 2; Wang und Strong (1996), S. 18-21; Wand und Wang (1996), S. 92-94; Wang et al. (1995b), S. 632.
[b] Vgl. Müller (2000), S. 14f.
[c] Vgl. Holthuis (1999), S. 33f.
[d] Vgl. Naumann und Rolker (1999), S. 102-106; Naumann und Rolker (2000), S. 152f.
[e] Vgl. Weikum (1999), S. 383ff.

Tabelle 3.1: Ausgewählte Ansätze im Bereich Datenqualität (Teil 1)

Autor(en)	Anwendungskontext	Definitionsansatz	Ableitung und Strukturierung der Datenqualitätskriterien
WOLF[a]	Informationslogistik	Anwenderorientiert, Zweckorientiert	Auflistung von Merkmalen nach Entscheidungsrelevanz, Inhalt, Zeit, Ort, Menge, Form und Wirtschaftlichkeit.
ENGLISH[b]	Data-Warehouse-Systeme, Informationssysteme	Result in user statisfaction; meeting expectations	Auflistung von Merkmalen der Kategorien Datendefinitions- und Architekturqualität, Qualität des Dateninhalts sowie der Datenpräsentation.
JARKE et al.[c]	Data-Warehouse-Systeme	Anwenderorientiert	Weitgehend auf den von Wang et al. erfassten Merkmalen basierend. Orientierung am Entwicklungsprozess für Data-Warehouse-Systeme.
WAND und WANG[d]	Informationssysteme	Abbildungsgüte zwischen realer Welt und der Repräsentation im Anwendungssystem	Eine qualitativ hochwertige Systemabbildung (Anwendungssystem ↔ reale Welt) ist vollständig, eindeutig, bedeutungsvoll und korrekt.
MILLER[e]	Informationssysteme	Anwenderorientiert	Auflistung von Qualitätsmerkmalen (Korrektheit, Aktualität, Vollständigkeit, Widerspruchsfreiheit, Format, Zugriffsfähigkeit, Kompatibilität, Sicherheit, Validität).

[a] Vgl. Wolf (1999), S. 93.
[b] Vgl. English (1999), S. 27-31 u. 83-118.
[c] Vgl. Jarke und Vassiliou (1997), S. 302f.; Jarke, Jeusfeld, Quix und Vassiliadis (1999), S. 238-242; Jarke et al. (2000), S. 137-140.
[d] Vgl. Wand und Wang (1996), S. 93.
[e] Vgl. Miller (1996), S.79-81.

Tabelle 3.2: Ausgewählte Ansätze im Bereich Datenqualität (Teil 2)

Autor(en)	Anwendungskontext	Definitionsansatz	Ableitung und Strukturierung der Datenqualitätskriterien
REDMAN et al.[a]	Informationssysteme	Datenschema; allgemeine Qualitätsmerkmale	Auflistung von Merkmalen bezogen auf die konzeptionelle Sicht, die Datenwerte und die Datenpräsentation.
LAUDON[b]	Kriminalstatistik	Anwenderorientiert	Erfassung von Merkmalen durch Interviews (Vollständigkeit, Korrektheit, Eindeutigkeit).
BALLOU und PAZER[c]	Datentransformations-prozesse in Informationssystemen	Anforderungen an Informationen für Entscheidungsprozesse	Auflistung von vier Merkmalen (Korrektheit, Aktualität, Vollständigkeit, Konsistenz).
HAUKE[d]	Informationsverarbei-tungsprozesse	Qualitativer „Nutzwert“ von Informationen	Auflistung von in der Literatur genannten Merkmalen. Diese werden in Raum- und Zeitkriterien, sach- und problembezogene Kriterien sowie personenbezogene Kriterien unterteilt.
MOREY[e]	Management-Informationssysteme	–	Bezieht sich ausschliesslich auf Korrektheit der Datenwerte.
GROTZ-MARTIN[f]	Entscheidungsprozesse	Pragmatischer Informations- oder Wirkungsgehalt	Auflistung von in der Literatur genannten Merkmalen und Strukturierung dieser in Relevanz, Rechtzeitigkeit, Aktualität, Zuverlässigkeit und Präzision.

[a] Vgl. Redman (1996), S. 245-267; Levitin und Redman (1995), S. 82-86; Fox, Levitin und Redman (1994), S. 13-17.
[b] Vgl. Laudon (1986), S. 6.
[c] Vgl. Ballou und Pazer (1985), S. 153.
[d] Vgl. Hauke (1984), S. 155-159.
[e] Vgl. Morey (1982), S. 338.
[f] Vgl. Grotz-Martin (1976), S. 15.

Tabelle 3.3: Ausgewählte Ansätze im Bereich Datenqualität (Teil 3)

geteilt werden.[246] Zunächst werden die von WAND und WANG vorgeschlagenen inneren Datenqualitätsmerkmale erläutert. Sie leiten die Qualitätsmerkmale im Sinne einer Abbildungsgüte zwischen Realwelt und Informationssystem ab. REDMAN et al. stellen eine Merkmalsliste auf und untergliedern diese in Merkmale des Datenmodells (konzeptionelle Sicht), der Datenwerte und der internen Darstellung. ENGLISH listet zahlreiche Qualitätsmerkmale auf und gliedert diese in Merkmale der Datendefinitions- und Architekturqualität, der Qualität des Dateninhalts sowie der Qualität der Datenpräsentation. Eine empirische Untersuchung wurde von WANG und STRONG durchgeführt und dient als Grundlage zahlreicher weiterer Forschungsarbeiten.[247] Abschliessend werden die von JARKE et al. vorgeschlagenen Datenqualitätsfaktoren für Data-Warehouse-Systeme erläutert.

3.2.1 Innere Datenqualität nach WAND und WANG

Ein auf die Entwicklung und den Betrieb eines Informationssystems fokussierter Ansatz wird von WAND und WANG vorgeschlagen.[248] Die Betrachtung lässt die externe Ebene, die den geforderten Zweck und die funktionalen Anforderungen an das Informationssystem vom Endnutzer zum Gegenstand hat, unberücksichtigt. Der Ansatz konzentriert sich auf die konzeptionelle und interne Ebene eines Informationssystems. Dem Ansatz liegt die Annahme zugrunde, dass ein Systemnutzer das Informationssystem mit der realen Welt vergleicht. Inkonsistenzen zwischen dem Informationssystem und der realen Welt führen zu Datenqualitätsmängeln (vgl. Abbildung 3.2). Eine Qualitätsbetrachtung findet dann im Sinne einer Abbildungsübereinstimmung zwischen Realwelt und Informationssystem statt, wobei

(a) jedem Zustand der realen Welt (RW-Zustand) mindestens ein Zustand im Informationssystem (IS-Zustand) gegenübersteht und

(b) jeder Zustand im Informationssystem auf den „richtigen“ Zustand der realen Welt zurückgeführt werden kann.

[246] Vgl. Wang und Strong (1996), S. 7; Naumann und Rolker (1999), S. 100.

[247] Vgl. z. B. Naumann und Rolker (1999), S. 100; Jarke und Vassiliou (1997), S. 302 f.; Kahn, Strong und Wang (1997), S. 87ff.

[248] Vgl. hierzu Wand und Wang (1996).

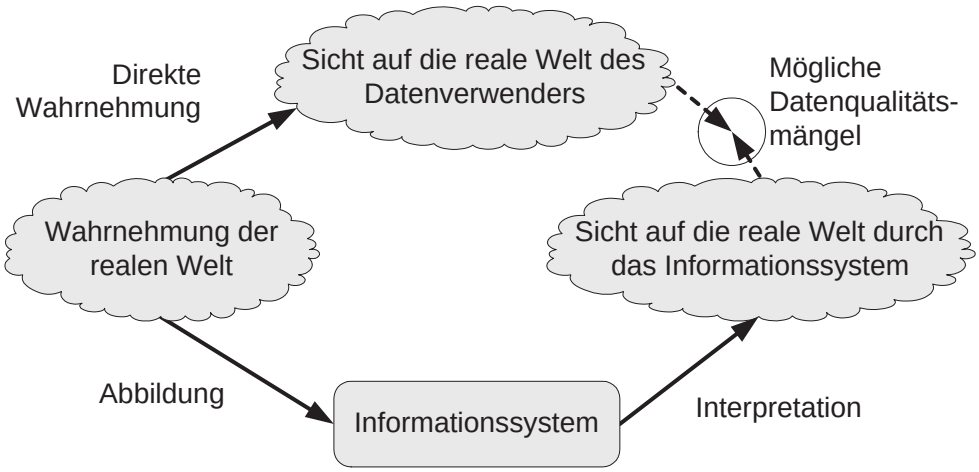


Abbildung 3.2: Mögliche Datenqualitätsmängel nach WAND und WANG (In Anlehnung an Wand und Wang (1996), S. 88)

Untersucht man mögliche Abbildungsmängel, können vier Fehlertypen identifiziert werden. Aus diesen werden innere Datenqualitätsmerkmale abgeleitet. Diese sind in Tabelle 3.4 zusammengefasst.

Merkmal	**Fehlerart**	**Ursache**
Vollständig	Ungenaue Abbildung: Fehlender Zustand im IS	Entwurf
Eindeutig	Ungenaue Abbildung: Mehrere RW-Zustände beziehen sich auf den gleichen IS-Zustand	Entwurf
Bedeutungsvoll	IS-Zustände ohne Bedeutung in der RW; Veränderung der Zustände (Abbildung auf einen IS-Zustand ohne Bedeutung in der RW)	Entwurf; Betrieb
Korrekt	Ein Zustand wird auf den falschen Zustand abgebildet	Betrieb

Tabelle 3.4: Innere Datenqualitätsmerkmale nach WAND und WANG (In Anlehnung an Wand und Wang (1996), S. 92)

3.2.2 Ansatz von REDMAN et al.

Im folgenden werden die von REDMAN et al. vorgeschlagenen Qualitätsmerkmale dargestellt. Grundlage bildet die Unterscheidung zwischen der Datendefini-

tion und den konkreten Datenwerten. Hierzu definieren sie zunächst ein Datum $< e, a, v >$ als ein Wert v aus dem Wertebereich eines Attributs a einer Entität e.[249] Sie folgen damit dem in Abschnitt 2.3.1 erläuterten Schemakonzept von Datenmodellen. Aufbauend auf diesem Konzept listen sie Qualitätskriterien für die konzeptionelle Sicht, die Dateninhalte und die Repräsentation der Daten auf. Die Qualitätskriterien sind zusammenfassend in Tabelle 3.5 dargestellt. Die konzeptionelle Sicht bezieht sich auf die Qualitätsanforderungen der Entitäten und Attribute im Datenschema. Ist das Datenschema aufgestellt, sollten die Dateninhalte korrekt, vollständig, aktuell und konsistent sein. Eine weitere Gruppe von Datenqualitätskriterien bezieht sich auf die Repräsentation der Daten durch Formate und deren physikalische Speicherung.

Konzeptionelle Sicht	Inhalt	*Relevanz, Zugriff, Klarheit der Definitionen*
	Umfang	*Vollständigkeit, Wesentlich*
	Detaillierungsgrad	*Attributgranularität, Domaingenauigkeit*
	Struktur	*Natürlich, Identifizierbar, Homogen, Minimum an Redundanz*
	Konsistenz	*Semantische Konsistenz, Strukturelle Konsistenz*
	Reaktionen auf Veränderungen	*Stabil, Flexibel*
Dateninhalte	*Korrekt, Vollständig (Entitäten und Attribute), Konsistenz, Aktualität*	
Daten-repräsentation	Formate	*Angemessen, Formatgenauigkeit, Effiziente Speichernutzung, Interpretierbarkeit, Formatflexibilität, Übertragbarkeit, Fähigkeit Nullwerte abzubilden*
	Physikalische Speicherung	*Darstellungskonsistenz*

Tabelle 3.5: Datenqualitätsmerkmale nach REDMAN et al. (In Anlehnung an Redman (1996), S. 267)

[249] Vgl. Redman (1996), S. 245-267; Levitin und Redman (1995), S. 82-86; Fox et al. (1994), S. 13-17.

3.2.3 Qualitätsmerkmale nach ENGLISH

ENGLISH listet eine Vielzahl von Qualitätsmerkmalen auf, deren Überschneidungen und Beziehungen jedoch kaum dargestellt werden. Grundsätzlich unterscheidet er zwischen der Datendefinitions- und Architekturqualität, der Qualität der Datenwerte sowie der Qualität der Datenpräsentation.[250] Im Sinne einer Spezifikation eines „Datenproduktes“ listet er für die *Datendefinitions- und Architekturqualität* zahlreiche Qualitätskriterien auf und gruppiert diese in:[251]

- Die Qualität der Datenstandards (Richtlinien, die eine konsistente, genaue, klare und verständliche Datendefinition unterstützen).
- Die Qualität der Datendefinitionen (Semantische Aspekte und Geschäftsregeln). Er nennt hier Merkmale wie bedeutungsvoll und verständlich.
- Die Qualität der Informationssystemarchitektur (Allgemeiner Entwurf der Datenmodelle und Datenbanken in bezug auf Wiederverwendung, Stabilität und Flexibilität).

Die *Qualität der Datenwerte* unterteilt er in innere und pragmatische Datenqualitätskriterien.[252] Diese sind zusammenfassend in Tabelle 3.6 aufgelistet. Während innere Datenqualitätskriterien weitgehend von der Datenverwendung unabhängig und allgemeingültig sind, hängen pragmatische Merkmale von der jeweiligen Anwendungssituation ab. Die *Qualität der Datenpräsentation* bezieht sich auf die Datenbereitstellung und die Datendarstellung. Er nennt hier Kriterien wie Zugriffsfähigkeit, Rechtzeitigkeit und Interpretierbarkeit.[253]

3.2.4 Empirische Untersuchung von WANG und STRONG

Grundlage zahlreicher weiterer Arbeiten bildet die von WANG und STRONG durchgeführte empirische Untersuchung.[254] Sie stellt eine umfangreiche empiri-

[250] Vgl. English (1999), S. 27-30.
[251] Vgl. English (1999), S. 87ff.
[252] Vgl. English (1999), S. 142ff.
[253] Vgl. English (1999), S. 29.
[254] Vgl. im folgenden Wang und Strong (1996).

Innere Datenqualitätskriterien
Übereinstimmung mit der Datendefinition
Vollständigkeit der Datenwerte
Plausibel und mit den Geschäftsregeln übereinstimmend
Genauigkeit zu Referenzdaten (Vergleichsdaten)
Genauigkeit zur Realität
Granularität der Datenwerte
Eindeutigkeit (Keine Duplikate)
Übereinstimmung von redundanten und verteilten Datenbeständen
Zeitliche Übereinstimmung von redundanten und verteilten Datenbeständen
Allgemeine Zugriffsfähigkeit
Pragmatische Datenqualitätskriterien
Aktualität und Pünktlichkeit
Klarheit / Interpretierbarkeit für den Datenanwender
Übereinstimmung zwischen abgeleiteten (berechneten) Daten zu den Ursprungsdaten
Nützlichkeit
Vollständigkeit bezogen auf den Informationsbedarf

Tabelle 3.6: Qualitätsmerkmale von Datenwerten nach ENGLISH (In Anlehnung an English (1999), S. 142f.)

sche Erhebung zur Bestimmung allgemeiner Datenqualitätsmerkmale dar.[255] Die Untersuchung fand in zwei Stufen statt:

- Zunächst wurden in einem ersten Schritt 179 Qualitätsmerkmale durch eine schriftliche Befragung unter 25 Endanwendern aus der Industrie und 112 Studenten erfasst. Dabei sollten die Teilnehmer in einer offenen Frage Datenqualitätsmerkmale nennen.[256] In einer zweiten Frage sollten zu einer Kriterienliste weitere Datenqualitätsmerkmale hinzugefügt werden. Ergebnis dieses ersten Schrittes ist eine Liste von Datenqualitätskriterien.

- Ausgehend von diesen Kriterien wurde eine zweite Untersuchung mit 1500 Personen durchgeführt. Die Teilnehmer sollten in einer schriftlichen Befragung die Relevanz einzelner Merkmale auf einer Skala von eins bis neun

[255] Vgl. Naumann und Rolker (1999), S. 102; Wang und Strong (1996), S. 7.
[256] Eine Liste der genannten Kriterien findet sich in Wang und Strong (1996), S. 29-31.

angeben. Eine nähere Beschreibung oder Erläuterung der Merkmale wurde nicht zur Verfügung gestellt.

Mit einer Rücklaufquote von 24% lagen 355 Fragebögen zur Auswertung vor. Als wichtige Datenqualitätsmerkmale wurden Genauigkeit (accuracy) und Korrektheit (correctness) genannt. Mit Hilfe einer Datenanalyse konnten dann 20 Merkmalsgruppen identifiziert und in einer weiteren Untersuchung mit 30 Teilnehmern in vier Kategorien untergliedert werden. Ergebnis dieser Untersuchung sind vier Kategorien mit zugeordneten Qualitätsmerkmalen, welche in Tabelle 3.7 zusammenfassend dargestellt werden.

Kategorie	**Datenqualitätsmerkmale**
Innere Datenqualität	Glaubwürdigkeit, Genauigkeit, Objektivität, Vertrauenswürdigkeit
Kontextabhängige Datenqualität	Zusatznutzen, Relevanz, Aktualität, Vollständigkeit, Angemessenes Datenvolumen
Darstellungsqualität	Interpretierbarkeit, Verständlichkeit, Konsistente Darstellung, Knappe Darstellung
Zugangsqualität	Zugriffsmöglichkeit, Zugriffssicherheit

Tabelle 3.7: Datenqualitätsmerkmale nach WANG und STRONG (In Anlehnung an Wang und Strong (1996), S. 20)

3.2.5 Qualitätsfaktoren für Data-Warehouse-Systeme nach JARKE et al.

JARKE et al. schlagen eine prozessorientierte Gliederung von Datenqualitätsmerkmalen für Data-Warehouse-Systeme vor.[257] Dabei werden aggregierte und historisierte Daten sowie Zugriffskonzepte besonders berücksichtigt.[258] Ausgehend von idealtypischen Rollen in einem Data-Warehouse-System werden die Prozesse

- der Entwicklung und Verwaltung,

[257] Vgl. im folgenden Jarke und Vassiliou (1997), S. 303; Jarke et al. (1999), S. 238-242.
[258] Vgl. Jarke und Vassiliou (1997), S. 302.

- der Softwareimplementierung sowie
- der Datennutzung

identifiziert. Die sich hieraus ergebende Liste von Qualitätsmerkmalen ist in Tabelle 3.8 dargestellt. Aufgrund der zentralen Bedeutung der gespeicherten Datenwerte werden die Merkmale der Datenqualität weiter konkretisiert. Im Gegensatz zu den prozessorientierten Qualitätsmerkmalen beziehen sich diese direkt auf den Zustand der Datenbestände. Es werden Vollständigkeit, Glaubwürdigkeit, Korrektheit, Konsistenz und die Möglichkeit der Dateninterpretation genannt.

Enwicklung und Verwaltung	Datenschema und Datenqualität	*Korrektheit, Vollständigkeit, Minimalität, Interpretierbarkeit, Datenrückverfolgung*
	Metadaten-Evolution	
Softwareimplementation		*Funktionalität, Zuverlässigkeit, Nützlichkeit, Software-Effizienz, Wartbarkeit, Portabilität*
Datennutzung	Zugriffsfähigkeit	*Systemverfügbarkeit, Transaktionsverfügbarkeit, Datensicherheit*
	Nützlichkeit	*Interpretierbarkeit, Antwortverhalten, Aktualität*

Tabelle 3.8: Qualitätsfaktoren nach JARKE et al. (Vgl. hierzu Jarke et al. (1999), S. 239-241)

3.3 Datenqualität in Data-Warehouse-Systemen

Die Vielfalt der oben dargestellten Ansätze zeigt die Vielschichtigkeit des Begriffs der Datenqualität. Im Rahmen dieser Arbeit soll daher nicht der Versuch unternommen werden, eine allgemeingültige Definition oder einen umfassenden Kriterienkatalog zu entwickeln. Es sollen vielmehr für praktische Problemstellungen einige relevante Qualitätskriterien identifiziert und soweit möglich nach den Qualitätssichten der Designqualität und Ausführungsqualität untergliedert werden. Aus der Vielzahl der genannten Datenqualitätsmerkmalen sollen diejenigen

ausgewählt und strukturiert werden, die sich für die Beschreibung der Datenqualität in Data-Warehouse-Systemen besonders eignen. Die Eignung bezieht sich dabei auf die Erfassung der subjektiven Qualitätsforderungen an die Daten sowie deren Spezifikation und Messung in einem Data-Warehouse-System. Anhand einer empirischen Untersuchung wird die Datenqualität in Data-Warehouse-Systemen erörtert sowie die Relevanz der Datenqualitätsmerkmale ermittelt.

Im Bereich der empirischen Sozialforschung ist der Aufbau eines problemadäquaten Untersuchungskonzeptes für die erfolgreiche Durchführung von Analysen von entscheidender Bedeutung. In dieser frühen Phase sind grundlegende Entscheidungen zur Forschungsmethodik, Stichprobe, Auswertung und nachgelagerten Ergebnisverwertung zu treffen. Diese werden durch das Forschungsvorhaben und den angestrebten Lösungsbeitrag determiniert.[259] Im folgenden wird daher zunächst der theoretische Bezugsrahmen und die zugrunde gelegte Untersuchungskonzeption beschrieben. In Abschnitt 3.3.2 werden dann anschliessend die Untersuchungsergebnisse analysiert und beschrieben.

3.3.1 Theoretischer Bezugsrahmen und Untersuchungskonzeption

Ein theoretischer Bezugsrahmen kann als Fundament empirischer Studien angesehen werden. Er besitzt die Aufgabe, das Vorwissen über das Forschungsobjekt zu strukturieren, relevante Variablengruppen abzugrenzen und vermutete Wirkungszusammenhänge darzulegen.[260] Für die durchgeführte empirische Untersuchung sind hier, neben den in Abschnitt 3.2 dargestellten Qualitätskriterien, für eine Einordnung der Ergebnisse übliche Anspruchsgruppen sowie typische Aufgaben eines Data-Warehouse-Systems zu identifizieren.

Wie oben gezeigt, existieren zahlreiche Ansätze zur Konkretisierung des Begriffs der Datenqualität.[261] Bislang lassen sich allerdings kaum Arbeiten zur empirischen Erfassung von Datenqualitätsforderungen und den Eigenschaften qualitativ

[259] Vgl. Atteslander (1995), S. 30ff.
[260] Vgl. Lamnek (1995), S. 113.
[261] Vgl. Abschnitt 3.2.

hochwertiger Daten in der Literatur finden. Meist werden häufig aufgeführte Qualitätskriterien genannt und für die jeweilige Arbeit strukturiert. Aufgrund dieses Defizits bezweckt die durchgeführte empirische Untersuchung daher, eine Basis für weitere Arbeiten zu schaffen. Der Anlass für die vorliegende empirische Untersuchung ist demnach, einen Beitrag zu diesem bislang nicht ausreichend erforschten Themenkomplex zu leisten. Sie ist, aufgrund mangelnder theoretischer Erklärungsmodelle und des relativ gering gesicherten Vorwissens im Forschungsbereich der qualitativen Sozialforschung zuzuordnen.[262] Mit ihrer Hilfe sollen Problembereiche der Datenqualität erfasst, Möglichkeiten zur Sicherstellung der Datenqualität ermittelt und die verschiedenen Datenqualitätsforderungen in Data-Warehouse-Systemen untersucht werden.

Aufgrund der bisherigen Erfahrungen in der Projektarbeit und dem Kompetenzzentrum kann davon ausgegangen werden, dass die Qualitätsforderungen von

- der jeweiligen Anspruchsgruppe innerhalb des Data-Warehouse-Systems und
- dem Aufgabenfokus des Data-Warehouse-(Teil)Systems

abhängen. Daher sind nach Möglichkeit verschiedene Anspruchsgruppen und Aufgabenbereiche von Data-Warehouse-Systemen in die Untersuchung einzubeziehen. Ausgangspunkt der Qualitätsbetrachtung ist der anwenderbezogene Qualitätsbegriff. Dieser bildet die Grundlage für die Qualitätssichtweise der Designqualität und führt zu einer Qualitätsspezifikation. Die Qualitätsspezifikation bildet ihrerseits die Grundlage für die Ausführungsqualität. Als wichtige Anspruchsgruppe für Qualitätsforderungen kann daher zunächst der Datenempfänger genannt werden. Neben diesem lassen sich noch weitere Anspruchsgruppen innerhalb eines Data-Warehouse-Systems finden. In Abschnitt 2.4.2.2 werden die Aufgabenträger anhand ihrer fachlichen Ausrichtung und bezüglich einzelner Systemkomponenten eingeordnet. Da insbesondere alle an der Entwicklung und dem Betrieb eines Data-Warehouse-Systems beteiligten Personengruppen sich hier wiederfinden, soll dieser Einordnungsrahmen zur Strukturierung der Anspruchsgruppen verwendet werden.

[262] Vgl. Lamnek (1995), S. 35ff.

Die Einsatzmöglichkeiten der Data-Warehouse-Systeme in der Praxis sind inzwischen sehr zahlreich.[263] Einer Untersuchung zufolge finden sich Einsatzgebiete von Data-Warehouse-Anwendungen in den Berichts- und Kontrollsystemen (Controlling), bei der Unterstützung der Geschäftsführung, im Marketing sowie im operativen Vertrieb.[264] Vor diesem Hintergrund erscheint es sinnvoll, die in der Praxis unter dem Begriff Data-Warehouse-Systeme vorherrschenden Informationssysteme anhand einer Strukturierungshilfe grob einzuordnen. In Abschnitt 2.4.1 führt die Untergliederung der Informationssysteme nach betriebswirtschaftlichen Grundfunktionen sowie nach der Art der informationsverarbeitenden Aufgabe zur Pyramide der Informationssysteme. Wenngleich diese Unterteilung lediglich als Strukturierungshilfe verstanden werden sollte, erscheint sie dennoch geeignet, die in der Praxis vorkommenden Informationssysteme einzuordnen.

Die Datenqualitätsmerkmale bedürfen ebenfalls einer Strukturierung. Eine erste Untergliederung bietet sich durch die Qualitätssichten in Design- und Ausführungsqualität an, wodurch sich insbesondere die Trennung von Datenschema und den Datenwerten auf Instanzebene ergibt.[265] Dieser Unterteilung folgen REDMAN et al., ENGLISH und JARKE et al., während bei WANG et al. dagegen der anwenderbezogene Datenqualitätsbegriff im Vordergrund steht. Durch eine Untersuchung der oben genannten Ansätze und weiterer Diskussionen im Rahmen der Projektarbeit mit verschiedenen Anwendergruppen wurde eine Kriterienliste erarbeitet. Interpretierbarkeit und Nützlichkeit ergaben sich als wichtige Kriterien für die Datenqualität eines Datenschemas. Als Qualitätsmerkmale für Datenwerte ergaben sich insbesondere Glaubwürdigkeit, zeitlicher Bezug, Nützlichkeit und Verfügbarkeit. Die erarbeitete Liste von Datenqualitätsmerkmalen für Datenschemata und Datenwerte ist in Tabelle 3.9 und Tabelle 3.10 zusammengestellt.

Auf Grundlage dieser Arbeit wurde ein neun Fragen umfassender Fragebogen erstellt.[266] Im Jahr 2001 wurde dieser an 110 Personen elektronisch per Email versendet. Die Personen wurden aus der Gruppe von Interessenten des Kompetenzzentrums „Data Warehousing 2“ ausgewählt. Auswahlkriterium bildete die

[263] Vgl. Holten, Knackstedt und Becker (2001a), S. 42.
[264] Vgl. Schelp (2000), S. 134.
[265] Vgl. Abschnitt 3.1.
[266] Dieser findet sich im Anhang A.1.

Kategorie	Merkmal	Beschreibung
Interpretierbarkeit	Semantik	Die Entitäten, Beziehungen und Attribute und deren Wertebereiche sind einheitlich, klar und genau beschrieben sowie dokumentiert.
	Identifizierbarkeit	Einzelne Informationsobjekte (z. B. Kunden) können eindeutig identifiziert werden.
	Synonyme	Beziehungen zwischen Synonymen sind bekannt und dokumentiert.
	Zeitlicher Bezug	Der zeitliche Bezug einzelner Informationsobjekte ist abgebildet.
	Repräsentation fehlender Werte	Fehlende Werte (Nullwerte / Default-Werte) sind definiert und können abgebildet werden.
Nützlichkeit (Zweckbezogen)	Vollständigkeit	Alle wesentlichen Entitäten, Beziehungen und Attribute sind erfasst. Die Daten ermöglichen die Erfüllung der Aufgabe.
	Erforderlichkeit	Definition von Pflicht- und Kannfeldern.
	Granularität	Die Entitäten, Beziehungen und Attribute sind im notwendigen Detaillierungsgrad erfasst.
	Präzision der Wertebereichsdefinition	Die Definition der Wertebereiche repräsentiert die möglichen und sinnvollen Datenwerte.

Tabelle 3.9: Qualitätsmerkmale bezogen auf das Datenschema

Firmengrösse, wobei mittelständische und kleine Unternehmen aufgrund der anzunehmenden geringen Komplexität nicht berücksichtigt wurden. Von den 110 angeschriebenen Personen antworteten 25 (Rücklaufquote 23%), so dass diese Fragebögen der Auswertung zur Verfügung standen.[267] Im Anhang A.2 ist die absolute und relative Verteilung der angeschriebenen Personen und der Antworten nach Branchen und Länder dargestellt.[268]

3.3.2 Ergebnis und Fazit der empirischen Untersuchung

Die ausgefüllten Fragebögen zeigen ein breites Spektrum an unterschiedlichen Data-Warehouse-Systemen auf. So finden sich in den Rückmeldungen langfristige Planungs- und Entscheidungssysteme, Analysesysteme, Berichts- und Kontrollsysteme, Systeme zur Unterstützung wertorientierter als auch mengenorien-

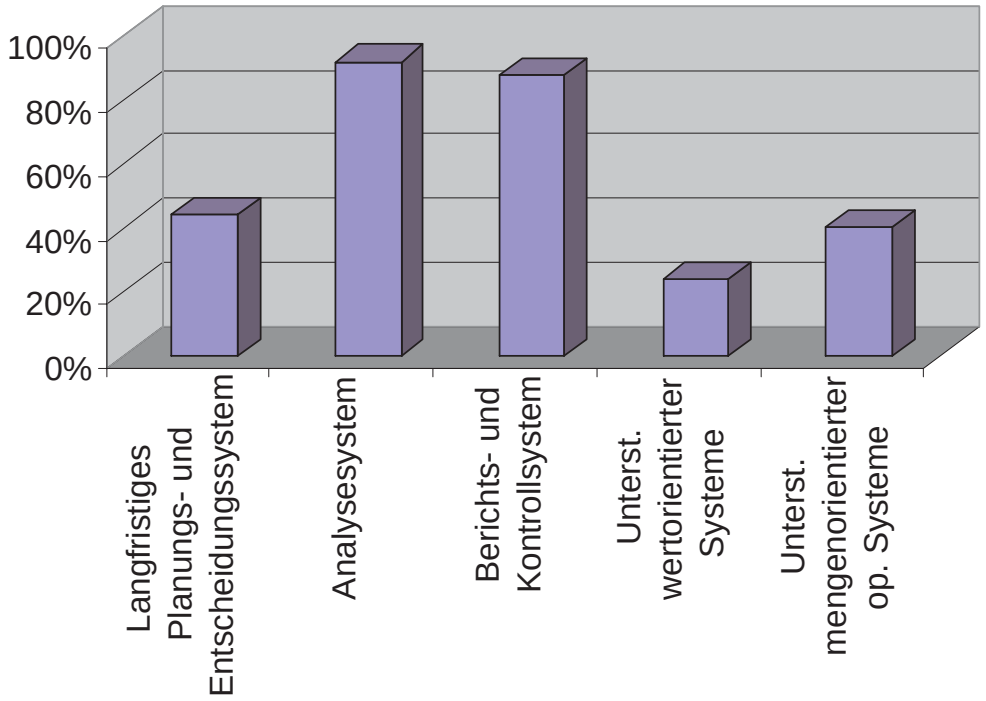
[267] Fünf Unternehmen hatten zur Zeit kein Data-Warehouse-System im Einsatz und haben daher den Fragebogen nicht beantwortet.

[268] Die Branchenzuordnung wurde in Anlehnung an die Klassifikation der Wirtschaftszweige vorgenommen; vgl. hierzu Statistisches Bundesamt (1999).

Kategorie	Merkmal	Beschreibung
Glaub-würdigkeit	Korrektheit	Die Daten stimmen inhaltlich mit der Datendefinition überein und sind empirisch korrekt.
	Datenherkunft	Die Datenherkunft und die vorgenommenen Datentransformationen sind bekannt.
	Vollständigkeit	Alle Daten sind gemäss Datenmodell erfasst.
	Widerspruchs-freiheit	Die Daten weisen keine Widersprüche zu Integritätsbedingungen (Geschäftsregeln, Erfahrungswerte) und Wertebereichsdefinitionen auf (innerhalb des Datenbestands, zu anderen Datenbeständen, im Zeitverlauf)
	Syntaktische Korrektheit	Die Daten stimmen mit der spezifizierten Syntax (Format) überein.
	Zuverlässigkeit	Die Glaubwürdigkeit der Daten ist konstant.
Zeitlicher Bezug	Aktualität	Datenwerte bezogen auf den gegenwärtigen Zeitpunkt sind erfasst.
	Zeitliche Konsistenz	Alle Datenwerte bzgl. eines Zeitpunktes sind gleichermassen aktuell.
	Nicht-Volatilität	Die Datenwerte sind permanent und können zu einem späteren Zeitpunkt wieder aufgerufen werden.
Nütz-lichkeit	Relevanz	Die Datenwerte können auf einen relevanten Datenausschnitt beschränkt werden.
	Zeitlicher Bezug	Die Datenwerte beziehen sich auf den benötigten Zeitraum.
Verfüg-barkeit	Zeitliche Verfügbarkeit	Die Daten stehen rechtzeitig zur Verfügung.
	System-verfügbarkeit	Das Gesamtsystem ist verfügbar.
	Transaktions-verfügbarkeit	Einzelne benötigte Transaktionen sind ausführbar, die Zugriffszeit ist akzeptabel und gleichbleibend.
	Zugriffsrechte	Die benötigten Zugriffsrechte sind ausreichend.

Tabelle 3.10: Qualitätsmerkmale bezogen auf die Datenwerte

tierter operativer Systeme (siehe Abbildung 3.3). Mit über 80% ist dabei eine Konzentration auf Analysesysteme, Berichts- und Kontrollsysteme zu verzeichnen. Allerdings traten häufig Mehrfachnennungen auf, so dass eine eindeutige Zuordnung zu den Typen nicht möglich ist.

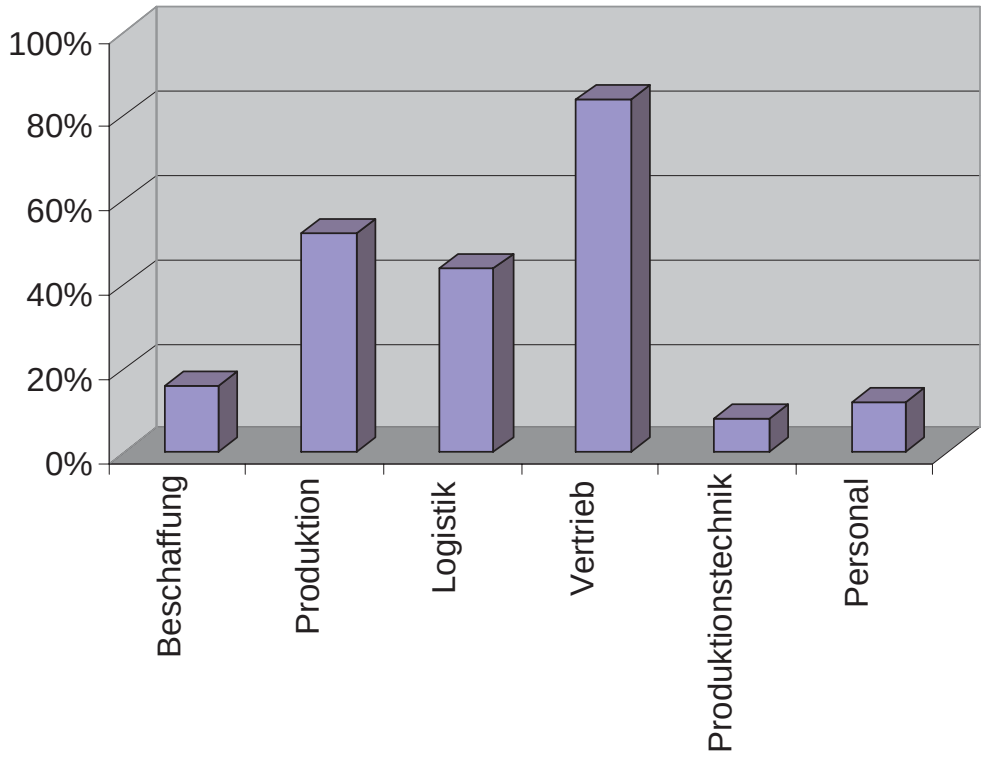


Angaben in % der ausgefüllten Fragebögen.

Abbildung 3.3: Zweck von Data-Warehouse-Systemen (Eigene Darstellung)

Werden die ausgefüllten Fragebögen nach den unterstützten betriebswirtschaftlichen Funktionen untersucht, stellt sich die Auswertung folgendermassen dar (siehe Abbildung 3.4): Auffallend ist die hohe Konzentration auf die Vertriebsunterstützung. Logistik und Produktionsabläufe werden ebenfalls durch Data-Warehouse-Systeme unterstützt. Die relativ geringe Unterstützung der Beschaffung und der Produktionstechnik lässt sich teilweise auf den hohen Anteil von Unternehmen des Kredit- und Versicherungsgewerbes zurückführen.

Bei der Analyse des Aufgabenfokus der befragten Personen, lässt sich ein breites Spektrum sowohl in fachlicher Hinsicht als auch in bezug auf die Verantwortungsbereiche feststellen (siehe Abbildung 3.5). Hierbei ist die Konzentration auf konzeptionelle Schwerpunkte im Bereich der zentralen Data-Warehouse-Datenbasis



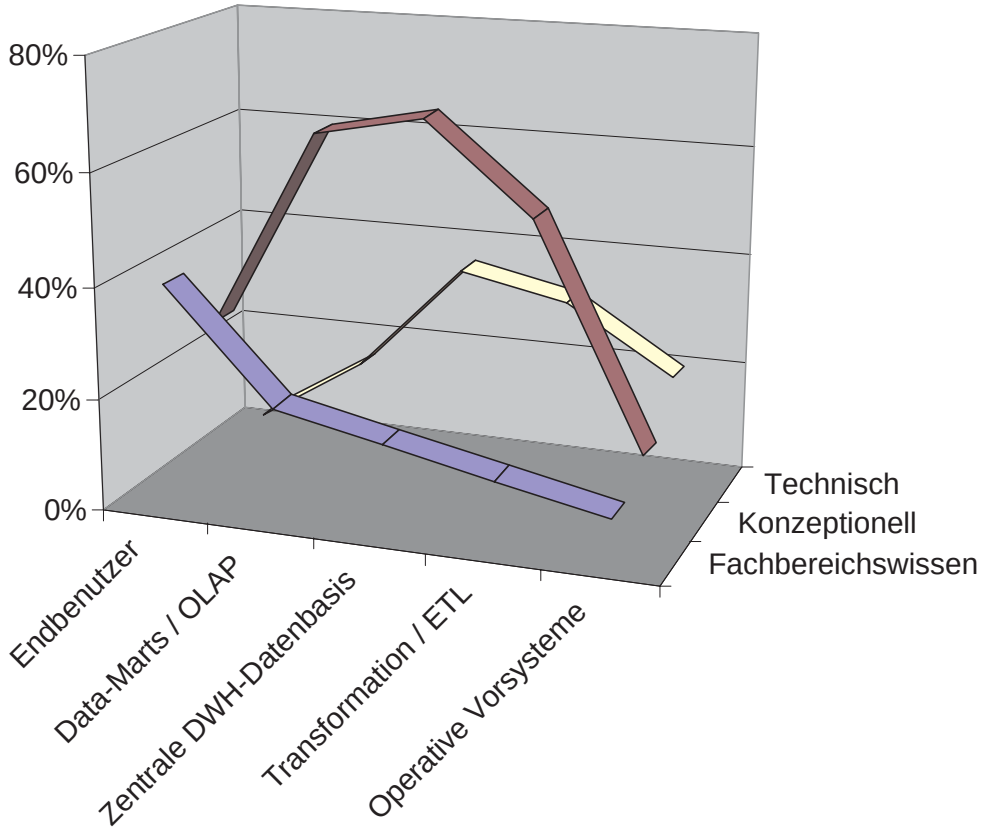
Angaben in % der ausgefüllten Fragebögen.

Abbildung 3.4: Unterstützte betriebswirtschaftliche Funktionen (Eigene Darstellung)

und der Data-Marts auffallend. Das spezifische Fachwissen nimmt erwartungsgemäss in Richtung operative Vorsysteme ab. Der technische Schwerpunkt liegt auf der Transformationskomponente und der zentralen Data-Warehouse-Datenbasis. Dies lässt Rückschlüsse auf die Komplexität dieser Transformationsaufgaben zu. Hier liegen die hauptsächlichen Probleme und Ursachen der Integration verschiedener Datenbestände.

3.3.2.1 Problembereiche und Massnahmen

Werden die Relevanz des Problembereichs und die Massnahmen zur Sicherstellung der Datenqualität untersucht, ergibt sich das in Tabelle 3.11 dargestellte Bild. Im Gegensatz zu 28%, für die Datenqualität kein besonderes Problem darstellt, ist Datenqualität für 60% der befragten Unternehmen ein grosses Problem. Für 8% der Rückmeldungen stellt Datenqualität sogar ein sehr grosses Problem dar. Die



Angaben in % der ausgefüllten Fragebögen.

Abbildung 3.5: Aufgabenfokus der befragten Personen (Eigene Darstellung)

Mehrzahl setzen zur Sicherstellung der Datenqualität sowohl Datenbereinigungstechniken als auch ein Datenqualitätsmanagement ein. Lediglich 12% geben an, dass in ihren Unternehmen keine besonderen Massnahmen zur Sicherstellung der Datenqualität getroffen werden.

Werden die Massnahmen zur Sicherstellung der Datenqualität näher untersucht, lassen sich neben den technisch orientierten Datenbereinigungsmassnahmen im wesentlichen organisatorische Regelungen erkennen. Teilweise werden organisatorische Einheiten in Form von fachübergreifenden oder fachspezifischen Datenqualitätsverantwortlichen, Datenbereinigungsteams, Business-Information-Managern, Data-Ownern und Data-Stewards genannt. Allerdings scheint die Abstimmung über die Verantwortungsbereiche derzeit bei der Mehrzahl der Unternehmen nicht abgeschlossen zu sein. Nicht alle Unternehmen haben klar geregelte Verantwortlichkeiten für die Datenqualität. Insbesondere sind Zielkonflikte

	Datenbe-reinigung	**DQM**	**DQM und Datenbe-reinigung**	**keine besonderen Massnahmen**	∑
sehr grosses Problem	0%	4%	4%	0%	8%
grosses Problem	4%	20%	36%	0%	60%
kein besonderes Problem	8%	4%	4%	12%	28%
kein Problem	0%	0%	0%	0%	0%
keine Angabe	4%	0%	0%	0%	4%
∑	16%	28%	44%	12%	100%

Angaben in % der ausgefüllten Fragebögen

Tabelle 3.11: Datenqualität in Data-Warehouse-Systemen

zwischen den Fachbereichen, den operativen Systemen sowie den Infrastruktureinheiten und den Anwendungssystem-Entwicklern zu erkennen. Häufig wird zur Abstimmung der Datenqualitätsziele und zur Problemidentifikation ein hoher Kommunikationsbedarf zwischen den verschiedenen Bereichen genannt.

Bei einigen Unternehmen wurden Qualitätssicherungsprozesse etabliert. Diese basieren auf speziellen Auswertungen und stichprobenhaften Datenqualitätskontrollen. Dabei wird die Datenqualität analysiert und schlussendlich durch den Datenqualitätsverantwortlichen freigegeben oder Verbesserungsmassnahmen eingeleitet. Für Änderungen der Transferprozesse oder der Datenmodelle existieren bei einigen Unternehmen organisatorische Regelungen zur Sicherstellung der Datenqualität. Dabei ist der Datenqualitätsbeauftragte bei Änderungen für die Qualitätssicherung verantwortlich und gibt erfolgreiche Anpassungen von Datenmodellen und Transferprozessen frei.

3.3.2.2 Datenqualitätsvorgaben

Bei der Frage nach den Qualitätsvorgaben und deren Überprüfung können die Antworten in zwei Kategorien unterteilt werden. Bei einigen Unternehmen findet bislang keine Vorgabe von Datenqualitätszielen statt, wobei diese allerdings für notwendig erachtet werden und geplant sind. Insbesondere scheint die Prüfung

der Datenqualität wichtig zu sein. Eine zweite Kategorie der Antworten nennt Qualitätsvorgaben durch festgelegte Standards und Liefervereinbarungen für Daten. In diesen Firmen werden

- eindeutig definierte Datenbeschreibungen,
- formale Angaben über die Syntax und die technische Bereitstellung,
- die Lieferzeitpunkte sowie
- bestimmte Angaben über ausgewählte charakteristische Eigenschaften der Daten (z. B. Anzahl von Fehler- oder Warnmeldungen)

festgelegt. Zur Bestimmung der Qualitätsvorgaben anhand von charakteristischen Eigenschaften werden historische Daten und Ladevorgänge im Zeitablauf analysiert und Vergleichswerte abgeleitet.

3.3.2.3 Datenqualitätsprüfungen

Qualitätsprüfungen werden meist durch Plausibilitätsprüfungen durchgeführt. Diese finden kontinuierlich oder in regelmässigen Abständen statt. Eine Analyse der Qualitätsprüfungen ergibt folgendes Bild:[269]

- 76% führen eine Qualitätsprüfung bei der Datenanlieferung (ETL-Prozess) durch.
- 68% prüfen die Datenqualität durch die Endanwender bei der Datenbereitstellung.
- 60% analysieren die Datenqualität des Data-Warehouse-Datenbestandes (Statistische Methoden, Data Mining und Integritätsbedingungen).
- 52% sichern die Datenqualität bereits bei der Datenmodellierung und der Systementwicklung (Standards und organisatorische Massnahmen).

[269] Angaben in Prozent der ausgefüllten Fragebögen.

- 20% setzen fest etablierte, organisatorisch geregelte und periodisch stattfindende Qualitätsprüfungen ein.

Bei den Qualitätsprüfungen werden folgende Kontrollen angegeben:

- Wertebereiche und Datentypen (technische Prüfung).
- Duplikate anhand von Schlüsselwerten.
- Referentielle Integrität.
- Vergleich mit Referenzdaten (Stammdaten und Schlüsseltabellen).
- Analyse der Systemprotokolle (z. B. Abgelehnte Datensätze im ETL-Prozess).
- Plausibilitäten (Datenbestand / Zeitvergleich).
  - Volumen- und Transaktionsmengen (z. B. Anzahl von Buchungen).
  - Summen über alle Datensätze einer Relation (z. B. Umsätze) und Datenverteilungen.
  - Datenabgleich mit anderen Systemen und den Datenquellen.
- Überprüfung durch den Endanwender (z. B. Reklamationen).
- Durch die Kunden, die Datenfehler melden oder bereits selbst korrigieren (z. B. durch engen Kontakt zum Kunden).

Neben einer bereits teilweise eingesetzten automatischen Prüfung werden die Qualitätsprüfungen meist durch standardisierte Auswertungen manuell vorgenommen.

3.3.2.4 Datenqualitätsmängel

Als wesentliche Probleme und Ursachen von Datenqualitätsmängeln werden aufgeführt:[270]

[270] Die Ergebnisse entsprechen im wesentlichen einer früheren Untersuchung über Datenqualitätsprobleme und deren Ursachen. Vgl. Helfert (2000a), S. 71f.; Helfert und Radon (2000), S. 115f.

- Inkorrekte Werte, fehlende Werte und Referenzen, Duplikate (diese behindern die operativen Systeme nicht oder nur unwesentlich).
- Inkonsistente Daten zwischen verschiedenen Anwendungssystemen.
- Falschbuchungen und Falscherfassungen.
- Ungenügende Plausibilitätsprüfungen in den operativen Systemen (z. B. bei der Dateneingabe).
- Änderungen in den operativen Systemen, die nicht dokumentiert und bekannt gegeben werden (z. B. bei Datenmodellen, Schlüsselwerten, Stammdaten, ...).
- Modellierungsfehler und redundant gehaltene Daten.
- Systemtechnische Probleme.

3.3.2.5 Sicherstellung der Datenqualität

Der Umgang mit den Datenqualitätsproblemen ist vielfältig und hängt von der Bedeutung des Problems, dem Anwendungskontext und der Problemursache ab. Als Möglichkeiten zur Sicherstellung der Datenqualität werden genannt:

- Qualitätsprüfungen und Abstimmung der Daten vor dem endgültigen Laden in die Data-Warehouse-Datenbasis.
- Datenbereinigung im ETL-Prozess.
- Laden und Kennzeichnen der problematischen Daten (z. B. bei Verletzung der referentiellen Integrität oder für Werte, die nicht in den Schlüsseltabellen oder Stammdaten vorhanden sind).
- Automatische Korrektur (z. B. bei Formatfehlern).
- Manuelle Korrektur (z. B. durch Interpretation der Datenwerte; häufig sind die Probleme bereits bekannt oder zumindest schnell identifiziert).

- Übermittlung der Prüfergebnisse an Datenlieferanten der operativen Systeme mit eventueller Datenkorrektur und erneuter Datenlieferung.
- Fehlersuche und Abstimmung mit den Datenlieferanten.
- Organisatorische Regelungen.

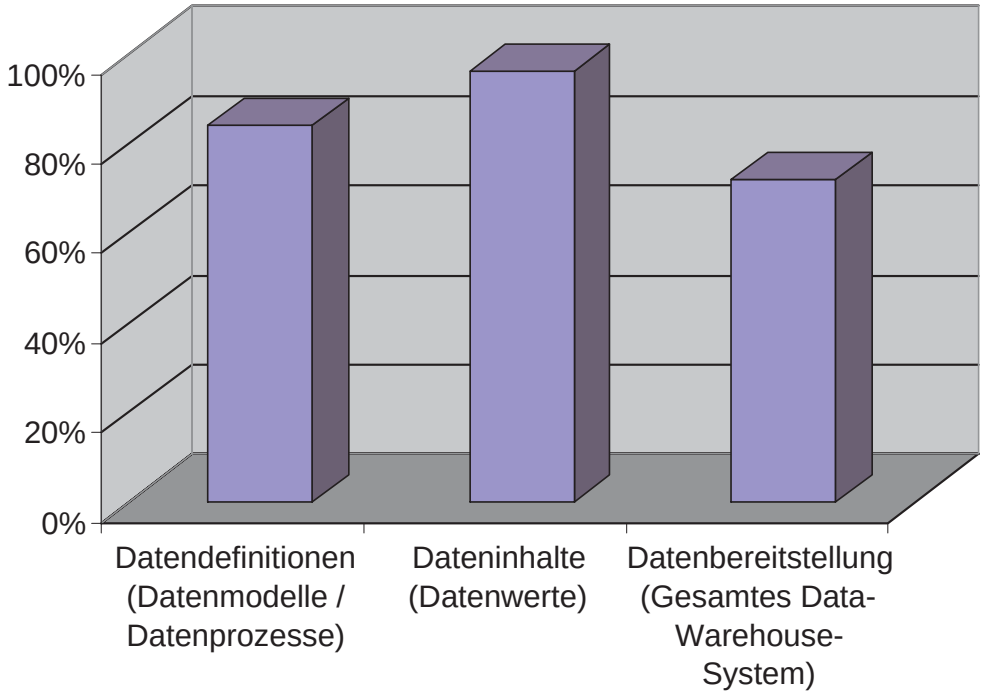
Interessanterweise hat kein Unternehmen die Integration der Qualitätsvorgaben und Qualitätsmessung in die Metadatenverwaltung aufgeführt.[271] Falls möglich, sollten die Datenqualitätsmängel an die Datenlieferanten gemeldet und dort deren Ursache behoben werden. Ein ständiger Kontakt zwischen den Verantwortlichen des zentralen Data-Warehouses und den Verantwortlichen der Quellsysteme ist hierbei nützlich.

3.3.2.6 Datenqualitätseigenschaften

Bei der Untersuchung des Begriffs der Datenqualität wird folgendes Ergebnis festgestellt. Zunächst ist zwischen Designqualität und Ausführungsqualität zu unterscheiden. Designqualität konkretisiert sich in der Angabe von Standards, Datendefinitionen und Dokumentationen. Bei der Ausführungsqualität liegt der Fokus auf den Dateninhalten und den Datenwerten. Ein weiterer wichtiger Aspekt der Datenqualität ist die Qualität der Datenbereitstellung, die insbesondere die Softwarekomponenten des Gesamtsystems berücksichtigt. Die Bedeutung der einzelnen Aspekte ist in Abbildung 3.6 zusammengefasst.

Werden die Eigenschaften zur Beschreibung qualitativ hochwertiger Daten analysiert, ist insbesondere die Widerspruchsfreiheit (Konsistenz) entscheidend. Der Datenbestand sollte inhaltlich und zeitlich konsistent sein. Bei den Datenwerten spielen Vollständigkeit und Korrektheit sowie die Repräsentation fehlender Werte eine grosse Rolle. Neben diesen ist die zeitliche Verfügbarkeit, Aktualität, referentielle Integrität und syntaktische Korrektheit der Datenwerte wichtig. Die Rückverfolgung der Datenherkunft und die Dokumentation von Abweichungen ist relevant.

[271] Allerdings wurde dies im Fragebogen nicht explizit untersucht.



Angaben in % der ausgefüllten Fragebögen.

Abbildung 3.6: Aspekte zur Beschreibung des Begriffs der „Datenqualität“ (Eigene Darstellung)

Weiter ist die Semantik und Identifizierbarkeit von Daten für die Datenqualität wichtig. Hier ist insbesondere eine einheitliche, klare und genaue Beschreibung der Datenmodelle und Datenflüsse zu nennen. Die Präzision der Wertebereiche und die Granularität der Datenmodelle scheinen weniger kritisch zu sein. Systemtechnische Aspekte, Datensicherheit und Zugriffsrechte sind weniger wichtig oder werden nicht als Eigenschaft von Datenqualität betrachtet. Die anhand einer offenen Frage genannten Datenqualitätseigenschaften sind in Tabelle 3.12 zusammengestellt. In Tabelle 3.13 sind die anhand einer vorgegebenen Liste eingeordneten Datenqualitätskriterien aufgeführt.[272]

[272] Eine vollständige Auflistung der Ergebnisse findet sich im Anhang A.2.

Eigenschaft	∑	**Eigenschaft**	∑
Konsistenz (inhaltlich, zeitlich, charakteristische Eigenschaften der Daten, Referenzdaten)	17	Verfügbarkeit	2
Richtig / korrekt / fehlerfrei	9	Wertebereichskonform	2
Vollständig	9	Zugriffsmöglichkeit / Zugänglichkeit	2
Genau definiert (charakteristische Eigenschaften der Daten, Schnittstellen, Dokumentation)	7	Datenqualitätskonzept	1
Aktuell	6	Deckungsgleich zu Erfahrungen (Statistik der Fachbereiche)	1
Einheitliche, standardisierte Form	4	Den operativen Daten entsprechend	1
Zeitlicher Bezug (historisiert)	4	Einheitliche Ermittlung	1
Interpretierbarkeit / aussagekräftig	3	Keine Dubletten	1
Datenherkunft nachvollziehbar / Abweichungen dokumentiert	3	Keine Datenleichen	1
Syntaktische Korrektheit	3	Motivation der DQ-Verantwortlichen	1
Abbild des Unternehmens	2	Pünktlichkeit	1
Anwenderbezogen / entscheidungsunterstützend	2	Sicherheit	1
Auswertbar	2	Überwacht	1
Granularität	2	Vergleichbar	1
Genügend präzise	2	Zentral vorgehalten	1
Referentielle Integrität	2	Zuverlässig	1
Relevanz	2		

Anzahl der Nennungen; ähnlich lautende Angaben in Gruppen zusammengefasst

Tabelle 3.12: In der Untersuchung genannte Datenqualitätseigenschaften

Rang	Welche Datenqualitätsmerkmale sind ...					
	entscheidend	**sehr wichtig**	**wichtig**	**vernachlässigbar**	**unwichtig**	**keine Eigenschaften von DQ**
1	Widerspruchs-freiheit (Datenbestand)	Identifizierbarkeit	Repräsentation fehlender Werte	Erforderlichkeit	Relevanz	Zugriffsrechte
2	Semantische Korrektheit	Widerspruchs-freiheit (Regeln)	Erforderlichkeit	Synonyme	Nicht-Volatilität	Transaktions-verfügbarkeit
3	Semantik	Zeitliche Konsistenz	Zeitliche Verfügbarkeit	Wertebereichs-definitionen	Wertebereichs-definitionen	System-verfügbarkeit
4	Identifizierbarkeit	Vollständigkeit	Synonyme	Aktualität	Aktualität	Zeitliche Verfügbarkeit
5	Datenherkunft	Nicht-Volatilität	Transaktions-verfügbarkeit	Vollständigkeit	Vollständigkeit	Relevanz

Die fünf häufig genannten Datenqualitätseigenschaften der jeweiligen Kategorie

Tabelle 3.13: Relevanz der Datenqualitätseigenschaften

3.4 Datenqualitätsmanagement

3.4.1 Historische Entwicklung des Qualitätsmanagements

Zunächst bietet sich eine historische Betrachtung des allgemeinen Qualitätsmanagements insofern an, dass hierdurch der Wandel und die Prinzipien derzeitiger Qualitätskonzepte deutlich werden. Die Entwicklung des Qualitätswesens nach 1950 ist grob in vier Stufen zu unterscheiden.[273] Diese sind in Abbildung 3.7 zusammengestellt.

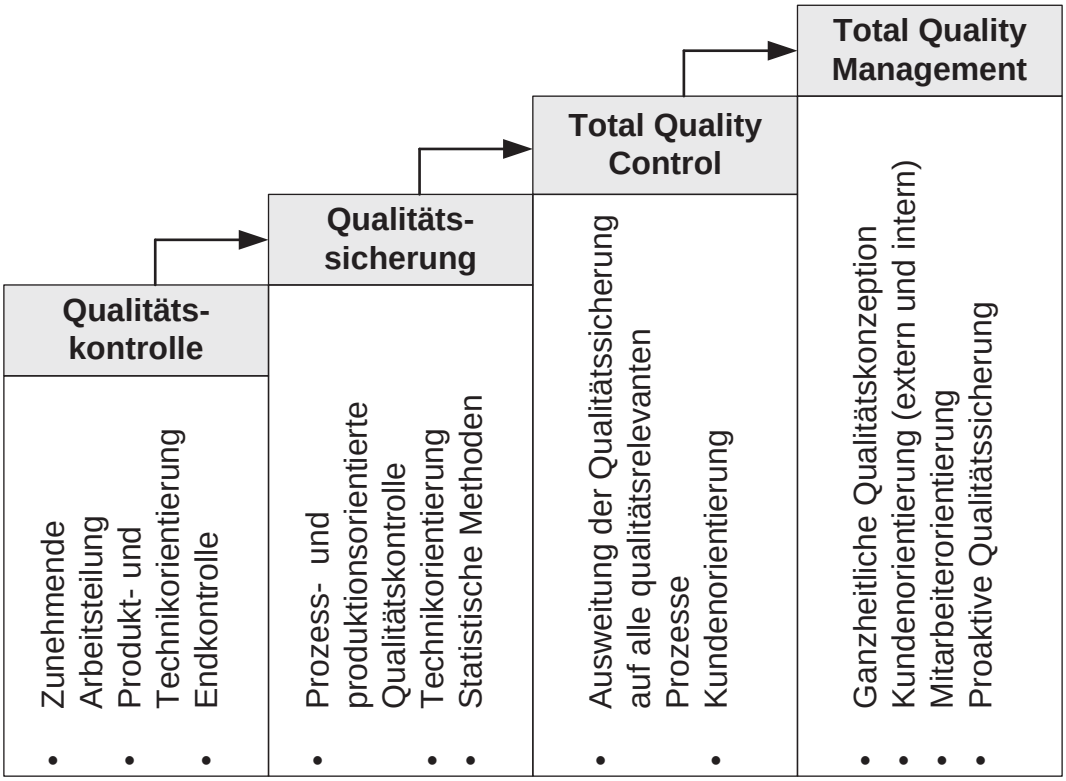


Abbildung 3.7: Entwicklungsstufen des Qualitätswesens (In Anlehnung an Wolf (1999), S. 63.)

In einer ersten Stufe, der *Qualitätskontrolle*, stand das Endprodukt mit dessen Produkteigenschaften im Vordergrund. Charakteristisch für diese Stufe ist die produktbezogene Qualitätssichtweise. Im Zentrum standen Qualitätsprüfungen

[273] Vgl. Ketting (1999), S. 17-30; Wolf (1999), S. 62-75; Seghezzi (1996), S. 5-9; Wallmüller (1995), S. 41f.

im Sinne einer Endkontrolle. Dies führte dazu, dass Qualitätsmängel dort behoben wurden, wo sie entdeckt wurden, am fertigen Produkt. Durch ausgeprägte Prüfabteilungen[274] versuchte man, die Qualität in die Produkte gewissermassen „hineinzuprüfen“ anstatt Qualität durch qualitätssichernde Prozesse zu produzieren.

Nach und nach wurde die Qualitätskontrolle im Sinne einer prozessorientierten Qualitätssicherung stärker in die Entwicklungs- und Produktionsprozesse integriert. Zunehmend fand eine Ausrichtung auf alle qualitätsrelevanten Produktionsprozesse bis hin zu einer als *Qualitätssicherung* bezeichneten zweiten Stufe des Qualitätswesens statt.[275] Ziel der Qualitätssicherung ist es, fehlerfreie Produktionsprozesse zu etablieren und die Fehler am Entstehungsort zu beheben. Um Abweichungen von Prozessparametern zu erfassen und regelnd in den Prozess einzugreifen, entstand der Ansatz der statistischen Prozesskontrolle (SPC)[276] als wesentliches Instrument der prozessorientierten Qualitätssicherung. Als Erweiterung zu den üblichen Abnahmeprüfungen, welche die Qualität erst zum Produktionsende feststellen, werden Prozessabweichungen regelmässig während der laufenden Produktion auf Grundlage von Stichproben untersucht. Grundlage der SPC bildet die Theorie des statistischen Testens,[277] die bestimmte Fragestellungen über ein Merkmal in einer Grundgesamtheit mit Hilfe statistischer Methoden auf Basis von Stichproben klärt. Hierzu wird eine formale Entscheidungsregel konstruiert, die einen Ablehnungsbereich für eine bestimmte Hypothese über die Grundgesamtheit ermittelt.

Zunehmend setzte sich die Erkenntnis durch, dass sich mit der prozessorientierten Qualitätssicherung nicht alle Qualitätsmängel verhindern lassen. Mängel, deren Ursache in den produktionsvorbereitenden Bereichen liegen, sind durch die prozessorientierte Qualitätssicherung nicht zu beheben. Es fand daher eine Ausweitung des Qualitätswesens auf alle qualitätsrelevanten Prozesse und Aktivitäten von der Forschung und Entwicklung über Konstruktion und Vertrieb bis zum Kun-

[274] Vgl. Ketting (1999), S. 24. Im deutschen Sprachgebrauch auch als Revision bezeichnet.

[275] Zur historischen Entwicklung und aktueller Entwicklungstendenzen der Qualitätssicherung vgl. Rinne und Mittag (1995), S. 24-33.

[276] Zur historischen Entwicklung der SPC vgl. z. B. Ketting (1999), S. 24f.; vgl. auch Osanna (2001), S. 1101-1105; Weihs, Jessenberger und Grize (1999), S. 24-27; Rinne und Mittag (1995), S. 77ff.

[277] Vgl. z. B. Fahrmeier, Künstler, Pigeot und Tutz (1997), S. 387ff.

dendienst statt. Wesentliche Kernpunkte dieser dritten Stufe des Qualitätswesens legte FEIGENBAUM bereits 1961 unter dem Begriff *Total Quality Control* fest.[278] An dieser Entwicklung hatte neben den USA und den wichtigen Industrieländern in Europa auch Japan einen entscheidenden Anteil. Insbesondere in den 70er und 80er Jahren wurden in Japan durch intensive Bemühungen nicht nur die Qualität der Produkte in entscheidendem Masse verbessert, sondern auch Methoden im Hinblick auf eine ganzheitliche Qualitätsbetrachtung entwickelt und zur Anwendungsreife gebracht. So entstanden beispielsweise auf Basis der Arbeiten von DEMING das „Kaizen“ oder die von AKAO propagierte Methode „Quality Function Deployment“.[279]

Nach und nach erfolgten gedankliche, konzeptionelle, methodische und technologische Erweiterungen und führten letztlich zu einem heute unter der Bezeichnung *Total Quality Management* (TQM) bekannten ganzheitlichen Qualitätskonzept. Total Quality Management stellt eine umfassende und methodenübergreifende Unternehmensphilosophie dar, in deren Mittelpunkt eine ganzheitliche Qualitätskonzeption unter Einbezug aller Mitarbeiter und gesamtgesellschaftlicher Aspekte mit dem Ziel der absoluten Kundenzufriedenheit steht.[280] In der Begriffsdefinition nach DIN ISO umfasst das Qualitätsmanagement alle Tätigkeiten der Gesamtführungsaufgabe, welche die Qualitätspolitik, die Qualitätsziele und die Verantwortungen für die Qualität festlegt.[281] Total Quality Management besteht im wesentlichen aus Methoden und Verfahren sowie Verhaltensweisen und Einstellungen, welche sich nach SEGHEZZI in das St. Galler Managementkonzept einordnen lassen.[282] Ergebnis ist das in Abbildung 3.8 dargestellte Konzept des integrierten Qualitätsmanagements. Es stellt ein dreidimensionales Gebilde dar, mit drei Ebenen als erste, drei Säulen als zweite und der im zeitlichen Ablauf stattfindenden Unternehmensentwicklung als dritte Dimension. Die Führungsaufgabe wird in die drei Ebenen des normativen, strategischen und operativen Managements untergliedert. Die mittlere Säule stellt die Aktivitäten dar, die einerseits

[278] Vgl. Feigenbaum (1961).

[279] Vgl. Ketting (1999), S. 27.

[280] Vgl. beispielsweise DIN, Deutsches Institut für Normung e. V. (1995), S. 318.

[281] Vgl. DIN, Deutsches Institut für Normung e. V. (1995), S. 244-246; Geiger (1994), S. 69-73.

[282] Vgl. hierzu Seghezzi (1996), S. 48ff. Auf eine detaillierte Betrachtung des Konzeptes soll verzichtet und lediglich auf Seghezzi (1996) verwiesen werden.

durch Strukturen unterstützt und andererseits durch das Verhalten der Führungskräfte und Mitarbeiter geprägt werden. Die dritte Dimension betrifft den zeitlichen Aspekt der Entwicklung der Qualitätsfähigkeit.

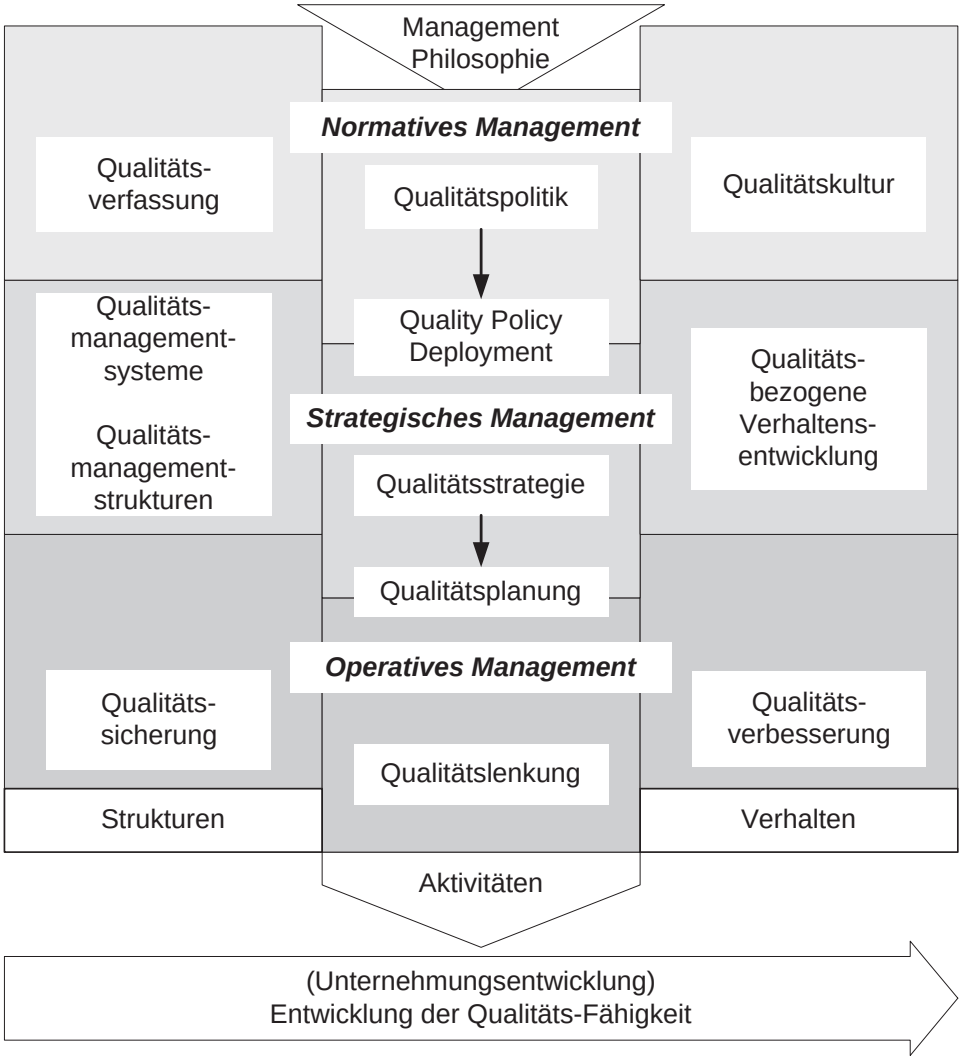


Abbildung 3.8: Integriertes Qualitätsmanagement (Vgl. Seghezzi (1996), S. 50)

Die bisherigen Ausführungen zeigen die Komplexität des Qualitätsmanagements und liefern die Basis für ein Konzept eines Datenqualitätsmanagements (DQM). Bisherige Konzepte zur Sicherstellung der Datenqualität stellen isolierte Techniken, Methoden und Organisationsansätze dar, die weder in sich geschlossen noch in allgemeine Managementkonzepte, insbesondere in das Informationsmanagement, eingeordnet sind. Im folgenden Abschnitt werden die Kernaspekte des als proaktiven Datenqualitätsmanagements bezeichneten ganzheitlichen Ansatzes

erläutert. Ausgehend von diesem Konzept fokussiert die Arbeit auf das operative Datenqualitätsmanagement für analytische Informationssysteme und ist insbesondere auf die Aufgaben der Qualitätsplanung und -lenkung in Data-Warehouse-Systemen bezogen.

3.4.2 Konzept eines proaktiven Datenqualitätsmanagements

Bevor das Konzept eines proaktiven Datenqualitätsmanagements dargestellt wird, soll zunächst der Begriff des Informationsmanagements im Überblick erläutert werden. Informationsmanagement ist bislang nicht einheitlich definiert, umfasst eine enorme Breite und unterliegt verschiedenen Interpretationen.[283] Allgemein versteht man darunter[284]

- primär die Aufgabe, Informationen für das Unternehmen zu beschaffen und in einer geeigneten Informationsstruktur bereitzustellen, und
- die hierfür erforderliche Infrastruktur langfristig zu planen und mittel- und kurzfristig zu beschaffen und einzusetzen.

Es umfasst die Gesamtheit aller Führungsaufgaben einer Wirtschaftseinheit bezogen auf den automatisierten bzw. automatisierbaren Teil des Informationssystems.[285] Die Aufgaben des Informationsmanagements kann man ähnlich der bei Managementkonzepten üblichen Einteilung in strategische sowie taktische und operative Aufgaben gliedern.[286]

Ausgehend von dieser kurzen Darstellung des Begriffs und der Aufgaben des Informationsmanagements kann man Datenqualitätsmanagement als Teilaspekt des Informationsmanagements eingliedern.[287] In Anlehnung an die Begriffsdefiniti-

[283] Vgl. z. B. Peterhaus (1995), S. 327ff.; Mertens et al. (2000), S. 172; Ferstl und Sinz (2001), S. 71-75; Heinrich (1992), S. 17ff.; Stahlknecht und Hasenkamp (1999), S. 452ff.; Beier, Gabriel und Streubel (1997), S. 1ff.

[284] Vgl. Stahlknecht und Hasenkamp (1999), S. 452.

[285] Vgl. Beier et al. (1997), S. 1.

[286] Vgl. Stahlknecht und Hasenkamp (1999), S. 453ff.; Ferstl und Sinz (2001), S. 73; Heinrich (1992), S. 20f.

[287] Vgl. auch Ausführungen in Heinrich (1992), S. 81-89.

on des Qualitätsmanagements nach DIN ISO[288] soll *Datenqualitätsmanagement* beschrieben werden

> als die Gesamtheit aller Tätigkeiten der Gesamtführungsaufgabe, welche die Datenqualitätspolitik, die Datenqualitätsziele und die Verantwortungen für die Datenqualität festlegt.

Es ist der Teil des Informationsmanagements, welcher die qualitativen Aspekte von Daten umfasst und insbesondere die Sicherstellung qualitativ hochwertiger Daten zum Ziel hat. Ausgehend von dieser begrifflichen Abgrenzung, den Grundsätzen des TQM und dem Konzept des integrierten Qualitätsmanagements lässt sich ein Ansatz für ein proaktives Datenqualitätsmanagement entwickeln, welcher sich grundsätzlich auf drei zentrale Bereiche stützt:[289]

- Die Verpflichtung des Managements, die Sicherstellung qualitativ hochwertiger Daten als Ziel festzulegen sowie Datenqualität als Teil der allgemeinen Qualitätsphilosophie und Unternehmenskultur zu akzeptieren und vorzuleben, so dass ein Bewusstsein und eine Kultur für qualitativ hochwertige Daten aufgebaut und gefördert wird. Auf Basis formulierter Unternehmensgrundsätze und -ziele ist eine Datenqualitätspolitik abzuleiten, welche die Qualitätsgrundsätze in bezug auf Daten festlegt und Ausgangspunkt zur Entwicklung einer Qualitätsstrategie für Daten bildet. Die Qualitätsplanung setzt Qualitätsstrategien in operationalisierbare Ziele um und bildet so den Übergang zum operativen Datenqualitätsmanagement. Eine derartige Operationalisierung strategischer Unternehmensziele über abzuleitende Ziele des Informations- und Datenqualitätsmanagements bis hin zu konkreten Zielvorgaben für einzelne Komponenten und Prozesse des Informationssystems basiert auf einem durchgängigen Zielsystem.[290] Dieses berücksichtigt die Sicherstellung einer unternehmensweiten Datenqualität. Bei der Bildung des Zielsystems sind insbesondere konfliktäre Beziehungen zwischen den Zielen einzelner Teilbereiche des Gesamtunternehmens zu beachten.

[288] Vgl. DIN, Deutsches Institut für Normung e. V. (1995), S. 244-246.

[289] Vgl. Helfert (2000a), S. 67f.; Wolf (1999), S.180ff.; Wallmüller (1995), S. 82ff.; English (1999), S. 70 ff.; vgl. auch Heinrich (1992), S. 86f.

[290] Vgl. auch Zielsystem des Informationsmanagements in Beier et al. (1997), S. 11-16.

- Ein Qualitätsmanagementsystem, welches den organisatorischen Rahmen verkörpert, ist zu etablieren und in die bisherige Organisationsstruktur zu integrieren. Dieses umfasst die Aufbau- und Ablauforganisation, die Zuständigkeiten, Verantwortlichkeiten, Prozesse und Mittel für das Qualitätsmanagement.[291] Es stellt sicher, dass in allen Bereichen geeignete Prozesse, Richtlinien, Pläne sowie Test- und Prüfverfahren zur Gewährleistung der Datenqualitätsziele etabliert sind. Im Sinne einer kontinuierlichen Qualitätsverbesserung ist eine ständige Überprüfung, Analyse und Verbesserung der festgelegten Massnahmen und durchzuführenden Prozesse erforderlich.

- Zur Unterstützung der Mitarbeiter bei der Ausübung der Qualitätsprozesse sind in allen Phasen und Bereichen der Datenversorgung geeignete Methoden, Verfahren und Werkzeuge zu entwickeln und zur Verfügung zu stellen.

Auf der operativen Ebene des Datenqualitätsmanagements können vier Aufgabenbereiche identifiziert werden.[292] Aufgabe der *Qualitätsplanung* ist es, unter Berücksichtigung einer Qualitätsstrategie, Bedürfnisse und Erwartungen an die Daten zu erfassen, diese in Vorgaben zu transformieren und Leistungen sowie Prozesse zu gestalten.[293] Im Rahmen der Qualitätsplanung werden Anforderungen an die Daten und deren Bereitstellungsprozesse festgelegt. Es sind dafür Qualitätsmerkmale auszuwählen, zu klassifizieren und zu gewichten sowie Sollwerte festzulegen.[294] Die *Qualitätslenkung* zielt auf die Einhaltung der in der Qualitätsplanung festgelegten Spezifikationen und der Beherrschung der Prozesse ab.[295] Hierfür sind zunächst geeignete Prozesse entlang des Datenflusses zu identifizieren und Massnahmen zur Erreichung der Prozesskonformität zu ergreifen. Die Qualität der Daten und der sie erzeugenden Prozesse müssen im Rahmen der Qualitätslenkung gemessen und in quantitativen Kennzahlen ausgedrückt werden. Wichtigstes Hilfsmittel für die Qualitätslenkung sind dabei Qualitätsprüfungen.[296] Schlussendlich sind Verantwortlichkeiten für die Qualitätslen-

[291] Vgl. DIN, Deutsches Institut für Normung e. V. (1995), S. 249; Geiger (1994), S. 145ff.
[292] Vgl. Heinrich, S. 86f.; Seghezzi, S. 53.
[293] Vgl. Seghezzi (1996), S. 72.
[294] Vgl. Wallmüller (1990), S. 19.
[295] Vgl. Seghezzi (1996), S. 76.
[296] Vgl. Wallmüller (1990), S. 19.

kung festzulegen und die Messergebnisse als Rückkopplung in Form von Regelkreisen zurückzuführen. Die *Qualitätssicherung* ist als strukturelle Unterstützung der Qualitätsplanung und Qualitätslenkung zu verstehen. Sie zielt darauf ab, Qualitätsrisiken systematisch zu erkennen, aufzudecken und ihre Wirkung zu bekämpfen.[297] Voraussetzung der Qualitätssicherung sind Risikoanalysen, wie beispielsweise die der Fehlermöglichkeits- und -einflussanalyse.[298] Der vierte Aufgabenbereich des operativen Datenqualitätsmanagements ist die *kontinuierliche Verbesserung* der Datenqualität.[299] Während Qualitätslenkung und Qualitätssicherung stabilisierend und veränderungshemmend wirken, fördert die kontinuierliche Verbesserung die dynamische Steigerung des Qualitätsniveaus. Als wichtigstes Instrumentarium der Qualitätsverbesserung sind Verbesserungsprojekte zu nennen.

Aufgrund der zentralen Bedeutung für ein proaktives Datenqualitätsmanagement bildet insbesondere die Qualitätsplanung und Qualitätslenkung den Schwerpunkt der Arbeit. Diese Aufgabenbereiche stellen die zentrale Voraussetzung für ein proaktives Datenqualitätsmanagement dar.[300] Weiter wird die Qualitätsplanung und -lenkung auf analytische Daten in Data-Warehouse-Systemen eingeschränkt, wenngleich die Ganzheitlichkeit des Ansatzes berücksichtigt bleiben soll.

3.4.3 Operatives Datenqualitätsmanagement

In Abschnitt 2.4.2 wurden die Komponenten eines Data-Warehouse-Systems erläutert sowie auf die potentiellen Problemfelder bezüglich der Sicherstellung der Datenqualität hingewiesen. Zweck des Data-Warehouse-Systems ist die Bereitstellung analytischer Daten in qualitativ hochwertiger Form. Die verschiedenen Sichtweisen und Qualitätsmerkmale wurden oben diskutiert. Ziel der folgenden Abschnitte ist es, Prinzipien der Qualitätsplanung und Qualitätslenkung als Kern des proaktiven Datenqualitätsmanagements darzustellen.[301] Diese bilden

[297] Vgl. Seghezzi (1996), S. 108.

[298] Vgl. Seghezzi (1996), S. 99.

[299] Vgl. Seghezzi (1996), S. 111.

[300] Vgl. English (1999), S. 70ff.; Huang, Lee und Wang (1999), S. 16, Helfert, Herrmann und Strauch (2001), S. 10.

[301] Für weitere Ausführungen siehe hierzu auch Helfert et al. (2001), S. 10ff.; Helfert (2000a), S. 66-68; Helfert und von Maur (2001), S. 65-68.

die Grundlage zur Ableitung der Anforderungen an das operative Datenqualitätsmanagement, welche in Abschnitt 3.5 betrachtet werden.

3.4.3.1 Qualitätsplanung

Aufgabe der Qualitätsplanung ist es, Qualitätsbedürfnisse und Qualitätserwartungen zu erfassen und diese in Vorgaben zu transformieren. Ergebnis der Qualitätsplanung ist eine Qualitätsspezifikation, welche die unterschiedlichen, teilweise konfliktären Qualitätsforderungen der Endbenutzer zusammenführt und Qualitätsvorgaben festlegt. Die Designqualität bezieht sich in diesem Zusammenhang auf die Differenz zwischen den Qualitätsforderungen der Endbenutzer und der festgelegten Spezifikation. Durch Interviews und Workshops mit Vertretern der Partnerunternehmen des Kompetenzzentrums „Data Warehousing 2“ konnte festgestellt werden, dass Qualitätsforderungen einerseits vom Anwender bzw. dessen Rolle als auch andererseits vom Anwendungskontext in Form der zu erfüllenden Aufgabe abhängen.[302] Die Erkenntnisse konnten im Rahmen der durchgeführten Fallstudie und der empirischen Untersuchung weiter gestützt werden.[303]

Die Informationsbedarfsanalyse bildet die Grundlage zur Erfassung von Qualitätsforderungen, indem in einem ersten Schritt die Informationsbedürfnisse der Fachbereiche und der Entscheidungsträger ermittelt werden.[304] In einem zweiten Schritt findet eine Quelldaten- und Prozessanalyse statt, die Probleme und Möglichkeiten zur Erfüllung der Informationsbedürfnisse untersucht. In einem iterativen Abgleich zwischen Anforderungen und Möglichkeiten sind diese gegebenenfalls anzupassen und in realistischen Qualitätsforderungen zu konkretisieren.[305] Ergebnis der Informationsbedarfsanalyse und der Qualitätsplanung ist ein Fachkonzept, das als Basis für die Umsetzung des Data-Warehouse-Systems ein qualitätsgeprüftes und in sich abgestimmtes Funktions- und Datenmodell bildet. Es stellt ein detailliert dokumentiertes Sollkonzept des zu erreichenden Systems dar, das insbesondere die bereitzustellenden Daten, deren Form und Bereitstel-

[302] Vgl. Helfert et al. (2001), S. 27.
[303] Vgl. hierzu Kapitel 4 und Abschnitt 3.3.
[304] Vgl. Helfert et al. (2001), S. 20 und Abschnitt 2.4.2.2.
[305] Vgl. Holthuis (1999), S. 227.

lungszeitpunkte als auch die hierfür notwendigen Softwarekomponenten und deren Funktionen festlegt. Als wesentliche Elemente der Spezifikation sind zu nennen:

- Die konzeptionellen und logischen Datenmodelle der Data-Warehouse-Datenbasis, der Data Marts sowie der Quellsysteme.
- Die syntaktischen und semantischen Beziehungen zwischen den Datenmodellen und deren Aggregationsstufen.
- Der Datenbereitstellungsplan, der insbesondere die zeitlichen und funktionellen Anforderungen an die Transferprozesse und Datenerfassungsprozesse festlegt.
- Die daraus abgeleiteten, funktionellen Anforderungen an notwendige Softwarekomponenten, wie Endbenutzerwerkzeuge, Transferprogramme und Datenbanksysteme.

Bestandteil der Datenmodelle sind syntaktische und semantische Beschreibungen und Anforderungen an die Datenwerte, welche im Rahmen der Datenbereitstellung zu erfüllen sind. Der Datenbereitstellungsplan legt die Datenquellen und Datensenken, die auszuführenden Funktionen sowie deren Transferzeitpunkte fest. Auf physischer Ebene werden die konzeptionellen Anforderungen durch Softwarekomponenten umgesetzt, deren funktionale Anforderungen in Pflichtenheften festgelegt sind.

Während sich Datenqualität i. e. S. lediglich auf die Datenmodelle und deren Wertausprägungen konzentriert, ist das Begriffsverständnis dieser Arbeit umfassend. Es berücksichtigt alle Einflussgrössen der Qualität auf analytische Daten eines Data-Warehouse-Systems entlang des Datenflusses von der Datenentstehung bis hin zur Datenanwendung. Aus diesem Grunde sind bei der Festlegung des Sollkonzeptes nicht nur Datenmodelle sondern auch die Anforderungen an die Transferprozesse und die abgeleitete Spezifikation für die Softwarekomponenten zu berücksichtigen. Auf Basis der Qualitätsforderungen können dann Qualitätsvorgaben für die Quellsysteme und die Datenerfassungsprozesse erarbeitet werden. Unter Berücksichtigung der jeweiligen fachspezifischen Anforderungen

der operativen Systeme können dann Sollkonzepte für die Quellsysteme und die Erfassungsprozesse spezifiziert werden. *Ergebnis der Qualitätsplanung* ist eine Qualitätsspezifikation für Data-Warehouse-Systeme, bestehend aus Datenmodellen und Anforderungen an die Transfer- und Datenerfassungsprozesse sowie Software- und Kommunikationskomponenten.

3.4.3.2 Qualitätslenkung

Nach Festlegung der Qualitätsforderungen in einem detaillierten Sollkonzept gilt es, diese im Rahmen der Systemumsetzung und des regelmässigen Betriebs einzuhalten. Diese Aufgabe wird als Qualitätslenkung bezeichnet, welche auf die Einhaltung der Spezifikationen und die Beherrschung der Prozesse abzielt. Die Qualitätslenkung betrifft die Ausführungsqualität des Data-Warehouse-Systems. Im allgemeinen Qualitätsmanagement hat sich die Übertragung des Prinzips der Regelung bewährt.[306] Nach einer Beschreibung der wichtigsten Grundbegriffe,[307] wird dessen Anwendung auf die Qualitätsplanung und -lenkung beschrieben.

Wie in Abbildung 3.9 dargestellt, findet bei der Steuerung im Gegensatz zum Prinzip der Regelung keine Rückkopplung über die aktuellen Zustände der Steuerstrecke statt. Unter Steuerung wird die Anweisung einer Systemkomponente an eine andere Komponente verstanden, sich in einer bestimmten Art zu verhalten. Das dabei unterstellte funktionale Verhalten beider Systemkomponenten wird in Form von Input-Output-Systemen beschrieben. Die zu beeinflussende Systemkomponente wird als Steuerstrecke bezeichnet. Diese wird durch die Steuereinrichtung aufgabenmässig beeinflusst. Eine von der Steuerung nicht unmittelbar beeinflusste, extern vorgegebene Zielgrösse, wird als Führungsgrösse *w* bezeichnet. Von der Führungsgrösse wird die Stellgrösse *y* abgeleitet. Diese überträgt dann ihre Wirkung auf die Steuerstrecke. Alle externen Grössen, welche die beabsichtigte Beeinflussung der Steuerstrecke beeinträchtigen, werden als Störgrössen *z* bezeichnet. In betrieblichen Systemen sind nicht exakt vorherbestimmte Störungen üblich, was auch für Data-Warehouse-Systeme gilt. Aus diesem Grunde eignet

[306] Vgl. Pfeifer (2001), S. 998-1002.

[307] Vgl. zum folgenden Haberfellner (1975), S. 49ff.; Ulrich (1970), S. 120-128; Baetge (1974), S. 23-36; Flechtner (1984), S. 27-43; Ferstl und Sinz (2001), S. 22-28.

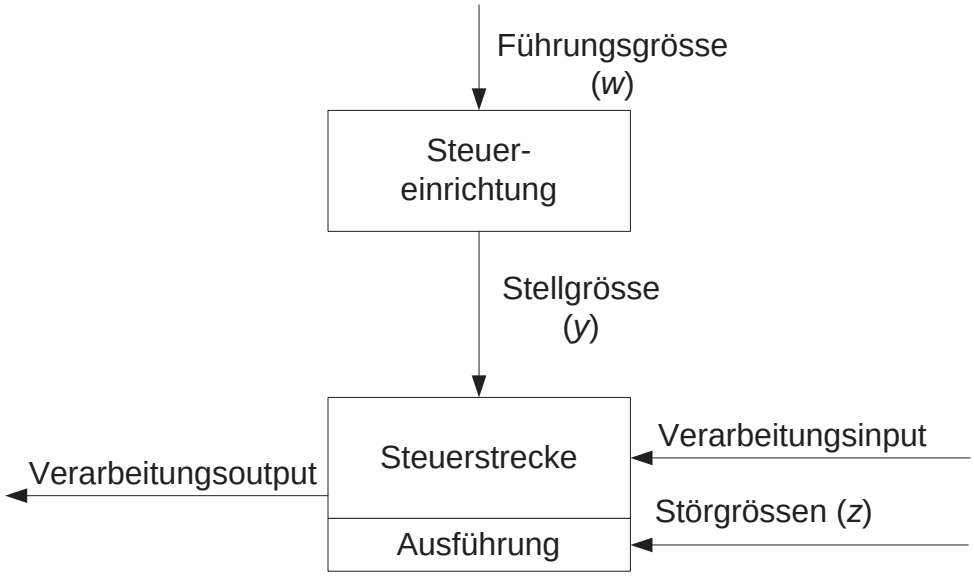


Abbildung 3.9: Prinzip der Steuerung (In Anlehnung an Haberfellner (1975), S. 49)

sich das Prinzip der Steuerung nur für begrenzte Anwendungen.

Im Gegensatz zur Steuerung wird beim Prinzip der Regelung fortlaufend der Systemzustand erfasst und mit der Führungsgrösse verglichen (vgl. Abbildung 3.10). Bei Abweichungen wird durch Anpassung der Stellgrösse *y* eine kontinuierliche Annäherung der zu regelnden Grösse an die Führungsgrösse bewirkt. So wird eine permanente, zyklische Interaktion der beiden Systemkomponenten eines Regelkreises über die Regelgrösse einerseits und die Stellgrösse andererseits ermöglicht. Die beiden Komponenten eines Regelkreises werden als Regelstrecke und Regler bezeichnet. Der aktuelle Zustand der Regelstrecke wird an den Regler durch Regelgrössen in Form von Prozess- und Zustandsbeschreibungen übermittelt.[308] Diese Informationen werden im Rahmen von betrieblichen Prozessen auch als Kontrollinformationen (KOI) bezeichnet.[309] Stellgrössen werden bei betrieblichen Prozessen als Sollwertinformationen (SOI) bezeichnet und umfassen als Oberbegriff die Stellgrösse *y* und die Führungsgrösse *w*. Im Rahmen der Planung und Kontrolle werden SOI ausgehend von der aktuellen KOI mit Hilfe eines

[308] Vgl. hierzu auch die Ausführungen zum Qualitätsmodell in Abschnitt 3.5.
[309] Vgl. hierzu Haberfellner (1975), S. 50ff.

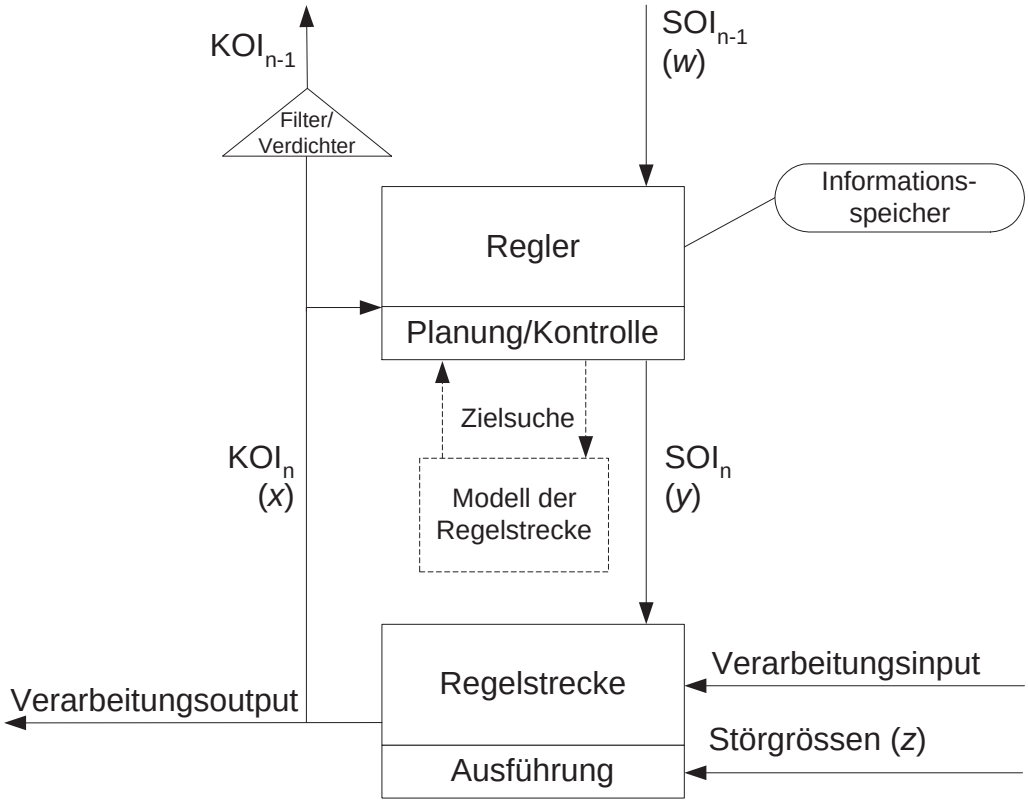


Abbildung 3.10: Prinzip der Regelung (In Anlehnung an Haberfellner (1975), S. 55)

Modells über die Regelstrecke bestimmt. Regelkreise können auch hierarchisch angeordnet sein. In diesem Fall bezeichnet *n* die Stufe der Hierarchie. Übergeordnete Regelungsstufen bestimmen entsprechende SOI für untergeordnete Stufen auf Basis der hierzu notwendigen KOI untergeordneter Regelkreise.

Bei der Qualitätsplanung für Data-Warehouse-Systeme entspricht die Stellgrösse den Qualitätsforderungen für analytische Daten. Aktionen oder Tätigkeiten, die Einfluss auf die Entstehung analytischer Daten nehmen, entsprechen der Regelstrecke. Im Fokus des operativen DQM sind hier die Datenerfassungs-, Transformationsprozesse, Datenmodelle und Datenbanksysteme als auch auf physischer Ebene die Software- und Kommunikationskomponenten zu nennen. Die Stellgrösse ist die Grösse, durch deren Änderung die Regelgrösse über die Regelstrecke beeinflusst werden kann. Sie resultiert aus dem Vergleich von Sollgrösse und Regelgrösse, also von gemessener Qualität und den Qualitätszielen. Ursachen

für Abweichungen bilden Störgrössen, die durch eine ungeplante und veränderte Einwirkung auf die Regelstrecke hervorgerufen werden. Sie beeinflussen so das die Regelgrösse tragende Datenqualitätsmerkmal. Die Regelstrecke der Qualitätsplanung kann, wie in Abbildung 3.11 dargestellt, in einem hierarchischen Regelkreismodell weiter detailliert werden.

Als Regelstrecke ist die für den Regelbereich betrachtungsrelevante Prozesskette, ein Ausschnitt hieraus oder gar ein einzelner Prozess zu verstehen. Ein Prozess kann jedoch nur korrekt ausgeführt werden, wenn die entsprechenden Anforderungen als Vorgaben definiert und bekannt sind; erst dann wird Qualität als Ergebnis von Prozessen messbar. Den in der Qualitätsplanung definierten Qualitätsforderungen sind konkrete Sollgrössen zuzuordnen. Durch die Qualitätsplanung werden Sollgrössen und Prüfmerkmale für Prozesse und Daten ausgewählt und gewichtet.[310] Die Sollgrössen gehören dann zu den wesentlichen Eingangsinformationen für den Gestaltungsrahmen der Qualitätslenkung und den darin festgelegten Qualitätssicherungsmassnahmen. Neben den Sollgrössen ist die Kenntnis der zu regelnden Grösse und ihrer Qualität wichtig. Diese Grössen werden durch Qualitätsprüfungen bestimmt und dienen zum Abgleich zwischen Qualitätsforderungen und erzeugter Qualität. Hierzu sind Kontrollpunkte in der Prozesskette zu bestimmen (Eigenschaften der Daten und Prozesse). Verantwortlich für die Ableitung und Bereitstellung der geeigneten Qualitätsmassnahmen ist der Regler der Qualitätslenkung. Über ihn werden Soll- und Regelgrössen miteinander verglichen, bewertet und aus der Differenz, dem erkannten Qualitätsdefizit, die Stellgrösse gebildet, d. h. geeignete Qualitätsmassnahmen ermittelt. Die Stellgrösse ist das Ergebnis des Abgleichs zwischen Sollgrösse und Regelgrösse und stellt letztlich die durchzuführenden Qualitätssicherungs- und Qualitätsverbesserungsmassnahmen zur Behebung der Qualitätsdefizite dar. Der Regler entspricht so der Menge aller geeigneten, qualitätsverbessernden Massnahmen und Instrumente.

Idealerweise sind alle Abläufe in einem Unternehmen in geeignete Qualitätsregelkreise verschiedenen Typs und mit unterschiedlichen Einflussbereichen auszugestalten.[311] Der kleinste Regelkreis ist die unmittelbare Rückmeldung eines Prüfer-

[310] Vgl. auch Abschnitt 4.2.2.

[311] Vgl. hierzu Wolf (1999), S. 148.

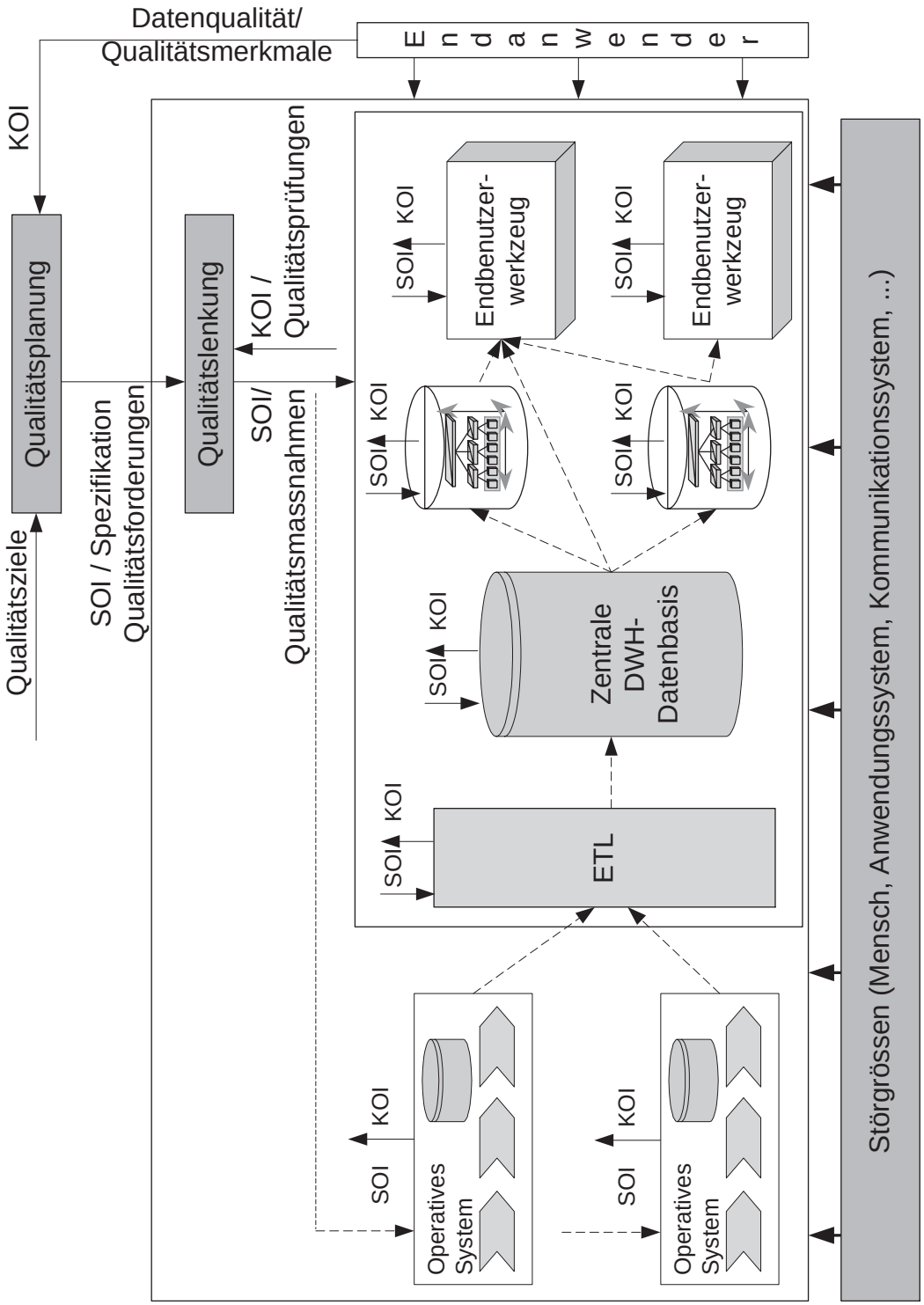


Abbildung 3.11: Regelkreismodell für das operative Datenqualitätsmanagement (Eigene Darstellung)

gebnisses. Ein Beispiel ist die Rückmeldung eines Fehlers nach einer unzulässigen Dateneingabe auf Grund einer Plausibilitätsprüfung in einem Datenfeld der Eingabemaske.[312] Des weiteren lassen sich „anwendungssystemnahe“ Regelkreise bilden, welche die Konsistenz zwischen verschiedenen Anwendungssystemen sichern. Weiter sind „ebeneninterne“ und „ebenenübergreifende“ Qualitätsregelkreise möglich.[313] Zur Umsetzung eines proaktiven Datenqualitätsmanagements, das die Bereitstellung von Daten in qualitativ hochwertiger Form zum Ziel hat, muss ein unternehmensweiter Qualitätsregelkeis aufgebaut werden, der im Sinne einer kontinuierlichen Verbesserung einen iterativen Prozess zur Sicherstellung und permanenten Verbesserung der Datenqualität ermöglicht.[314] Ausgehend von den Qualitätszielen gliedert sich der kontinuierliche Verbesserungsprozess der Datenqualität grob in die folgenden Phasen:[315]

- Spezifikation von Qualitätsforderungen unter Berücksichtigung der anwenderspezifischen Qualitätserwartungen, den gegenwärtigen Möglichkeiten und zu beachtenden Restriktionen (insbesondere in den operativen Vorsystemen).
- Ermittlung von Qualitätsmängeln durch Aufnahme und Analyse der gegenwärtigen Abläufe und Prozesse sowie der bereitgestellten Daten.
- Ableitung geeigneter Strategien und Massnahmen zur Behebung der Datenqualitätsmängel.
- Schwachstellen- und Ursachenanalyse zur Identifikation von Auswirkungen mangelnder Datenqualität.
- Planung, Umsetzung und Integration der Massnahmen in die bestehenden Strukturen.

[312] Vgl. Wolf (1999), S. 148.

[313] Vgl. Pfeifer (2001), S. 1000f. Im allgemeinen Qualitätsmanagement lassen sich „maschineninterne“, „maschinennahe“, „ebeneninterne“ und „ebenenübergreifende“ Qualitätsregelkreise voneinander abgrenzen. Der Schwierigkeitsgrad zum Aufbau solcher Regelkreise steigt aufgrund der erhöhten Komplexität mit zunehmender Spannweite. Im Rahmen von Informationssystemen soll begrifflich die Fokussierung auf Anwendungssysteme zum Ausdruck gebracht werden, so dass hier der Begriff des „anwendungssystemnahen“ Regelkreises verwendet werden soll.

[314] Vgl. hierzu Wolf (1999), S. 149ff., der einen unternehmensweiten Regelkreis für das Informationsqualitätsmanagement beschreibt.

[315] Vgl. Wolf (1999), S. 150.

- Permanente Kontrolle der aus der Umsetzung resultierenden Auswirkungen sowie Rückmeldung der Ergebnisse in den kontinuierlichen Verbesserungsprozess.

Das Prinzip der kontinuierlichen Qualitätsverbesserung lässt sich in den prozessorientierten Qualitätsansatz von DEMING, der die Schritte Plan, Do, Check und Act umfasst, veranschaulichen und einordnen.[316] Zunächst sind in einem ersten Schritt Qualitätsforderungen festzulegen (Plan) und in einer weiteren Phase umzusetzen (Do). Das Rückmelden der erreichten Situation und die Prüfung im Hinblick auf die gesetzten Ziele (Check) bewirkt wiederum eine Analyse des Zustandes und die Suche nach Ursachen und weiteren Verbesserungsmöglichkeiten (Act).

Die Konstruktion der Phasen des Regelkreises liefert die grundlegende Basis, um das Prinzip der kontinuierlichen Verbesserung umzusetzen. Innerhalb des unternehmensweiten Regelkreises sind weitere untergeordnete Regelkreise zur Detaillierung und Konkretisierung notwendig, um ein für den jeweiligen Anwendungsfall erforderliches Abstraktionsniveau zu erreichen. Ziel ist die Ableitung konkreter Qualitätsmassnahmen und Spezifikationen, die eine praktische Umsetzung in der Regelstrecke ermöglichen. Die durchgängige Anwendung des Prinzips der Regelung ermöglicht ähnlich strukturierte untergeordnete Regelkreise auf allen Ebenen. Im folgenden Abschnitt sollen die von diesem Prinzip abgeleiteten Anforderungen an das operative Datenqualitätsmanagement erläutert und zusammenfassend dargestellt werden.

3.5 Anforderungen an das operative Datenqualitätsmanagement

Grundlage des oben aufgezeigten Prinzips sind Sollwert- und Kontrollinformationen zur Qualitätsvorgabe und deren Kontrolle. Im folgenden sollen die daraus abgeleiteten grundsätzlichen Anforderungen des operativen Datenqualitätsmanagements und insbesondere die der Qualitätsplanung und Qualitätslenkung erörtert

[316] Vgl. Seghezzi (1996), S. 53; English (1999), S. 42f.; Wolf (1999), S. 152.; Helfert et al. (2001), S. 9f.

werden. Wie oben dargestellt, ist die Aufgabe der Qualitätsplanung die Erfassung von Qualitätsbedürfnissen und Qualitätserwartungen sowie die Transformation dieser in Vorgaben und Spezifikationen. Die Qualitätsvorgaben sind im Rahmen der Qualitätslenkung einzuhalten. Hierfür wird ein Hilfsmittel zur Spezifikation und zur Prüfung des Erfüllungsgrades von Qualitätsforderungen benötigt. Möglichkeiten zur Vorgabe und Prüfung liegen einerseits in der Beschreibung von Prozessergebnissen durch Produktmerkmale und andererseits in der Festlegung von Prozesseigenschaften, den Prozessmerkmalen.[317] Hierfür sind die wesentlichen Datenqualitätskriterien auszuwählen, zu klassifizieren und zu gewichten sowie in einem Qualitätsmodell abzubilden. Dessen grundsätzliche Struktur ist in Abbildung 3.12 dargestellt.[318] Ein zentraler Aspekt des Qualitätsmodelles ist die Zerlegungssystematik von Qualitätskriterien in Prozess- und Produktmerkmale sowie deren Operationalisierung auf unterster Ebene anhand von Qualitätskennzahlen. Durch den Einsatz solcher Qualitätsmodelle können[319]

- Qualitätsforderungen durch aussagekräftige und quantifizierbare Qualitätskennzahlen für Produkt- und Prozessmerkmale angegeben,
- die Qualität weitgehend transparent und objektiv bewertet sowie
- die Qualität im Sinne eines Qualitätsmanagements geplant, gesteuert und kontrolliert werden.

Im Falle eines Data-Warehouse-Systems beziehen sich die Produktmerkmale auf Eigenschaften der Datenbestände und die Prozessmerkmale auf die Transfer- und Erfassungsprozesse. Produktmerkmale können weiter anhand der Granularität unterschieden werden. So können beispielsweise bei relationaler Datenhaltung Eigenschaften auf Attributebene, auf Tupelebene, auf Relationenebene, auf Datenbankebene und zwischen verschiedenen Datenbanken festgelegt sein. Einige dieser Eigenschaften werden in Abschnitt 4.2.2.1 für das Datenqualitätsmerkmal der Glaubwürdigkeit aufgeführt. Ein wesentliches Merkmal der Transfer- und Erfassungsprozesse ist deren Ausführungszeit, die insbesondere die Aktualität der Daten beeinflusst. Dieser Aspekt wird in Abschnitt 4.2.2.2 erörtert.

[317] Vgl. Kaposi und Myers (1994), S. 97ff; Seghezzi (1996), S. 82.
[318] Vgl. Wallmüller (1990), S. 46ff.
[319] Vgl. Wallmüller (1990), S. 54.

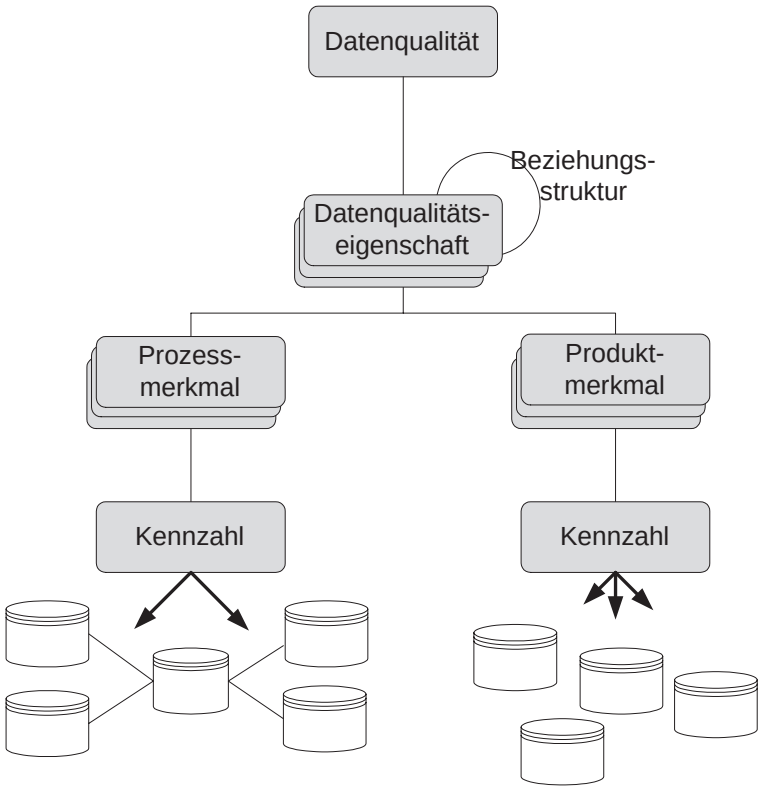


Abbildung 3.12: Struktur des Qualitätsmodells (In Anlehung an Wallmüller (1990), S. 46)

Wie bereits die allgemeine Literatur zur Qualität zeigt, ist sowohl die Begriffsbildung als auch die Gestaltung objektiver Bewertungskriterien anhand von Referenzmodellen noch nicht vollständig gelöst.[320] Dies trifft auch für Datenqualität zu, wobei die Festlegung geeigneter Messgrössen und Qualitätsindikatoren aktueller Forschungsgegenstand verschiedener Projekte ist.[321] Formal bedeutet Messung die Festlegung von Merkmalsausprägungen zu einem Merkmal.[322] Obwohl der Einsatz von Kennzahlen immer wieder in Frage gestellt wird, finden diese seit längerem in der Praxis Anwendung[323] und sollen daher als mögliche Basis für die operative Steuerung und Regelung im Rahmen des Datenqualitätsmanagements erörtert werden. Aus diesem Grund werden zunächst in Abschnitt 3.5.1

[320] Vgl. z. B. Caduff (1997), S. 49ff.

[321] Vgl. Abschnitt 3.6 und auch beispielsweise Soler und Yankelevich (2001); Milek et al. (2001); Jin und Embury (2001); Pokorny (2000); Kopcso, Pipino und Rybolt (2000); Dedeke (2000).

[322] Vgl. Rinne und Mittag (1995), S. 39.

[323] Vgl. Pfeifer (1996), S. 193f.; Botta (1997), S. 4.

Kennzahlen und die daraus abgeleiteten Anforderungen dargestellt, bevor in Abschnitt 3.5.2 auf die zentralen Anforderungen an ein Datenqualitätsmanagement eingegangen wird.

3.5.1 Kennzahlen und Kennzahlensysteme

Hauptsächliches Einsatzgebiet von Kennzahlen ist die operative Steuerung und Regelung von Unternehmen, indem Sollvorgaben und Istbeschreibungen durch Kennzahlen ausgedrückt werden.[324] Im oben aufgezeigten Regelkreismodell haben Kennzahlen daher eine hohe Bedeutung, denn einerseits können mit ihrer Hilfe Sollgrössen festgelegt und andererseits das Verhalten der Regelstrecke beschrieben werden. Aufgrund der Schwankungen der Kennzahl können dann die Entscheidungsparameter der zu regelnden Einheit entsprechend angepasst werden. Weitere Anwendungsgebiete lassen sich im innerbetrieblichen Einsatz durch Vergleiche im Zeitablauf sowie zwischen Soll- und Normgrössen finden.[325] Neben der innerbetrieblichen Anwendung sind Kennzahlen häufig auch für überbetriebliche Vergleiche geeignet.[326]

Eine geeignete Basis für die Ermittlung von Qualitätskennzahlen und darauf aufbauenden Kennzahlensystemen bildet die Kennzahlentheorie.[327] Über den Begriff der Kennzahl herrscht bislang allerdings wenig Einigkeit.[328] Übereinstimmung besteht weitgehend darüber, dass Kennzahlen im betriebswirtschaftlichen Verständnis der Operationalisierung von Unternehmenszielen dienen und somit die Überprüfung der Zielerreichung unterstützen sollen.[329] Allgemein kann unter einer Kennzahl eine Wertausprägung eines Merkmals für den inner- und zwischenbetrieblichen Vergleich verstanden werden und grob anhand

- des betriebswirtschaftlichen Tatbestands,

[324] Vgl. Mutscheller (1996), S. 41.
[325] Vgl. Mutscheller (1996), S. 41; Siegwart (1998), S. 13ff.
[326] Vgl. Mutscheller (1996), S. 41.
[327] Für das allgemeine Qualitätsmanagement existieren bereits Kennzahlensysteme wie z. B. von Bartram (1992), S. 226ff.
[328] Vgl. Botta (1997), S. 8f.; Meyer (1994), S. 1f.; Siegwart (1998), S. 5; Reichmann (1995), S. 19ff.
[329] Vgl. Mutscheller (1996), S. 32.

- ihrer Struktur und
- deren Zahlenwerte

untergliedert werden.[330] Der Einsatzbereich der Kennzahlen ist durch deren betriebswirtschaftlichen Tatbestand gekennzeichnet. Qualitätskennzahlen sind dabei der betriebswirtschaftlichen Funktion des Qualitätsmanagements zugeordnet. Bei der inneren Struktur von Kennzahlen unterscheidet man zwischen quantitativen, inhaltlichen und zeitlichen Komponenten. Durch den Messvorgang wird die Kennzahl durch den zahlenmässigen Aspekt konkretisiert, wobei die Zahl aus dem Ergebnis des zu messenden Sachverhaltes mit einer Messskala gebildet wird. Für diese lassen sich vier Grundarten unterscheiden.[331] Bei einer *Nominalskala* werden Klassen ohne jegliche Rangordnung gebildet, die lediglich nach dem Kriterium gleich oder verschieden differenziert werden. Die Merkmalsausprägungen sind Namen oder Klassenbezeichnungen. Werden die Klassen nach einer Rangfolge angeordnet, spricht man von einer *Ordinalskala*. Meist wird die Rangfolge durch eine natürliche Zahl, die man als Rangzahl bezeichnet, ausgedrückt. Eine Klassengrösse wird jedoch nicht festgelegt. Bei *Kardinalskalen* werden zusätzlich Verhältnisse von Merkmalsausprägungen spezifiziert. Es handelt sich um reelle Zahlen, die zudem noch eine Dimensionsangabe haben. Man unterscheidet Kardinalskalen weiter in:

Intervallskalen: Hier tritt zur Ordnungseigenschaft der Ordinalskalen die Berücksichtigung von Abständen zwischen den Merkmalsausprägungen hinzu.

Verhältnisskalen: Zur Intervallskala tritt ein natürlicher Nullpunkt hinzu, so dass auch Verhältnisse von Merkmalsausprägungen messbar werden.

Absolutskalen: Zur Verhältnisskala kommt eine natürliche Einheit, die mit dem betrachteten Ereignis in Beziehung steht. Absolute Kennzahlen lassen sich weiter in Einzelzahlen, Summen, Differenzen oder Mittelwerte unterteilen.

[330] Vgl. hierzu Mutscheller (1996), S. 32-34.
[331] Vgl. hierzu z. B. Rinne und Mittag (1995), S. 39f.; Meyer (1994), S. 2ff.; Bohley (1996), S. 65f.

Verhältniszahlen stellen wichtige Kennzahlen dar, die durch Bildung eines Quotienten ermittelt werden. Diese können generell in drei Arten unterteilt werden:[332]

- Beziehungszahlen
- Gliederungszahlen
- Mess- und Indexzahlen

Mit Hilfe von *Beziehungszahlen* können Zusammenhänge und Entwicklungen abgebildet werden, indem eine Grösse zu einer weiteren in Beziehung gesetzt wird. Typische Beziehungszahlen sind beispielsweise der Kapitalumschlag, die Liefertreue oder im Bereich des Qualitätsmanagements die Reklamationsrate. Wird eine Grösse G in mehrere Teilgrössen T unterteilt und wiederum zur Ausgangsgrösse in Beziehung gesetzt, so spricht man von *Gliederungszahlen*, die häufig in Prozent angegeben werden:[333]

$$K = \frac{T \times 100}{G}$$

Messzahlen dienen dem Entwicklungsvergleich, indem sie die Veränderung bestimmter betrieblicher Daten aufzeigen. Messzahlen werden gebildet, indem eine Basiszahl B aus der Zahlenreihe gewählt und mit 100 Punkten gleichgesetzt wird. Alle weiteren Zahlen Z, werden auf diese Basiszahl bezogen:[334]

$$K = \frac{Z \times 100}{B}$$

Werden mehrere Zeitreihen berücksichtigt, spricht man von *Indexzahlen* als spezielle Messzahlen.[335] Aufgrund der zentralen Bedeutung von Kennzahlen, sollen aus der Vielzahl von Anforderungen an Kennzahlen einige wesentliche ausgewählt und in Tabelle 3.14 beschrieben werden.[336]

[332] Vgl. Hartung, Elpelt und Klösener (1998), S. 55-60; Bohley (1996), S. 29; Mutscheller (1996), S. 35ff.

[333] Vgl. Hartung et al. (1998), S. 55; Siegwart (1998), S. 6-10; Meyer (1994), S. 3f.

[334] Vgl. Siegwart (1998), S. 9f.

[335] Vgl. Mutscheller (1996), S. 37f.

[336] Vgl. z. B. Mutscheller (1996), S. 38f.; Gillies (1992), S. 45f.; Fries (1994), S. 95ff.; Bieri (1995), S. 40ff.; vgl. auch Anforderungen an Messgrössen in Seghezzi (1996), S. 38.

Anforderung	Beschreibung
Validität	Eine Kennzahl sollte das dem Messvorgang tatsächlich zugrundeliegende Ereignis erfassen.
Eindimensionalität	Eine Kennzahl soll lediglich ein zentrales Merkmal messen.
Objektivität	Die Kennzahl soll die Realität des betrachteten Ereignisses widerspiegeln und durch den Messenden und seine Einrichtung zur Messung weitgehend nicht beeinflussbar sein.
Stabilität / Präzision	Bei wiederholter Messung unter identischen Bedingungen und gleichen gegebenen Anfangsbedingungen muss das Messergebnis reproduzierbar sein.
Automatisierbarkeit	Nach Möglichkeit sollte die Kennzahl automatisch ermittelt werden können.
Sensitivität	Bei Veränderung des zugrundeliegenden Merkmals soll diese Änderung durch den Messwert repräsentiert sein.
Reaktionszeit	Die Kennzahl sollte die Veränderung des zugrundeliegenden Merkmals zeitnah ausdrücken.
Messaufwand	Der Aufwand zur Ermittlung der Kennzahl sollte in einem ökonomischen Verhältnis zum Nutzen stehen.
Verständlichkeit	Die Kennzahl und ihre Wirkungszusammenhänge sollen verständlich sein.

Tabelle 3.14: Zentrale Anforderungen an Kennzahlen

3.5.2 Bewertungsrahmen

Neben den Anforderungen an Kennzahlen existieren Ansprüche bezüglich des Datenqualitätsmanagements im allgemeinen und der Qualitätsplanung und -lenkung im besonderen. Diese sollen im folgenden erläutert und anschliessend zusammenfassend dargestellt werden.

In Abschnitt 3.1 wurden die verschiedenen *Qualitätssichtweisen* und die Unterscheidung zwischen Designqualität und Ausführungsqualität erörtert, die grundlegend für das Datenqualitätsmanagement sind. So beeinflusst die Qualitätsplanung die Designqualität, während die Qualitätslenkung auf die Ausführungsqualität Einfluss ausübt. Daher sind beim Datenqualitätsmanagement diese beiden Sichtweisen grundsätzlich zu unterscheiden. Wie sich gezeigt hat, existieren innerhalb eines Data-Warehouse-Systems verschiedene Anspruchsgruppen mit individuel-

len Qualitätsforderungen. Bei der Qualitätsplanung sind die *anspruchsgruppen- und anwendungsspezifischen* Qualitätsforderungen zu berücksichtigen und in eine Qualitätsspezifikation zu überführen. Allerdings können sich die Qualitätsforderungen im Zeitverlauf verändern, was durch eine *dynamische Anpassung der Qualitätsforderungen* zu beachten ist. Die Qualitätsspezifikation sollte *konkrete und detaillierte Sollvorgaben* enthalten, welche die Qualitätsforderungen für einzelne Komponenten in operationalisierbaren Aussagen ausdrücken und Anhaltspunkte zur Ursachenanalyse bieten. Zur Verständlichkeit dieser Aussagen müssen die Beziehungen zwischen den subjektiven Qualitätsforderungen in Form von Qualitätsmerkmalen und den operationalisierten Kennzahlen bekannt sein. Kennzahlen sollten auf unterschiedlichen *Aggregationsstufen* abgebildet sein und so ein geschlossenes Kennzahlensystem bilden. Dabei ist auf die *Verständlichkeit* der Qualitätsaussagen zu achten.

Aussagen über die Datenqualität sind durch Qualitätsprüfungen zu ermitteln. Die Qualitätsaussagen sind dann in entsprechender Form bei der *Datenverwendung* zur Qualitätsbeurteilung heranzuziehen. Hierbei sind eventuell anwendungsspezifische Besonderheiten zu berücksichtigen. Es sind Möglichkeiten zur *Aufnahme von erkannten Qualitätsmängeln* zu schaffen. Gegebenenfalls ist ein *Benachrichtigungssystem* zu implementieren, das bestimmte Personenkreise beim Überschreiten kritischer Qualitätswerte benachrichtigt. Zur weiteren *Analyse der Datenqualität* sind entsprechende Verfahren anzubieten. Bei der Qualitätsprüfung sind *objektive Qualitätsprüfungen* den subjektiven Qualitätsprüfungen vorzuziehen. Gerade aufgrund der im allgemein sehr grossen Datenmengen ist insbesondere ein *hoher Automatisierungsgrad* bei der Qualitätsprüfung notwendig. Dies bedingt die *Integration in die Metadatenverwaltung*, bei der die bisherige Architektur zu berücksichtigen ist und flexibel für zukünftige Änderungen gestaltet werden sollte. Zur Unterstützung der Tätigkeiten sollten entsprechende *Werkzeuge* zur Verfügung gestellt werden.

Das operative Datenqualitätsmanagement ist im Rahmen des oben dargestellten *ganzheitlichen proaktiven Datenqualitätsmanagements* zu verstehen, das insbesondere die umfassende Betrachtung der gesamten Datenversorgung eines Un-

ternehmens miteinbezieht.[337] So ist auch das operative Datenqualitätsmanagement eines Data-Warehouse-Systems nicht nur auf die zentrale Data-Warehouse-Datenbasis beschränkt, sondern betrachtet den gesamten Datenfluss, von der Datenentstehung bis zur Datenverwendung. Grundsätzlich sind die Ursachen mangelnder Datenqualität zu beheben und nur in Ausnahmefällen Datenbereinigungsmassnahmen anzuwenden. Hierzu sind *Methoden und Techniken* bereitzustellen sowie ein *Qualitätsmanagementsystem* zu etablieren. Insbesondere ist die Qualitätsplanung methodisch zu unterstützen sowie Ansätze und Techniken zur Qualitätsprüfung zur Verfügung zu stellen, wobei auch organisatorische Aspekte des Qualitätsmanagementsystems zu berücksichtigen sind. In diesem Zusammenhang ist auch ein Vorgehensmodell zur Einführung eines Datenqualitätsmanagements zu nennen.

Zusammenfassend sind in Tabelle 3.15 die zentralen Anforderungen an ein Datenqualitätsmanagement und insbesondere an die Qualitätsplanung und -lenkung aufgeführt, die so einen Bewertungsrahmen für die im nachfolgenden Abschnitt dargestellten Ansätze bilden.

3.6 Ausgewählte Ansätze zum operativen Datenqualitätsmanagement

Nachdem im obigen Abschnitt die Anforderungen an das Datenqualitätsmanagement erläutert wurden, sollen im folgenden einige ausgewählte Ansätze zur Erfassung und Modellierung von Datenqualitätsforderungen sowie zur Qualitätsbeurteilung der Datenqualität aufgeführt werden. Im Gegensatz zum Qualitätsmanagement des produzierenden Bereichs und bei Dienstleistungen haben sich bislang noch keine Qualitätsnormen für Datenqualität durchgesetzt. Bislang gibt es für das Qualitätsmanagement von Daten keine fundierten und etablierten Lösungen. Wie in Abschnitt 2.4.2.3 aufgeführt, findet zur Verbesserung der Datenqualität meist eine Datenbereinigung im Rahmen der Transferprozesse durch Beseitigung von syntaktischen und teilweise semantischen Heterogenitäten

[337] Vgl. auch ganzheitliche Sichtweise des Qualitätsmanagements im Rahmen des Informationsmanagements in Heinrich (1992), S. 83.

Kategorie	**Anforderung**
Allgemein	Ganzheitliches, proaktives DQM
	Methodische Unterstützung und Techniken
	Qualitätsmanagementsystem (organisatorische Aspekte)
	Qualitätssichten (Designqualität / Ausführungsqualität)
Qualitäts-planung	Anspruchsgruppen- und anwendungsspezifische Betrachtung von Qualitätsforderungen
	Dynamische Anpassung der Qualitätsforderungen
	Konkrete und detaillierte Sollvorgaben
Qualitäts-lenkung	Objektive Qualitätsprüfungen
	Angabe von Qualitätsaussagen bei der Datenverwendung
	Qualitätsaussagen auf unterschiedlichen Aggregationsstufen
	Verständliche Qualitätsaussagen
	Möglichkeit zur Aufnahme von erkannten Qualitätsmängeln (subjektiv)
	Benachrichtigungssystem
	Weitere Analysemöglichkeiten über die Datenqualität
Technisch	Integration in die Metadatenverwaltung
	Werkzeugunterstützung
	Hoher Automatisierungsgrad

Tabelle 3.15: Zentrale Anforderungen an ein Datenqualitätsmanagement, insbesondere der Datenqualitätsplanung und -lenkung

statt. Viele Ansätze fokussieren sich dabei auf die Zusammenführung isolierter Datenbestände, die Datenbereinigung und die Beseitigung von Duplikaten in Adressbeständen.[338] Es findet, wie die empirische Studie in Abschnitt 3.3 gezeigt hat, häufig kein ganzheitliches Datenqualitätsmanagement im obigen Sinne statt. Insbesondere ist die Qualitätsplanung und -überprüfung, sowie die Analyse von Ursachen mangelnder Datenqualität und deren kontinuierlichen Verbesserung bislang nicht methodisch etabliert. Vielmehr wird eine nachträgliche Qualitätsverbesserung und kein proaktives Datenqualitätsmanagement durchgeführt. Zwar beschäftigen sich zahlreiche Forschungsprojekte mit Fragestellungen im Bereich der Datenqualität,[339] jedoch finden sich nur beschränkt praktikable Lösungskonzepte und Ansätze.

Am MIT beschäftigt sich seit 1992 das Forschungsprojekt „Total Data Quality Management“ (TDQM) mit dem Thema Datenqualität und dem Ziel, eine theoretische Basis für das Datenqualitätsmanagement zu erarbeiten.[340] Der Ansatz basiert im wesentlichen auf dem anwendungsbezogenen Qualitätsbegriff und der Übertragung der Konzepte des Total Quality Managements auf Datenqualität. Zwar werden Ansätze zur Qualitätsplanung und zur -messung dargestellt, jedoch ist das Ziel eines umfassenden Datenqualitätsmanagements leider bislang noch nicht vollständig konkretisiert.

Ziel des europäischen Forschungsprojektes DWQ (Foundations of Data Warehouse Quality) war es, eine Grundlage für Data-Warehouse-Entwickler bereitzustellen, die es erlaubt, Qualitätsfaktoren in Datenmodellen, Datenstrukturen und Implementationstechniken zu integrieren.[341] Schwerpunkt der Arbeit bildete dabei der Entwicklungsprozess für Data-Warehouse-Systeme und dessen Repräsentation in einem Metadatenmodell, wobei Datenqualität als Teilaspekt betrachtet wurde. Der Ansatz basiert im wesentlichen auf der Betrachtung von Beziehungen zwischen den unterschiedlichen Datenmodellen in einem Data-Warehouse-System. Schwerpunktmässig wird dabei die Realisierung durch eine Abfragesprache und

[338] Vgl. Dasu, Johnson und Koutsofios (2000), S. 191.

[339] Vgl. z. B. in Eppler und Wittig (2000), S. 84ff.; Wang et al. (1995b), S. 623ff.

[340] Eine Zusammenfassung über die Forschungsergebnisse findet sich in Wang et al. (2001).

[341] Vgl. Jarke et al. (2000); Jarke et al. (1999), S. 229ff.; Jarke und Vassiliou (1997), S. 299ff. Das Forschungsprojekt wurde von 1996 bis 1999 durchgeführt.

die Integration in die Metadatenverwaltung untersucht.

Die Ansätze von ENGLISH und REDMAN stellen zwei bedeutende Konzepte aus dem nicht akademischen Bereich dar und sollen daher erwähnt werden. ENGLISH orientiert sich bei seinem Ansatz an dem aus dem Total Quality Management übertragenen Vorgehen und entwickelt hierzu eine Vorgehensbeschreibung zur Umsetzung eines qualitätsorientierten Informationsmanagements.[342] Wenngleich das Verfahren in zahlreichen Praxisprojekten Anwendung fand, ist ein theoretisch fundiertes, methodisches Vorgehen nicht durchgängig berücksichtigt. Der Ansatz von REDMAN basiert im wesentlichen auf der aus der Qualitätskontrolle bekannten statistischen Prozesskontrolle, wobei er hierzu den Datenfluss im Sinne eines Produktionsprozesses strukturiert und für einzelne Prozesse Qualitätsvorgaben festlegt.[343]

Ein Projekt, das eine weitgehend automatisierte Qualitätsprüfung bezweckt, ist das Projekt CARAVEL.[344] Allerdings wird dabei im wesentlichen auf die Identifizierung von Duplikaten fokussiert. Ein Konzept für das Datenqualitätsmanagement in Data-Warehouse-Systemen, welches weitgehend in die Richtung des hier vorgeschlagenen geht, stellt der Ansatz im Rahmen des Forschungsprojektes CLIQ (Data Cleansing mit intelligentem Qualitätsmanagement) dar.[345] Ziel ist es, Datenqualitätsmetriken auf verschiedenen Granularitätsstufen für Datenqualitätsmerkmale bereitzustellen und hierzu eine softwarebasierte Unterstützung anzubieten. Der Fokus der Arbeit bezieht sich insbesondere auf die Operationalisierung von Datenqualitätsmetriken für die zentrale Data-Warehouse-Datenbank. Schwerpunkt bilden dabei die Transformations- und Ladeprozesse im Vorfeld der Data-Warehouse-Datenbank, so dass eine Gesamtbetrachtung des Datenflusses noch zu berücksichtigen ist. So ist die Teilbetrachtung des Data-Warehouse-Systems und die sich daraus ergebenden Folgerungen[346] ein wesentlicher Unterschied zum hier vorgeschlagenen Konzept. Weiter ist, im Gegensatz zum hier beschriebenen Konzept, bei dem die Explikation von Qualitätsforderungen mitberücksichtigt wird,

[342] Vgl. English (1999).
[343] Vgl. Redman (1996).
[344] Vgl. Galhardas, Florescu, Shasha und Simon (2000).
[345] Vgl. Hinrichs (2001); Hinrichs und Aden (2001), Grimmer und Hinrichs (2001).
[346] Vgl. auch Abschnitt 2.4.2.5.

die Arbeit eher auf die technische Repräsentation der Datenqualitätsmessung in Datenbanken bezogen. Das hier vorgeschlagene Konzept bezweckt insbesondere ausdrucksstarke, für Endanwender interpretierbare Qualitätsaussagen bereitzustellen. Hierzu sind die granularen, technischen Qualitätsmessungen zu einem Kennzahlensystem zusammenzuführen. Aufgrund der eher technischen Orientierung des Projektes ist dies noch umzusetzen.

Im folgenden Abschnitt sollen exemplarisch Ansätze der Qualitätsplanung und Qualitätsprüfung beschrieben werden. Abschliessend werden zusammenfassend die Defizite bisheriger Ansätze und die sich daraus ergebenden Folgerungen dargestellt.

3.6.1 Erfassung und Modellierung von Datenqualitätsforderungen

Bei der Gestaltung von Informationssystemen wurde die Erfassung und Modellierung von Datenqualitätsforderungen bislang wenig beachtet.[347] Exemplarisch sollen drei wesentliche Ansätze aufgeführt werden. Der Ansatz von WANG et al. basiert auf der Repräsentation von Datenqualitätsforderungen in Entity-Relationship-Modellen und der Erweiterung relationaler Modelle um Qualitätsattribute.[348] Anhand eines vierstufigen Prozesses schlagen sie die Entwicklung eines auf Attributen basierenden Datenmodells vor. Dieses kann in relationale Modelle umgesetzt werden. Zunächst wird im ersten Schritt mit Hilfe der herkömmlichen Entity-Relationship-Modellierung ein anwendungsspezifisches Datenmodell entwickelt. Anschliessend werden Qualitätsparameter den Entitäts- und Beziehungstypen sowie den Attributen zugeordnet. Die subjektiven Qualitätsparameter werden dann in objektive Qualitätsindikatoren transformiert. In einem abschliessenden Schritt sollen die so erstellten anwendungsspezifischen Qualitätssichten in ein einheitliches Qualitätsschema integriert werden. Wenngleich das vorgeschlagene Vorgehen grundsätzlich geeignet erscheint, ist die Erfassung der subjektiven Qualitätskriterien und deren Überführung in objektive

[347] Vgl. Wang, Kon und Madnick (1993), S. 672.

[348] Vgl. Wang et al. (2001), S. 19ff.; Wang, Reddy und Kon (1995a); Storey und Wang (1998); Wang et al. (1993).

Qualitätsindikatoren bei zahlreichen Anwendern und komplexen Datenmodellen problematisch. Insbesondere ist die Aussagekraft des erweiterten Datenmodells fraglich. Zwar werden Qualitätsindikatoren im Modell erfasst, jedoch wird keine Aussage über deren Zielgrösse getroffen. Weiter ist die Kopplung zwischen den Qualitätsforderungen und der Qualitätsbeurteilung nicht explizit berücksichtigt.

Ein weiteres in der Literatur diskutiertes Konzept zur Erfassung von Datenqualitätsforderungen ist die aus dem Qualitätsmanagement bekannte Methode „Quality Function Deployment".[349] Der konzeptionelle Kern der Methode ist das „House of Quality", welches als Grundlage zur strukturierten Diskussion dient. Mit dieser Strukturierungshilfe werden subjektive Qualitätsforderungen erfasst und in Beziehung zu objektiven Zielgrössen einzelner Qualitätsmerkmale gesetzt. Neben diesem Ansatz wird beispielsweise im Projekt DWQ auch die Übertragung der aus dem Software-Engineering stammenden Methode „Goals-Question-Metrics" vorgeschlagen.[350] Basis bilden Qualitätsziele, deren Erreichung häufig nicht direkt messbar ist. Für jedes Ziel sollen dann Fragen abgeleitet werden, welche die Zielerreichung kontrollieren sollen. Aus den überprüfbaren Fragen sollen dann Qualitätsmetriken konkretisiert werden.

3.6.2 Qualitätsbeurteilung und Qualitätsprüfung

In der Literatur findet sich eine Vielzahl von Ansätzen zur Qualitätsbeurteilung von Daten.[351] Neben den berücksichtigten Datenqualitätsmerkmalen[352] unterscheiden sich die oben dargestellten Forschungsprojekte im Ansatz zur Qualitätsbeurteilung. Es finden sich sowohl subjektive Qualitätsbeurteilungen durch den Endanwender, Qualitätsprüfung durch die Analyse des Datenbestandes als auch prozessorientierte Ansätze, welche die Qualität anhand von Prozesseigenschaften betrachten. Im folgenden werden einige ausgewählte Ansätze dargestellt.

Das TDQM Forschungsprogramm verfolgt insbesondere die subjektive Qualitätsbeurteilung durch die Endanwender, welche durch eine Analyse des Datenbestan-

[349] Vgl. Redman (1996), S. 140ff.; Jarke et al. (2000), S. 142f.; Helfert und Radon (2000), S. 117-119.
[350] Vgl. Jarke et al. (2000), S. 114ff.; Bobrowski, Marre und Yankelevich (1999), S. 119ff.
[351] Vgl. z. B. Ansätze in Wang et al. (1995b), S. 633f.
[352] Vgl. Abschnitt 3.2.

des anhand von Plausibilitätsregeln ergänzt wird. Hierzu wird eine Qualitätsbeurteilung durch drei Elemente vorgeschlagen.[353] Die Qualitätseinschätzung findet durch eine umfassende Befragung der Endanwender nach deren subjektiver Qualitätseinschätzung statt. Als weiteres Element wird eine objektive, anwendungsunabhängige sowie eine anwendungsabhängige Qualitätsanalyse des Datenbestandes anhand von Plausibilitätsregeln vorgeschlagen.

Kern des Ansatzes des Forschungsprojektes DWQ ist das in Abbildung 3.13 dargestellte Qualitätsmodell für Data-Warehouse-Systeme.[354] Die „messbaren Objekte“, als zentraler Bestandteil des Modells, beziehen sich auf die Komponenten eines Data-Warehouse-Systems in Form von Elementen der Schemata und Datenhaltungssysteme sowie der Transferprozesse. Der obere Teil des Modells beschreibt die Qualitätsziele, von denen Qualitätsfragen abgeleitet werden (Mitte des Modells). Der untere Teil bezieht sich auf die Qualitätsmessung der „messbaren Objekte“. Indem das Qualitätsmodell in die Metadatenverwaltung integriert ist, sind alle qualitätsrelevanten Angaben in dieser verfügbar. Durch eine zweistufige Instanziierung des Qualitätsmodells können sowohl Muster als auch deren konkrete, anwendungsspezifische Realisierung in der Metadatenverwaltung abgelegt werden.

Beim Forschungsprojekt CLIQ wird die Datenqualitätsmessung in die Extraktions-, Transformations- und Ladephase im Vorfeld der zentralen Data-Warehouse-Datenbank eingebettet.[355] Die Qualitätsprüfung findet sowohl in der temporären Datenhaltung im Transformationsbereich als auch in der zentralen Data-Warehouse-Datenbank statt. Im wesentlichen handelt es sich um Prüfungen anhand von Vorgaben im Datenschema und von festgelegten Integritätsbedingungen und Erfahrungswerten. Im Gegensatz zum hier vorgeschlagenen Konzept sind allerdings die Betrachtungen jeweils auf einzelne Datenbestände bezogen, so dass Beziehungen zwischen verschiedenen Datenbeständen im gesamten Data-Warehouse-System nicht direkt betrachtet werden. Weiter werden die Ausführung der Transferprozesse und die Datenvolumen nicht explizit berücksichtigt.[356]

[353] Vgl. Huang et al. (1999), S. 60ff.; Wang, Strong, Kahn und Lee (1999).
[354] Vgl. Jarke et al. (2000), S. 123ff.
[355] Vgl. Hinrichs (2001); Grimmer und Hinrichs (2001).
[356] Vgl. in diesem Zusammenhang Abschnitt 4.2.2.2 und 4.2.2.1.4.

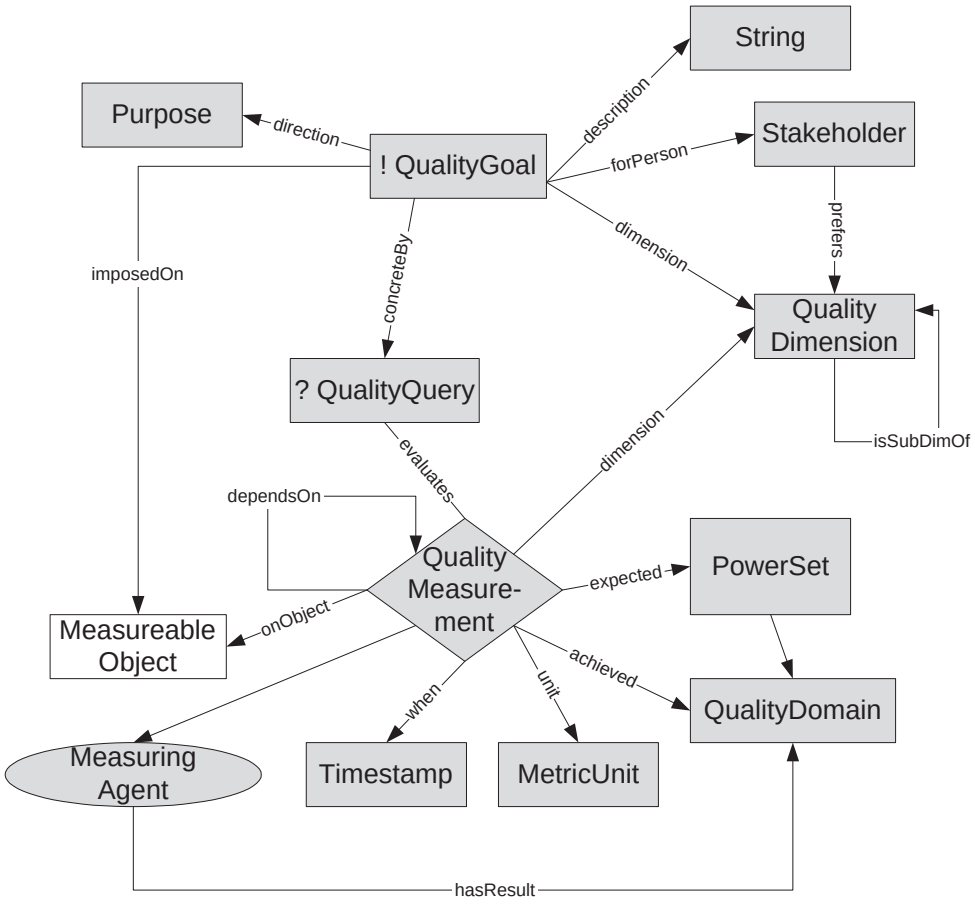


Abbildung 3.13: Qualitätsmodell des Forschungsprojekts DWQ (In Anlehnung an Jarke et al. (2000), S. 148)

3.7 Schlussfolgerungen

Im folgenden Abschnitt soll ein Überblick über die zentralen Erkenntnisse dieses Kapitels dargestellt werden. Zunächst wurde in Abschnitt 3.1 der allgemeine Qualitätsbegriff erläutert und die sich daraus ergebenden verschiedenen Qualitätsauffassungen dargestellt. Aufbauend auf diesen wurden dann die Qualitätssichten der Designqualität und Ausführungsqualität abgeleitet. In Abschnitt 3.2 wurden Ansätze zur Konkretisierung des komplexen Begriffes der Datenqualität untersucht. Es zeigte sich, dass der Begriff durch eine Vielzahl von Qualitätsmerkmalen beschrieben wird. Einzelne Merkmale hängen in ihrer Bedeutung und Intensität erheblich vom Anwendungskontext ab. Es sind Überschneidungen vorhanden und es existiert keine einheitliche Beschreibung oder Definition. Allgemein werden

Korrektheit, Vollständigkeit, Konsistenz und Aktualität aufgeführt.

Im Rahmen der empirischen Untersuchung wurde ein Kriterienkatalog entwickelt, der eine für die Arbeit geeignete Basis zur Konkretisierung des Begriffes der Datenqualität ermöglicht. Dabei wurde sowohl die Unterscheidung zwischen Design- und Ausführungsqualität als auch die These von anwender- und anwendungsspezifischen Qualitätsforderungen unterstützt. Als wesentliche Datenqualitätsmerkmale wurde die Widerspruchsfreiheit innerhalb und zwischen Datenbeständen und die zeitliche Konsistenz der Datenwerte genannt. Darüber hinaus ist Vollständigkeit und Korrektheit sowie die Repräsentation fehlender Werte wichtig. Es zeigte sich auch, dass das Thema der Datenqualität in Data-Warehouse-Systemen sehr grosse Relevanz besitzt. Bei fast allen Unternehmen stellt Datenqualität ein Problem dar. Bislang existieren bei den Unternehmen einzelne Ansätze zur Qualitätsprüfung im Rahmen der Datentransformation und der Analyse des Datenbestandes. Die umfassende Betrachtung des Datenflusses hinsichtlich qualitativer Zielsetzungen und die Umsetzung von organisatorischen Regelungen scheint derzeit nicht bei allen Unternehmen etabliert zu sein. Werden die Qualitätsprüfungen näher betrachtet, sind subjektive Qualitätsbeurteilungen durch den Datenverwender, Prüfungen anhand von Integritätsbedingungen und Vergleiche zwischen Datenbeständen als auch eine Analyse der Transferprozesse über Systemprotokolle zu finden. Beim Umgang mit den Datenqualitätsproblemen werden vor allem korrigierende Massnahmen (Datenbereinigung) eingesetzt. Die Beseitigung der Ursachen (z. B. in den operativen Vorsystemen) wird allerdings als geeignetere Lösungsmöglichkeit erkannt. Es fehlt hier im wesentlichen an organisatorischen Konzepten, methodischer Unterstützung und eine Möglichkeit, Qualitätsvorgaben durch konkrete Sollgrössen anzugeben und diese durch objektive Qualitätsprüfungen zu kontrollieren. Aufgrund der meist nur wagen Einschätzung der Datenqualität durch Endanwender wäre eine Qualitätsangabe über die Daten bei deren Verwendung wichtig.

Aufbauend auf diesen Erkenntnissen wurde in Abschnitt 3.4 ein Konzept für ein proaktives Datenqualitätsmanagement erarbeitet. Zunächst wurde der aktuelle Stand des Qualitätswesens anhand des Total Quality Managements und des integrierten Qualitätsmanagements nach SEGHEZZI dargestellt. Die drei zentralen

Bereiche,

- die Verpflichtung des Managements, Datenqualität als Ziel festzulegen und daraus ein durchgängiges Zielsystem abzuleiten, das die Sicherstellung einer unternehmensweiten Datenqualität berücksichtigt,
- die Etablierung eines Qualitätsmanagementsystems in die bisherige Organisationsstruktur und
- die Unterstützung durch Methoden, Verfahren und Werkzeuge

bilden den Kern des proaktiven Datenqualitätsmanagements. Auf der operativen Ebene wurden die Aufgaben der Qualitätsplanung, der Qualitätslenkung, der Qualitätssicherung sowie der kontinuierlichen Qualitätsverbesserung identifiziert. In Abschnitt 3.4.3 wurden dann die Qualitätsplanung und Qualitätslenkung als Kernaufgaben erörtert und ein hierarchisches Regelkreismodell entwickelt. Daraus wurden dann Anforderungen an das operative Datenqualitätsmanagement abgeleitet und in Abschnitt 3.6 mit aktuellen Forschungsprojekten verglichen. Eine vorgenommene Einschätzung der Abdeckung des Erfüllungsgrades dieser Anforderungen ist in Tabelle 3.16 dargestellt. Es zeigt sich, dass keines der Projekte alle die hier identifizierten Anforderungen abdeckt, wenngleich Einzelanforderungen von verschiedenen Projekten angegangen werden. Auffallend ist, dass Benachrichtigungssysteme und Eingabemöglichkeiten zur Aufnahme erkannter Qualitätsmängel bislang kaum vorhanden sind. Weiter gibt es eine Vielzahl von Ansätzen zur Qualitätsprüfung, wobei derzeit leider noch keine zufriedenstellenden und praktikablen Lösungskonzepte existieren. Interessant ist auch, dass ein Teil der Projekte sich eher auf Aspekte des Managements fokussiert (TDQM, ENGLISH, REDMAN) und ein anderer sich von der eher technischen Seite dem Problem annähert (DWQ, CLIQ).

Wenngleich die im nächsten Kapitel dargestellte Fallstudie nicht alle Aspekte des proaktiven Datenqualitätsmanagements umfassend betrachtet, werden dennoch wesentliche Aufgabenbereiche erörtert. Insbesondere wird die Integration des Datenqualitätsmanagements in die Metadatenverwaltung, konkrete Datenqualitätsforderungen im Rahmen der Qualitätsplanung sowie deren Prüfung durch die Qualitätslenkung dargestellt.

Anforderungen	TDQM	English	Redman	DWQ	CLIQ
Ganzheitliches, proaktives DQM	●	●	⊙	·	·
Methodische Unterstützung und Techniken	⊙	⊙	⊙	·	·
Qualitätsmanagementsystem	·	⊙	⊙	○	○
Qualitätssichten	·	●	●	·	·
Anspruchsgruppen- und anwendungsspezifische Betrachtung	●	●	○	●	○
Dynamische Anpassung der Qualitätsforderungen	-	·	·	●	·
Konkrete und detaillierte Sollvorgaben	-	⊙	⊙	●	●
Objektive Qualitätsprüfungen	⊙	⊙	⊙	●	●
Angabe von Qualitätsaussagen bei der Datenverwendung	○	○	○	○	○
Qualitätsaussagen unterschiedlicher Aggregationsstufen	-	○	⊙	○	●
Verständliche Qualitätsaussagen	●	⊙	⊙	○	⊙
Möglichkeit zur Aufnahme von erkannten Qualitätsmängeln	·	○	·	·	·
Benachrichtigungssystem	-	-	-	·	·
Weitere Analysemöglichkeiten über die Datenqualität	⊙	⊙	○	⊙	○
Integration in die Metadatenverwaltung	·	·	·	●	●
Werkzeugunterstützung	●	⊙	○	●	●
Hoher Automatisierungsgrad	○	·	○	●	●

Anmerkungen:

- ● berücksichtigt
- ⊙ teilweise berücksichtigt
- ○ ansatzweise berücksichtigt
- · nicht direkt berücksichtigt
- - keine Angaben

Tabelle 3.16: Erfüllungsgrad ausgewählter Forschungsprojekte

Kapitel 4

Ansatz für ein operatives Datenqualitätsmanagement

Im folgenden wird das im Rahmen eines Forschungsprojektes bei einer Schweizer Universalbank erarbeitete Konzept eines operativen Datenqualitätsmanagements zur Planung und Messung der Datenqualität beschrieben. In Abschnitt 4.1 wird zunächst der Handlungsbedarf anhand der Ausgangssituation und den Rahmenbedingungen verdeutlicht sowie die zugrundeliegende Architektur des Data-Warehouse-Systems dargestellt. Anschliessend wird in Abschnitt 4.2 das innerhalb der Projektarbeit in enger Zusammenarbeit entwickelte Konzept eines in die Metadatenverwaltung integrierten Datenqualitätssystems aufgezeigt. Ausgehend von projektspezifischen Datenqualitätsforderungen werden anschliessend exemplarisch die Möglichkeiten für eine Qualitätsspezifikation und deren Prüfung erläutert. Zusammenfassend werden in Abschnitt 4.2.3 Möglichkeiten zur Auswertung der Datenqualität innerhalb eines Data-Warehouse-Systems erörtert.

4.1 Rahmenbedingungen und Data-Warehouse-System

4.1.1 Architektur

Das Data-Warehouse-System der Universalbank ist durch eine einheitliche Architektur gekennzeichnet, die in den letzten Jahren sukzessive aufgebaut wurde.[357] Die bestehenden isolierten Anwendungssysteme waren für verschiedenen Hardware- und Betriebssystemplattformen implementiert. Diese wurden in eine einheitliche Gesamtarchitektur überführt, welche in Abbildung 4.1 dargestellt

[357] Vgl. im folgenden Wegener (2000), S. 81ff.

ist. Kennzeichen des Konzepts ist eine komponentenbasierte Data-Warehouse-Architektur, die insbesondere über eine integrierte Metadatenverwaltung und ein einheitliches Metadatenmodell verfügt. Im wesentlichen finden sich darin die in Abschnitt 2.4.2.1 dargestellten Elemente eines Data-Warehouse-Systems wieder, so dass hier lediglich einige Besonderheiten erläutert werden.

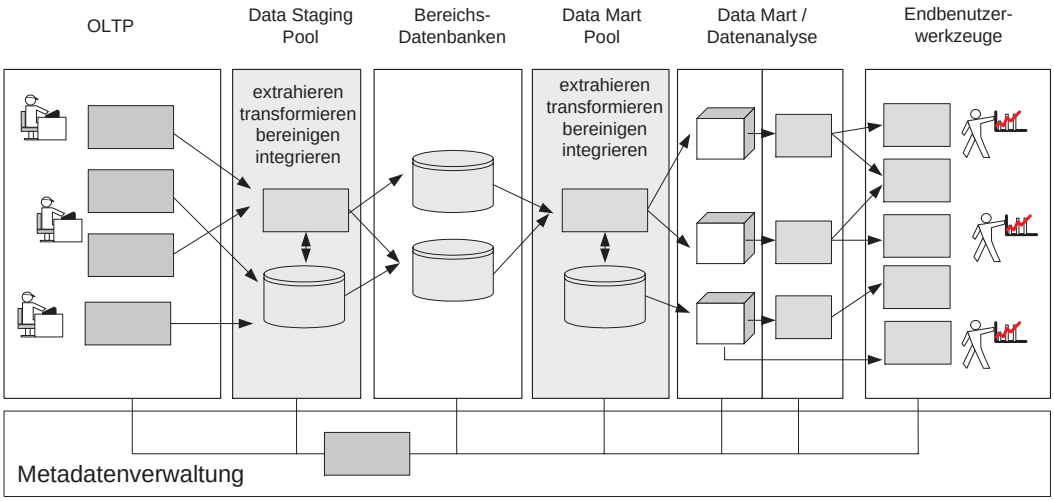


Abbildung 4.1: Architektur des Data-Warehouse-Systems (In Anlehnung an Wegener (2000), S. 82)

Die zahlreichen Quellsysteme stellen Daten in einem definierten Format in die als Feeder bezeichneten Datenbereitstellungsdateien ein. Eine auf Tagesbasis stattfindende Überführung der Daten in die Feeder-Dateien soll eine einheitliche Aktualität und zeitliche Konsistenz der Datenbasis gewährleisten. Die eingestellten Daten werden dann extrahiert und erste, einfache Qualitätssicherungs- und Integrationsoperationen darauf vorgenommen. Dabei sind allerdings zeitliche Restriktionen für die Transferprozesse zu beachten, so dass lediglich einfache Qualitätsprüfungen durchgeführt werden können. Anschliessend werden die Daten in sogenannte Bereichsdatenbanken (BDB) geladen. Bereichsdatenbanken stellen Teilsichten auf einen bankfachlich als weitgehend isoliert betrachteten Bereich dar. Derzeit existieren Bereichsdatenbanken für

- das Kreditwesen (Credit & Risk),
- die Kontenverwaltung (Accounting),

- den Zahlungsverkehr (Payments) und
- den Vertrieb (Marketing).

Neben den vier Bereichsdatenbanken existiert ein zentraler Bereich, der die Informationsobjekte Partner, Produkte und Portfolio enthält. Partner sind ehemalige, derzeitige und zukünftige Geschäftspartner der Bank. Die über Portfolios zu den Partnern in Beziehung stehenden Produkte repräsentieren die von der Bank angebotenen Leistungen. Eine lose gekoppelte Datenbank enthält zudem zentrale Stammdaten, wie beispielsweise Länderkennungen, Währungssorten sowie die zur Verfügung stehenden Kundentypen. Abfragen sollen grundsätzlich nicht direkt auf den Bereichsdatenbanken erfolgen, sondern auf den für einen bestimmten Analysezweck optimierten Data Marts. Als letzte Stufe stellen Präsentationssysteme Daten und Analysen für die Datenverwender bereit. Da sich bislang noch kein Standard für die Metadatenverwaltung in kommerziellen Produkten etabliert hat, werden bislang proprietäre Metadatenverwaltungs-Werkzeuge eingesetzt und über eine eigene Kontrollumgebung integriert.

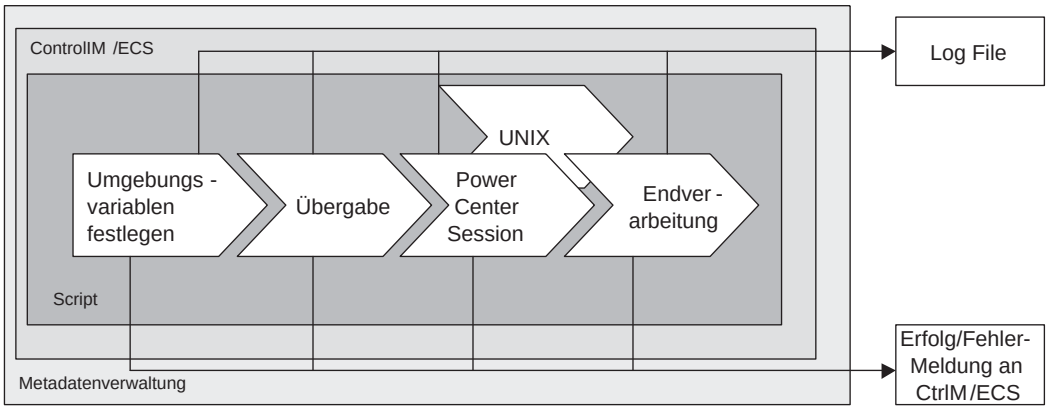


Abbildung 4.2: Steuerung des Transformationsprozesses (In Anlehung an Kaminsky (2000), S. 14)

Die zwischen den Feeder-Dateien und den Bereichsdatenbanken stattfindenden Transformationsprozesse werden durch das Werkzeug „POWERCENTER“ von INFORMATICA oder in UNIX ablaufenden Transformationsroutinen ausgeführt. Eine zentrale Kontrollumgebung (ControlM/ECS) initiiert die spezifizierten Trans-

formationsprozesse und übernimmt deren zeitliche Steuerung. Der grobe Ablauf ist in Abbildung 4.2 dargestellt. Ausgelöst durch festgelegte Ereignisse und Regeln, wird ein Skript aufgerufen, welches dann die für den Transformationsprozess notwendigen Programmaufrufe startet und überwacht. Es werden zunächst Kontrollinformationen aus einer Metadatenbank gelesen, notwendige Variablen definiert und einige Programmroutinen ausgeführt. Anschliessend wird der eigentliche Transformationsprozess in Form einer POWERCENTER-Session oder eines UNIX-Programms gestartet.[358] Abschliessende Routinen archivieren die Protokolldateien und geben Statusinformationen über die Transformationsläufe an die Steuerumgebung zurück.

Die für die Transformationsprozesse notwendigen Kontrolldaten werden in einer Datenbank abgelegt,[359] deren Datenschema in Abbildung 4.3 dargestellt ist. Zentrale Entität ist `SESSION`. Sie repräsentiert einen isoliert betrachteten Transformationsprozess. Die aufzurufende POWERCENTER-Session ist durch das Attribut `SESSION_NAME` bezeichnet und führt die einzelnen Transformationsschritte durch. Das Attribut `WIEDERHOLBAR` mit Wahrheitswerten gibt an, ob ein identischer Transformationsprozess (d. h. Session mit identischem `GEPLANT_ZEITPUNKT`) mehrmals aufgerufen werden kann ohne inkonsistente Datenbankzustände hervorzurufen. Die Entität `DATENQUELLEN` und das Attribut `QUELLE_NAME` im besonderen spezifiziert die Dateinamen der Quelldateien. Die Zieldatenstrukturen sind in der Entität `ZIELDATEN` im Attribut `ZIEL_NAME` abgelegt. Informationen über den Ablauf einzelner Transformationsprozesse sind in der Entität `SESSION_LOG` zusammengefasst. Diese sind eindeutig über eine `SESSION_ID` und dem in der Kontrollumgebung geplanten `GEPLANT_ZEITPUNKT` zu identifizieren.[360] `START_ZEITPUNKT` gibt den Startzeitpunkt des Transformationsprozesses an. Zur Verbindung der hier verwendeten Metadaten mit den in POWERCENTER bereits vorhandenen Metadaten wird der im Attribut `PC_SESSION_ZEITPUNKT` abgelegte Zeitstempel benötigt. Das Attribut `END_ZEITPUNKT` gibt den Endzeitpunkt des Transformationsprozesses an.

[358] Im Rahmen der Arbeit werden vereinfachend lediglich POWERCENTER-Sessions betrachtet.

[359] Das hier dargestellte Datenschema enthält aus Vereinfachungsgründen lediglich die wesentlichen Entitäten und Attribute des tatsächlichen Schemas.

[360] Die Identifikation einzelner Transferprozesse ist zur Qualitätsprüfung der Aktualität wichtig; vgl. auch Abschnitt 4.2.2.2 und Abschnitt 4.2.3.

Die möglichen Ausprägungen {`run, completed, failed`} des Attributs `STATUS` geben Aufschluss über den jeweiligen Zustand einzelner Transformationsprozesse.

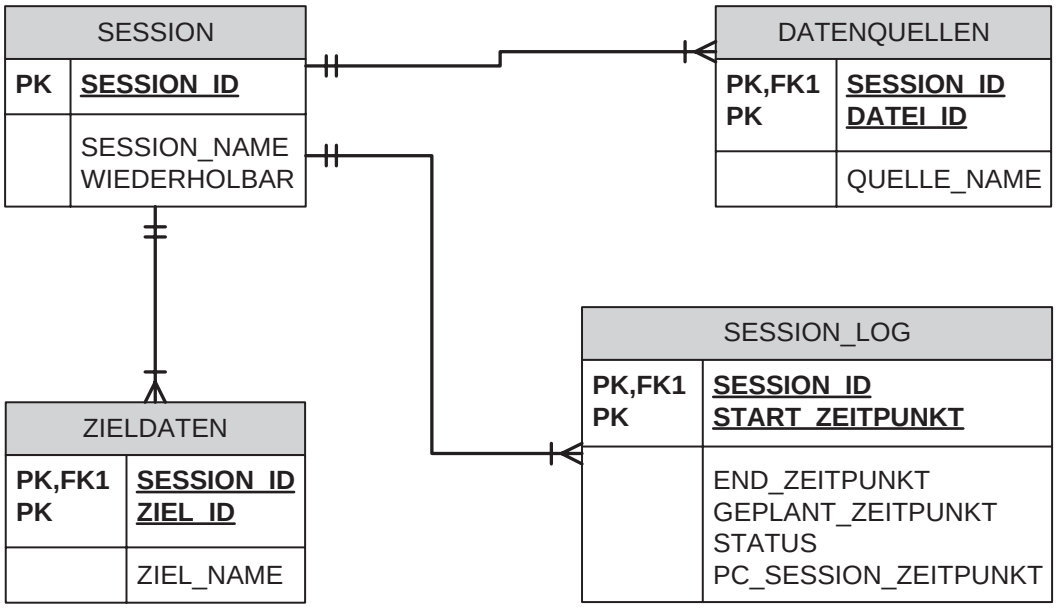


Abbildung 4.3: Medatdatenschema für den ETL-Prozess (Stark vereinfachte Darstellung des Medatdatenschemas der Universalbank)

Für die Bereichsdatenbank „Produkte, Partner, Portfolio“ werden bereits einige technische Qualitätsüberprüfungen im Rahmen des Transformationsvorgangs durchgeführt und in einer Datei protokolliert. Damit die Performanz des Transformationsprozesses nicht zusätzlich beeinträchtigt wird, werden lediglich einige wenige Qualitätsüberprüfungen vorgenommen. Komplexere Überprüfungen können im Anschluss an die Transformationsprozesse auf den Datenbeständen stattfinden. Die Qualitätsprüfungen im Rahmen des Transformationsprozesses konzentrieren sich vor allem auf eher technische Prüfungen einzelner Attributwerte und technisch orientierten Integritätsbedingungen. So werden Schlüsselbedingungen, Pflichtfeldern und Wertebereiche sowie Datenformate überprüft. Die jeweiligen Fehlerarten haben entsprechende Reaktionen zur Folge und werden einer Fehlerkategorie (Fehler oder Warnung) zugeordnet. Beispielhaft sind einige dieser Qualitätsüberprüfungen in Tabelle 4.1 aufgeführt.

Die im Rahmen der Transformation ermittelten und abgelegten Daten geben Auf-

Beschreibung der Operation	**Reaktion auf Verletzung**	**Fehler-Kategorie**
Einfügen eines Schlüsselduplikats	Datensatz nicht laden	Fehler
Aktualisieren eines Datensatzes, dessen Schlüsselwert nicht vorhanden ist	Einfügen eines neuen Datensatzes	Warnung
Löschen eines Datensatzes, dessen Schlüsselwert nicht vorhanden ist	Operation nicht möglich	Fehler
Einfügen eines Nullwertes in ein Schlüsselfeld	Daten nicht laden	Fehler
Einfügen eines Nullwertes in ein Feld „gültig bis“	Einfügen des Defaultwertes (31.12.9999)	Warnung
Ein Wert im falschen Format	Einfügen eines Nullwertes	Warnung
Einfügen eines ungültigen Wertes für ein Feld „gültig ab “	Daten nicht laden	Fehler
Einfügen eines ungültigen Wertes für ein Feld „gültig bis “	Laden des Defaultwertes (31.12.9999)	Warnung
Laden eins Attributwertes, für das keine Quelle existiert	Einfügen eines Nullwertes	Warnung

Tabelle 4.1: Qualitätsüberprüfungen im Transformationsprozess und deren Reaktionen bei Verletzung der Bedingung

schluss über den Ablauf geplanter Transformationsprozesse und die Verletzung einiger wichtiger Integritätsbedingungen.[361] Aufbauend auf diesen Daten werden derzeit bereits einige Aussagen über die Qualität der Datenbasis bereitgestellt.[362] Als Indikator für die Aktualität stehen Aussagen über die letzte erfolgreiche Prozessausführung zur Verfügung. Angaben über die Anzahl der geladenen und abgelehnten Datensätze dienen als Indikator für die Vollständigkeit der Datenbasis. Die hierbei gemachten Erfahrungen zeigen, dass zeitliche Aspekte im Vergleich zur Glaubwürdigkeit und Interpretierbarkeit der Daten leichter zu bewerten sind. Im Anhang B sind einige Beispiele dieser Angaben dargestellt. Die Aussagen sind bislang allerdings technisch orientiert und werden der Forderung nach verständlichen und verwenderorientierten Qualitätsaussagen nicht gerecht.

[361] Vgl. in diesem Zusammenhang auch Abschnitt 4.2.3.
[362] Vgl. auch Helfert et al. (2001), S. 13.

Zur Interpretation der zur Verfügung gestellten Qualitätsaussagen ist Fachwissen über die Transformationsprozesse und die zugrundeliegenden Datenmodelle notwendig. Häufig sind die Datenverwender jedoch wenig technisch fokussiert und können daher die Qualitätsaussagen nicht ausreichend interpretieren. Von den Datennutzern wird dagegen eine aussagekräftige Angabe über die Datenqualität mit geringer Komplexität erwartet. Hierzu ist eine hohe Aggregation unterschiedlicher Qualitätskennzahlen zu einer komprimierten Qualitätsaussage notwendig.

Ausgehend von den Rahmenbedingungen werden im folgenden Abschnitt einige zentrale Problemfelder der Datenqualität bei der Universalbank dargestellt. Anschliessend wird das Konzept eines metadatenbasierten Datenqualitätssystems erläutert und insbesondere auf die Prüfung der Ausführungsqualität eingegangen. Die Ergebnisse der Qualitätsprüfungen können dann durch Kennzahlen den Datenverwendern als verständliche Datenqualitätsaussagen zur Verfügung gestellt werden.

4.1.2 Zentrale Problemfelder im Bereich der Datenqualität

Erste Diskussionen mit den Projektverantwortlichen der entsprechenden Bereiche zeigten bereits zu Beginn, dass Handlungsbedarf bezüglich der Datenqualität besteht. Nicht selten wird, insbesondere von den Datennutzern, die Vermutung falscher oder unbrauchbarer Daten geäussert. Aufgrund der bislang nicht ermittelten Datenqualität wird von den Datenverwendern häufig die Frage nach der Nutzbarkeit der Daten gestellt. Wünschenswert wäre eine verständliche Kennzeichnung der Daten und Berichte, anhand derer die Dateneignung einfach zu erkennen ist (z. B. farbliche Kennzeichnung für „gut“, „ausreichend“ und „nicht zu verwenden“). Eine Analyse der in einem Bereich vorhandenen Fehlerprotokolle veranschaulicht einige zentrale Datenqualitätsprobleme.[363] Diese sind exemplarisch in Tabelle 4.2 aufgeführt. Vor allem sind Interpretationsprobleme der Daten, uneinheitliche Definitionen, nicht eingehaltene Datenspezifikationen, unvollständige und inkorrekte Datenwerte und Datenbeziehungen sowie dynamisch

[363] Die Auswertung fand im Jahr 2001 statt und basiert auf 51 Entitäten, wovon lediglich bei 17 Entitäten ausführliche Testprotokolle vorlagen.

schwankende Qualitätsniveaus auffallend. Eine detaillierte Analyse der Datenqualitätsforderungen ausgewählter Projekte findet sich in Abschnitt 4.2.1. Resultat dieser Problematiken in bezug auf die Datenqualität ist geringes Vertrauen in die bereitgestellten Daten des Data-Warehouse-Systems. Folge sind ein Rückgang in der Nutzung des Data-Warehouse-Systems und fehlerhafte Entscheidungen.

4.2 Ein metadatenbasiertes Datenqualitätssystem

Neben der Bereitstellung geeigneter Qualitätsaussagen für die Datenverwender, beabsichtigt das Projekt die kontinuierliche Verbesserung der Datenqualität mittels eines proaktiven Datenqualitätsmanagements. Langfristiges Ziel ist dabei die umfassende Qualitätsbewirtschaftung des gesamten Datenflusses, von der Datenentstehung in den operativen Systemen bis hin zur Datenverwendung. Aufgrund sowohl organisatorischer als auch zeitlicher Restriktionen musste die Projektarbeit exemplarisch auf ein Teil des in Abschnitt 3.4 erarbeiteten Konzeptes begrenzt werden. Zunächst wurde eine Fokussierung auf die als wichtig erachteten Datenqualitätskriterien vorgenommen und durch eine Kriterienliste konkretisiert.[364] Aufgrund bestehender organisatorischer Beschränkungen und den schwer abschätzbaren zeitlichen Risiken konzentrierte sich die Betrachtung auf die den operativen Systemen nachgelagerten Elemente des Data-Warehouse-Systems. In dieser ersten Projektphase wurde die Betrachtung der operativen Systemen explizit ausgeschlossen. Durch strukturierte Interviews wurden anschliessend für einzelne Fachbereiche beispielhaft deren Qualitätsforderungen erhoben und Möglichkeiten zur Datenqualitätsmessung diskutiert. Wenngleich lediglich Teilaspekte erörtert wurden, ist konzeptionell der Ansatz des proaktiven Datenqualitätsmanagements berücksichtigt.[365] In einem Folgeschritt können so die Ergebnisse der ersten Projektphase zu einem ganzheitlichen Datenqualitätsmanagement erweitert werden.

Ausgehend von der oben beschriebenen Architektur des Data-Warehouse-Sys-

[364] Vgl. Kriterienliste in Abschnitt 3.3.

[365] Eine Diskussion des Konzepts fand im Rahmen eines Workshops mit Partnerunternehmen des Kompetenzzentrums CC DW2 am 6. und 7. Juni 2001 in Ermatingen / Schloss Wolfsberg (Schweiz) mit 30 Teilnehmern statt. Vgl. hierzu Helfert und von Maur (2001), S. 67f. und Helfert et al. (2001), S. 18ff.

Exemplarische Datenqualitätsprobleme
Ein Vergleich der Auswertungen zweier Systeme ergab, dass bei zahlreichen Krediten der Betrag in beiden Systemen nicht übereinstimmt.[a] Eine weitere Analyse zu einem späteren Zeitpunkt ergab dagegen bei weitaus weniger Krediten Differenzen. Die Ursachen konnten nach einer Detailanalyse geklärt werden.
Beim Vergleich von Daten zwischen zwei Systemen auf Attributebene wurden die Daten als • gut (identische Datenwerte, nicht Null), • kritisch (nicht identische Datenwerte, nicht Null), • kritisch-verschlechtert (nur im Altsystem vorhanden) und • akzeptabel (nur im Neusystem vorhanden) eingeteilt. In Abhängigkeit einzelner Attribute ergaben sich unterschiedliche Übereinstimmungsgrade. Zahlreiche Attribute wurden als kritisch und kritisch-verschlechtert eingestuft. Teilweise konnten die Differenzen auf technische Probleme (z. B. Repräsentation von Sonderzeichen) zurückgeführt werden. Beim gleichen Test zu einem späteren Zeitpunkt ergab sich ein ähnliches Bild. Es zeigte sich jedoch eine nochmalige, deutliche Qualitätsverschlechterung zweier Attribute.
Aufgrund unterschiedlicher Definitionen bestimmter Schlüsselattribute ist eine exakte Überprüfung des Geburts- und Gründungsdatum zwischen unterschiedlichen Systemen nicht möglich.
Bei einer Überprüfung der Attribute gemäss Datendefinition konnten die eingetragenen Werte nicht sinnvoll interpretiert werden.
Attributwerte einer Entität existieren lediglich auf der Entität „Konto“und nicht wie in der Datenbeschreibung angegeben auf der Entität „CIF“.
Häufig weisen Kredite, denen ein Konto zugeordnet sein sollte, keinen Eintrag in der entsprechenden Relation auf.

[a] Eine Saldodifferenz von 5 SFR ist dabei akzeptabel.

Tabelle 4.2: Analyse von Fehlerprotokollen

tems, kommt der Metadatenverwaltung bei der Ermittlung geeigneter Datenqualitätskennzahlen besondere Bedeutung zu. Durch die Metadaten sind die Transformationsprozesse festgelegt und werden durch die Metadatenverwaltung ausgelöst sowie überwacht. Dabei entstehen Daten über den Ablauf des Datenflusses. Zur Beurteilung der Datenqualität sind, neben der Nutzung von Metadaten, auch manuelle und werkzeugunterstützte Analysen von Datenqualitätsexperten und das Urteil des Datenverwenders einzubeziehen. Das so entwickelte Konzept, das alle qualitätsrelevanten Daten entlang des Datenflusses ermittelt, ist in Abbildung 4.4 dargestellt.

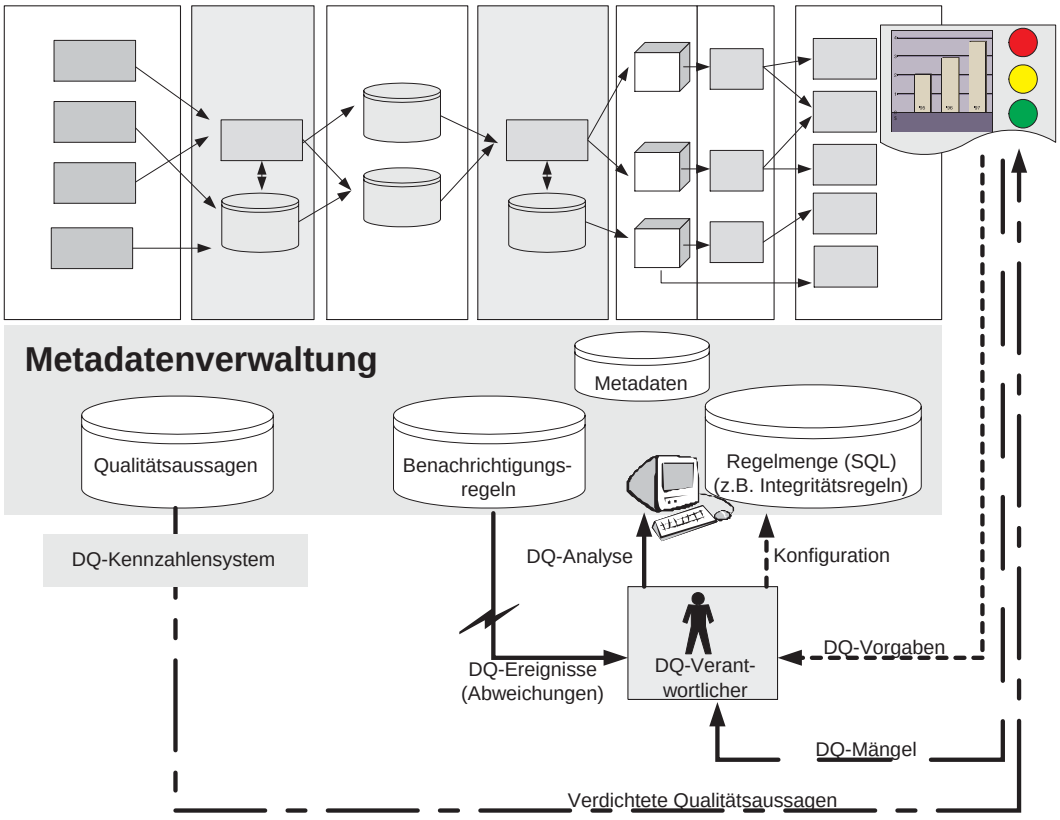


Abbildung 4.4: Konzept eines metadatenbasierten Datenqualitätssystems (Eigene Darstellung)

Kern des Ansatzes ist ein in die Metadatenverwaltung integriertes Datenqualitätssystem. Hier werden alle relevanten Qualitätsaussagen verwaltet. Eine *Regelmenge*, in der bestimmte Regeln zur Prüfung der Datenqualität hinterlegt sind, ist wesentlicher Bestandteil des Systems. Neben den zu berücksichtigenden Mes-

sobjekten in den Regelbedingungen und den Zielwerten, werden hier auch deren Ausführungszeitpunkte spezifiziert. Möglichkeiten zur Regelbildung werden in Abschnitt 4.2.2 erörtert. Die sich aus den Qualitätsprüfungen ergebenden *Messwerte* werden gespeichert und sind für Qualitätsaussagen verfügbar. Die Bildung von Qualitätskennzahlen wird in Abschnitt 4.2.3 diskutiert. Prinzipiell können die Messwerte durch Erweiterung der Datenmodelle in den bereits vorhandenen Datenhaltungssystemen oder in einem separaten Datenhaltungssystem verwaltet werden. Aufgrund der Flexibilität wird hier die getrennte Datenhaltung für Qualitätswerte bevorzugt. Eine weitere Komponente des Datenqualitätssystems sind *Benachrichtigungsregeln*. Hier werden Regeln und Ereignisse zur Benachrichtigung entsprechender Personen oder Personengruppen festgelegt. Qualitätsverantwortliche können dann bei Unterschreiten bestimmter Qualitätswerte auf elektronischem Wege (z. B. Email), mobilem Telefon (z. B. SMS) oder sonstigen Kommunikationskanälen über das Ereignis in Kenntnis gesetzt werden. Sie können dann problemadäquate Massnahmen einleiten und so im Sinne einer Qualitätslenkung regelnd in den Prozess eingreifen.

4.2.1 Datenqualitätsforderungen

Ausgehend von den in Abschnitt 3.2 erarbeiteten allgemeinen Datenqualitätsmerkmalen sollen im folgenden die für ausgewählte Projekte spezifischen Datenqualitätsforderungen dargestellt werden. Exemplarisch wurden hierfür strukturierte Interviews mit Fachbereichsvertretern von drei Projekten auf Ebene der Data Marts durchgeführt:

Projekt A im Bereich Zahlungsverkehr.

Projekt B im Bereich Kredit-Controlling.

Projekt C im Bereich Werbung und Kampagnenmanagement.

Die Interviews fanden für die einzelnen Projekte gesondert statt, wobei die Dauer der Interviews in der Regel auf eine Stunde angesetzt wurde. Neben den Fachbereichsvertretern nahm ein Bereichsvertreter des Architekturkonzeptes teil.

Zunächst wurden das Konzept des Datenqualitätsmanagements und wichtige Datenqualitätsmerkmale vorgestellt. Von wichtigen Datenqualitätsmängeln ausgehend wurden dann im Rahmen einer Diskussion grobe projektspezifische Anforderungen an die Datenqualität erörtert. Dabei konnten bereits erste Ansätze für eine eventuelle Qualitätsprüfung identifiziert werden.

Bei *Projekt A* fällt das Gesamtsystem durch den Ausfall einzelner Systemkomponenten im Schnitt zwei bis drei Stunden pro Tag aus. Problematisch ist dabei, wenn Komponenten unerkannt ausfallen und daher Transformationsprozesse und Berechnungsroutinen nicht korrekt ausgeführt werden. Eine wichtige Aufgabe ist daher das Erkennen des Ausfalls einzelner Systemkomponenten und das Nicht-Ausführen von Transferprozessen. Aus diesem Grund wird von Seiten der Anwender ein weitgehend automatisierter Verfügbarkeitstest für einzelne Systemkomponenten gefordert. Die Glaubwürdigkeit der Daten wird derzeit subjektiv durch die Datenanwender eingeschätzt („Bauchgefühl“). Für die subjektive Einschätzung der Datenqualität durch den Datenanwender ist die Datenherkunft sehr wichtig. Mit Hilfe dieser kann der Datenanwender aufgrund seines fachlichen Wissens und der Erfahrung die Datenqualität einschätzen. Vergleiche zu anderen Systemen sind aufgrund unterschiedlicher Datendefinitionen bislang nur eingeschränkt möglich.[366] Da logische Zusammenhänge zwischen den Daten derzeit nicht berücksichtigt sind, wäre eine Überprüfung anhand von Integritätsbedingungen äusserst nützlich. Aufgrund der Datenverwendung sind die bereitgestellten Daten nicht sehr zeitkritisch. Eine zwei bis drei Tage verzögerte Aktualität der Daten ist ausreichend und wird bereits weitgehend erfüllt. Erfahrungsgemäss ist die Vollständigkeit der Daten mit einer Verzögerung von zwei bis drei Wochen weitgehend sichergestellt und unproblematisch. Werden die Probleme im Bereich der Datenmodellierung untersucht, ergibt sich folgendes Bild. Bislang sind keine einheitliche Standards zur Datenmodellierung etabliert, so dass häufig für semantisch gleiche Attribute unterschiedliche syntaktische Definitionen verwendet werden. Standards zur syntaktischen Spezifikation sowie der Festlegung von Wertebereichen und Pflichtfeldern sind nicht vorgegeben. Zwischen unterschiedlichen Datenmodellen bestehen syntaktische und semantische Heterogenität, was zu Un-

[366] Beispielsweise werden in verschiedenen operativen Systemen unterschiedliche Attribute erfasst oder unterschiedliche Datenquellen für die Datenlieferung herangezogen.

klarheiten und Widersprüchen führt. Ein einheitliches, konsistentes Datenmodell sowie abgestimmte Modellierungsstandards wären hier zur Erhöhung der Datenqualität erforderlich.

Bei *Projekt B* stehen insbesondere Probleme in bezug auf die Glaubwürdigkeit der Daten im Vordergrund. Im wesentlichen handelt es sich hierbei um unvollständige Datenwerte einzelner Datenobjekte und deren Beziehungen. Ebenso finden sich widersprüchliche Daten in einzelnen oder zwischen verschiedenen Datenbeständen. Qualitätsprüfungen auf syntaktische Korrektheit und Einhaltung der Wertebereiche finden bislang nicht oder nur unzureichend statt. Eine hohe Glaubwürdigkeit der Daten zeichnet sich bei Projekt B insbesondere durch widerspruchsfreie Datenbestände aus.[367] Bislang wird die Glaubwürdigkeit der Daten subjektiv durch die Datenverwender geschätzt, wofür die Datenherkunft wichtig ist. Sollten jedoch ausreichende Qualitätsaussagen über die Daten zur Verfügung gestellt werden, erscheint die Datenherkunft nicht mehr von hoher Bedeutung zu sein. Neben der Glaubwürdigkeit wird Handlungsbedarf bei den Zugriffszeiten genannt, die häufig zu lange sind. Die derzeitige monatliche Aktualisierung wird für das Projekt B als ausreichend angesehen. Wichtiger als eine höhere Aktualisierungsfrequenz ist die zeitliche Zuverlässigkeit der Datenbereitstellung. Aufgrund umfangreicher Datenbeschreibungen und etablierter Modellierungsstandards haben Qualitätsaspekte der Datenmodellierung lediglich einen geringen Stellenwert. Wenngleich weitere Richtlinien zur Datenbeschreibung nützlich wären, ist die Semantik der Daten im Projekt B ausreichend gut beschrieben.

Im Unterschied zu den bisherigen Projekten sind die direkten Fachbereichsvertreter beim *Projekt C* nicht die letztendlichen Datenanwender. Während die anderen Projekte die Daten für Entscheidungsprozesse direkt nutzen, werden hier Daten an Datenabnehmer gemäss Liefervereinbarungen bereitgestellt. Standen bei den oben genannten Projekten Qualitätsforderungen bezüglich der Glaubwürdigkeit im Vordergrund, sind hier vor allem Verfügbarkeit und die syntaktische Korrektheit wichtig. Die Datenabnehmer spezifizieren ihre Datenbedürfnisse und legen so die Qualitätsforderungen fest. Durch das Projekt C werden dann die geforder-

[367] Beispielsweise im Vergleich zum Buchhaltungssystem oder zu den bisher eingesetzten Altsystemen.

ten Daten aus der Datenbasis generiert und in die Zieldatenstrukturen überführt. Aufgrund der Anforderungen der Datenabnehmer ist die Systemverfügbarkeit und die Aktualität der Daten (nach Möglichkeit tagesaktuell) besonders wichtig. Die Übereinstimmung der Daten mit den festgelegten Zieldatenstrukturen in bezug auf Syntax und Semantik ist dabei wichtig. Da keine direkte Datennutzung im Rahmen des Projektes stattfindet, wird die semantische Korrektheit und Widerspruchsfreiheit zu anderen Systemen wenig beachtet. Aufgrund des fehlenden Anwendungsbezugs ist die Glaubwürdigkeit der Daten von geringer Bedeutung.

Die geäusserten Qualitätsforderungen repräsentieren die jeweiligen Datenqualitätsprobleme der einzelnen Projekte. Während sich das Projekt C auf die Erfüllung geforderter Datenspezifikationen hinsichtlich der Verfügbarkeit, Aktualität und Konformität zu den Zieldatenstrukturen konzentriert, ist für die anderen Projekte die Glaubwürdigkeit und Zuverlässigkeit wichtig. Die Glaubwürdigkeit wird insbesondere durch widerspruchsfreie Daten charakterisiert. Widerspruchsfreiheit bezieht sich dabei auf die Konsistenz in und zwischen Datenbeständen, zu allgemein gültigen Geschäftsregeln und Plausibilitätsbedingungen, allgemeine Integritätsbedingungen und Wertebereichsdefinitionen. Weiter ist die Vollständigkeit und die syntaktische Korrektheit bezüglich des Datenmodells wichtig. Der zeitliche Bezug, insbesondere die Aktualität der Daten, ist bei den Projekten weitgehend gewährleistet. Allerdings ist die Lieferzuverlässigkeit der geplanten Aktualisierungsfrequenzen wichtig. Allen Projekten gemein sind Probleme bezüglich der Verfügbarkeit und Systemzuverlässigkeit. Der Datenzugriff dauert häufig zu lange oder ist wegen unzureichender Systemverfügbarkeit nicht möglich. Auffallend ist, dass Zugriffsrechte ein geringes Problem darstellen. Die Interpretierbarkeit der Daten und Einheitlichkeit der Datendefinitionen ist bei ausreichend beschriebenen Datenmodellen und etablierten Modellierungsstandards ein geringes Problem. Soweit diese noch nicht vorhanden sind, werden hier Richtlinien zur Datenmodellierung und -beschreibung gefordert. Interessant erscheint weiter, dass die Nützlichkeit der im Datenmodell spezifizierten Daten bislang unproblematisch erscheint. In Tabelle 4.3 sind zusammenfassend die wesentlichen Qualitätsforderungen der einzelnen Projekte sowie die erörterten Anforderungen an das Datenqualitätsmanagement dargestellt.

Merkmal		Projekt A	Projekt B	Projekt C	DQM und Qualitätsprüfung
Schema	Interpretierbarkeit	Einheitliches Datenmodell und Standards	Bereits sehr gute Datenbeschreibung	Einheitliche Datenbeschreibung (Semantik) wichtig	Einheitliche, übergreifende Modellierungsstandards und -methoden
	Nützlichkeit			Angeforderte Daten müssen im Datenmodell erfasst sein	Prüfung der Entitäten und Attribute durch Diskussionen mit den Endanwendern
Werte	Glaubwürdigkeit	Widerspruchsfreiheit innerhalb und zwischen Datenbeständen sowie im Zeitverlauf, Vollständigkeit, Wertebereichskonformität, syntaktische Korrektheit, (Datenherkunft)		Vollständigkeit bzgl. der angeforderten Daten, kein Vergleich unterschiedlicher Auswertungen oder zwischen Systemen (Widerspruchsfreiheit), Einhaltung der Syntax, Wertebereichsverletzungen ein geringes Problem	Prüfung bzgl. der Einhaltung des Datenschemas (Syntax, Pflichtfelder) und der Widerspruchsfreiheit (Integritätsbedingungen)
	Zeitlicher Bezug	2 bis 3 Tage aktuelle Daten sind ausreichend	monatliche Aktualisierung ist ausreichend	Sicherstellung der zeitlichen Konsistenz zwischen verschiedenen Datenquellen, hohe Aktualität	Sicherstellen der geplanten Aktualisierungsfrequenz (Lieferzuverlässigkeit), um so zeitlich konsistente Datenbestände zu gewährleisten
	Verfügbarkeit	Häufiger Systemausfall einzelner Komponenten	Zugriffszeit zu lange	Verfügbarkeit wird vom Datenabnehmer vorgegeben, deren Einhaltung sehr wichtig ist	Prüfung auf Verfügbarkeit einzelner Systemkomponenten und Transferprozesse

Tabelle 4.3: Qualitätsforderungen ausgewählter Projekte

Sowohl die empirische Untersuchung als auch die Diskussionen im Rahmen der Projektarbeit zeigten folgende Erkenntnisse: Zunächst ist zwischen Design- und Ausführungsqualität zu unterscheiden. Während die Designqualität eines Data-Warehouse-Systems durch die Datenschemata beeinflusst wird, wird die Ausführungsqualität durch die konkreten Datenwerte und die Funktionsfähigkeit der Systemkomponenten bestimmt.

Für die auf das Datenschema bezogenen Datenqualitätsmerkmale sind insbesondere semantische Aspekte wichtig. Es sind einheitliche und standardisierte, syntaktische Beschreibungen sowie widerspruchsfreie, aussagekräftige und genau Datendefinitionen gefordert. Diese können durch organisatorische Regelungen sowie einheitliche Modellierungsstandards und -methoden weitgehend sichergestellt werden. Ergebnis der Datenmodellierung sollten konsistente Datenschemata auf unterschiedlichen Beschreibungs- und Architekturebenen sein.[368] Die Nützlichkeit der in Datenschemata spezifizierten Daten kann durch Diskussionen mit den jeweiligen Endanwendern geprüft und so Mängel erkannt werden. Zu kontinuierlichen Verbesserung ist hier ein iterativer Abgleich zwischen Anforderungen und Spezifikation zu initiieren.

Im Gegensatz zur Designqualität scheint die Sicherstellung und Verbesserung der Ausführungsqualität in gegenwärtigen Data-Warehouse-Systemen problematischer zu sein. Bei der Ausführungsqualität ist als wesentliches Datenqualitätsmerkmal die Glaubwürdigkeit der Datenwerte zu nennen. Die Datenwerte sollten in bezug auf die Spezifikation im Datenschema korrekt und vollständig sein. Kennzeichen glaubwürdiger Daten ist deren Widerspruchsfreiheit. Daten sollten im wesentlichen widerspruchsfrei zu Integritätsbedingungen sein, die insbesondere Abhängigkeiten innerhalb und zwischen Datenbeständen, im Zeitablauf sowie zu syntaktischen und semantischen Festlegungen berücksichtigen. Bislang wird die Glaubwürdigkeit der Datenwerte weitgehend durch den Endanwender subjektiv eingeschätzt. Dabei ist die Kenntnis der Datenherkunft wichtig. Die Sicherstellung eines zeitlich konsistenten Datenbestandes ist im Vergleich zur Forderung nach Aktualität der Datenwerte problematischer. Für die zeitliche Konsistenz ist

[368] Vgl. in diesem Zusammenhang die Betrachtungsebenen für Data-Warehouse-Systeme in Abschnitt 2.4.2.5.

die erfolgreiche Ausführung der geplanten Transferprozesse wesentlich (Lieferzuverlässigkeit). Weiter scheint die Systemverfügbarkeit einzelner Softwarekomponenten ein Problem zu sein, das allerdings häufig nicht als Datenqualitätsproblem sondern als Problem der Softwarequalität erkannt wird.

Aufgrund der hohen Relevanz im Rahmen der empirischen Untersuchung und der Projektarbeit werden im folgenden einige Ansätze zur Messung der Ausführungsqualität in Data-Warehouse-Systemen erörtert. Der Schwerpunkt liegt dabei auf den Qualitätsmerkmalen Glaubwürdigkeit, Aktualität und zeitliche Konsistenz der Datenwerte sowie der Möglichkeit einer automatischen Qualitätsprüfung im Rahmen der Metadatenverwaltung.

4.2.2 Prüfung der Ausführungsqualität

Die verschiedenen Qualitätssichten und die daraus abgeleitete Unterscheidung zwischen Design- und Ausführungsqualität wurden in Abschnitt 3.1 erläutert. Während *Designqualität* die Erfassung von Qualitätsforderungen und deren Umsetzung in eine Spezifikation umfasst, bezieht sich Ausführungsqualität auf die Einhaltung und Erreichung der in der Spezifikation genannten Qualitätsforderungen. Zur Beurteilung der Designqualität sind neben Funktionsspezifikationen der Anwendungssysteme, die hier nicht betrachtet werden sollen, Datenschemata relevant. Insbesondere ist das globale, auf die Data-Warehouse-Datenbasis bezogene Datenschema wichtig.[369] Als zentrales Bezugsobjekt sind die mit ihm in Beziehung stehenden Datenschemata konsistent zu halten. In den Datenschemata wird neben den Datenobjekten deren Syntax und Semantik festgelegt, welche so die Interpetierbarkeit und Nützlichkeit als zentrale Datenqualitätsmerkmale beeinflussen. Konkrete Datenwerte werden durch den Betrieb des Data-Warehouse-Systems erzeugt, wobei sich dies auf die Qualitätssichtweise der *Ausführungsqualität* bezieht. Glaubwürdigkeit, zeitlicher Bezug, Nützlichkeit und Verfügbarkeit wurden in Kapitel 3 als wesentliche Qualitätsmerkmale der Datenwerte erarbeitet.

Zur Messung der Datenqualität werden in der Literatur bislang lediglich Einzelaspekte durch Messansätze konkretisiert. Aufbauend auf diesen Erkenntnissen

[369] Vgl. Abschnitt 2.4.2.5.

soll im folgenden exemplarisch ein Ansatz zur Messung der Ausführungsqualität erläutert werden. Aufgrund der fallspezifischen Restriktionen wird eine Eingrenzung auf die von den Datenverwendern als wichtig eingestuften Qualitätskriterien vorgenommen. Insbesondere werden in den folgenden Abschnitten die Qualitätsmerkmale Glaubwürdigkeit, Aktualität und zeitliche Konsistenz der Datenwerte betrachtet.

Zur Prüfung der *Glaubwürdigkeit* von Datenwerten kommen prinzipiell neben der empirischen Prüfung und der subjektiven Qualitätseinschätzung durch den Datenverwender die Qualitätsprüfung durch Integritätsbedingungen in Frage. Hierfür sind geeignete Integritätsbedingungen auszuwählen sowie eine geeignete Beschreibungsform zu finden. Dieser Aspekt wird in Abschnitt 4.2.2.1 betrachtet. Neben Glaubwürdigkeit ist der *zeitliche Bezug* in Form von Aktualität und zeitlicher Konsistenz eine zentrale Einflussgrösse der Datenqualität. In einem Data-Warehouse-System werden Daten zwischen den verschiedenen Datenhaltungssystemen durch Transferprozesse überführt. Es sind notwendige Daten zu extrahieren, syntaktisch und semantisch zu transformieren und in eine Datenbank zu überführen. Transferprozesse beeinflussen dabei nicht nur inhaltliche Aspekte der Datenqualität, sondern auch insbesondere zeitliche Merkmale. Abschliessend werden die einzelnen Qualitätsforderungen in Abschnitt 4.2.3 zusammengeführt und eine Möglichkeit der *Qualitätsauswertung* erörtert.

4.2.2.1 Glaubwürdigkeit

Datenverwender haben aufgrund ihrer Erfahrungen und dem fachlichem Wissen bestimmte Vorstellungen über die Daten und deren Zusammenhänge. Hierzu zählen beispielsweise Vergleichswerte mit anderen Datenquellen, Erfahrungen über Datenverteilungen und Datenmuster, Geschäftsregeln und Wertebereiche sowie allgemein übliche Datenformate. Differieren diese Vorstellungen mit den zur Verfügung gestellten Daten, wirken sie für den Datenverwender unglaubwürdig. Eine Möglichkeit wäre daher die manuelle Prüfung der Daten durch den Datenverwender. Allerdings sind die Datenvolumen in einem Data-Warehouse-System im allgemeinen für eine manuelle Datenanalyse zu gross und die Strukturen zu

komplex, so dass eine weitgehend automatische Prüfung notwendig wird. Auch die empirische Prüfung der Datenwerte ist, wenn überhaupt, dann nur auf Stichprobenbasis möglich.[370] Aufgrund der Erkenntnis, dass empirisch korrekte Daten in sich widerspruchsfrei sind,[371] stellt Widerspruchsfreiheit einen guten Indikator für empirische Korrektheit dar. Während empirische Korrektheit auf die Übereinstimmung mit den Werten der realen Welt abzielt, bezieht sich die Widerspruchsfreiheit auf die Konsistenz des Datenbestandes. Diese kann in Form von Integritätsbedingungen angegeben und überprüft werden. Daher ist die Prüfung auf Widerspruchsfreiheit im Gegensatz zur empirischen Prüfung und zur subjektiven Qualitätseinschätzung durch den Datenverwender relativ einfach durchzuführen. Dieser Ansatz soll im folgenden für Data-Warehouse-Systeme weiter verfolgt werden.

Ziel des Ansatzes ist es, die „allgemeinen Erfahrungen“ und das anwendungsspezifische „Wissen“ der Fachexperten über die Daten zu explizieren und so einen Indikator für die Glaubwürdigkeit von Datenbeständen zu entwickeln. Dieses „Wissen“ ist allerdings meist implizit vorhanden und nicht formal ausgedrückt.[372] Es sind daher Techniken und Modelle zur Explikation und Beschreibung des Wissens notwendig. Wenngleich Konsistenzbedingungen für verschiedene Datenbanksysteme möglich sind, sollen die folgenden Ausführungen aufgrund der hohen Relevanz relationaler Datenbanksystemen auf diese beschränkt werden.

In Abschnitt 2.3.1 werden bereits einige allgemein übliche Integritätsbedingungen für relationale Datenbanken erläutert. Einfache Datenqualitätsprobleme, wie fehlende Werte, Datenformate und Fremdschlüsselbeziehungen, sind durch diese zu erkennen und zu prüfen. Allerdings stellen diese lediglich eine Teilmenge aller möglichen Bedingungen dar. Es existieren noch zahlreiche weitere Integritätsbedingungen,[373] wie beispielsweise:

- Ein Wert steht mit anderen Werten in Beziehung (z. B. Summe der Kontensalden einer Kundengruppe in System A ist [grösser, gleich oder kleiner] als

[370] Vgl. Busatto (2000), S. 130.
[371] Vgl. Mandke und Nayar (1998), S. 234
[372] Vgl. Busatto (2000), S. 132.
[373] Vgl. Elmasri und Navathe (1994), S. 149.

die Summe der Guthaben einer Kundengruppe in System B abzüglich der Summe der Kredite einer Kundengruppe in System C).

- Die Anzahl der Tupel einer Relation steht in Beziehung zur Anzahl der Tupel einer anderen Relation (z. B. Anzahl der Portfolios [grösser, gleich oder kleiner] der Anzahl der Kunden).
- Ein Wert ist zeitinvariant (z. B. Anzahl der Kantone).
- Ein Attributwert zeigt im Zeitablauf ein ähnliches Verhalten wie ein zweiter Attributwert (z. B. durchschnittliches Kreditvolumen verhält sich linear zur Anzahl der Kunden).

Dabei liegt eine zentrale Annahme zugrunde.[374] Es wird angenommen, dass bestimmte Eigenschaften der Daten zeitinvariant sind und so eine geeignete Vergleichsmöglichkeit für verschiedener Datenbestände im Zeitablauf bieten. Es sind daher zunächst geeignete, zeitinvariante Eigenschaften von Datenbeständen zu identifizieren und in Bedingungen zu spezifizieren.[375] Die Datenbestände, die zur Ermittlung der qualitätsrelevanten Eigenschaften dienen, werden als Qualitätsreferenzdaten bezeichnet. Diese sollten qualitativ hochwertig sein, wobei im allgemeinen die Qualität von häufig genutzten Datenbeständen höher ist. Zur Auswahl der Qualitätsreferenzdaten und der Ermittlung von charakteristischen Eigenschaften sind Fachexperten zu involvieren. Sie besitzen das notwendige fachspezifische Wissen über die Datenbestände. Die charakteristischen Eigenschaften der Qualitätsreferenzdaten können dann in Modellen und Bedingungen abgebildet und zur Prüfung zukünftiger Daten herangezogen werden. In den folgenden Abschnitten werden, neben einigen zentralen Beschreibungsmöglichkeiten der deskriptiven Statistik, Methoden des Data Minings für komplexe Datenbeziehungen und -strukturen untersucht.

Allerdings unterliegen Daten in betrieblichen Systemen dynamischen Veränderungen und Schwankungen aufgrund:[376]

[374] Vgl. Milek et al. (2001), S. 193f.

[375] Das methodische Vorgehen und die Techniken zur Erstellung sollen nicht primärer Fokus der Arbeit sein.

[376] Vgl. Milek et al. (2001), S. 191.

- individuellen Kundenverhaltens,
- des Marktverhaltens und
- saisonaler Entwicklungen.

Diese Schwankungen sind in den Datenbeständen enthalten und wirken sich auf deren Eigenschaften im Zeitablauf aus. Aus diesem Grund sind die ermittelten Eigenschaften der Qualitätsreferenzdaten mit denen neuer Datenbestände nicht direkt verglichen werden. Es sind die jeweiligen Schwankungen zu berücksichtigen. In Abhängigkeit der Schwankungsart sind dabei unterschiedliche Möglichkeiten denkbar.[377] Zunächst können relativ kleine Schwankungen in grossen Datenbeständen durch Datenaggregationen ausgeglichen und so unterdrückt werden. Es bietet sich eine Aggregation über mehrere Datentupel im Zustandsraum und / oder über die Zeit an. Bei grösseren oder zyklischen Schwankungsbreiten können diese durch stochastische Modelle abgebildet werden. Hierzu zählen beispielsweise Wahrscheinlichkeitsverteilungen über das Kundenverhalten sowie Zeitreihen zur Abbildung von saisonalen Zyklen und Marktschwankungen. Auf Grundlage dieser Modelle können dann Plausibilitätsintervalle ermittelt werden, in denen sich die Datenausprägungen wahrscheinlich befinden.

Neben den sich aus der realen Welt ergebenden Schwankungen sind in betrieblichen Daten auch qualitätsrelevante Störgrössen enthalten.[378] Ziel ist nun, die qualitätsrelevanten Störgrössen zu isolieren und durch geeignete Kennzahlen auszudrücken. Allerdings sollten die Kennzahlen einerseits die reellen Schwankungen berücksichtigen, aber andererseits dennoch sensitiv auf qualitätsrelevante Störungen reagieren. Diese Differenzierung zwischen qualitätsrelevanten Störgrössen und reellen Schwankungen ist in der Praxis äusserst schwierig. Aus diesem Grund sollten bei Ermittlung von entsprechenden Modellen und charakteristischen Eigenschaften Fachexperten und Datenanalysten einbezogen werden. Das grundsätzliche Prinzip zur Prüfung der Glaubwürdigkeit von Daten ist in Abbildung 4.5 dargestellt und wird in den folgenden Abschnitten erläutert.

[377] Vgl. Milek et al. (2001), S. 193.
[378] Vgl. auch Milek et al. (2001), S. 191.; vgl. Dasu und Johnson (1999), S. 90.

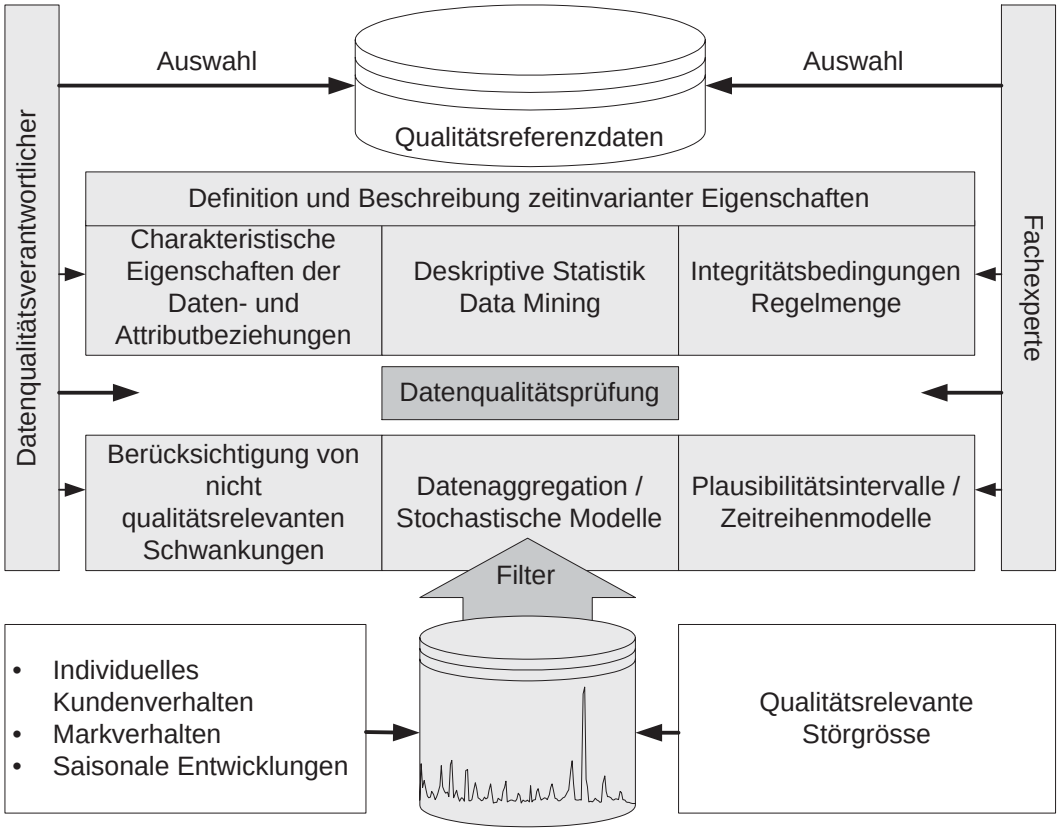


Abbildung 4.5: Prinzip und Vorgehen zur Prüfung der Glaubwürdigkeit (Eigene Darstellung)

4.2.2.1.1 Ausgewählte charakteristische Eigenschaften Die deskriptive Statistik dient zunächst zur beschreibenden und graphischen Aufbereitung von Daten.[379] Sie umfasst graphische Darstellungen wie Diagramme und Verlaufskurven sowie Mittelwerte und Streuung als Kenngrössen der Datenverteilung. Mit ihrer Hilfe lassen sich gegebene Datenbestände durch bestimmte Eigenschaften beschreiben und analysieren. Datenbeschreibungen in der deskriptiven Statistik erfolgen häufig in Form von Grafiken, wie beispielsweise Stabdiagramme, Kreisdiagramme und Histogramme. Wie in Abbildung 4.6 exemplarisch gezeigt, lassen sich durch einfache Datenanalysen anhand der Verteilung der Wertausprägungen relativ leicht unregelmässige Daten finden. So lässt sich beispielsweise beim Attribut „Kanton“ der unterdurchschnittlich geringe Anteil der Ausprägung „St. Gal-

[379] Vgl. im folgenden z. B. Fahrmeier et al. (1997), S. 11f.; Ester und Sander (2000), S. 30ff; Bohley (1996), S. 117ff.

len“ in einer Menge von Tupeln einfach erkennen. Beim Attribut „Geburtsdatum“ lässt sich die grosse Abweichung der Ausprägung „01.01.1800“ vom Zentrum der Daten feststellen. Solche unregelmässigen Datenausprägungen werden allgemein als Datenausreisser bezeichnet.

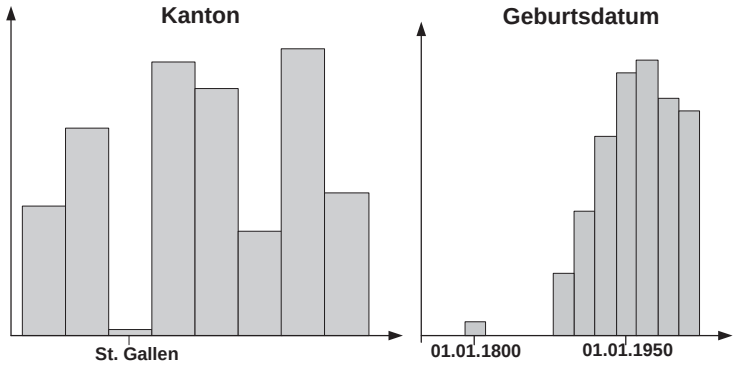


Abbildung 4.6: Beispiel für univariate Analysen (In Anlehnung an Seidl (2001), S. 19)

Allerdings benötigen die visuellen Analysen Expertenwissen über allgemein übliche und anwendungsspezifische Datenverteilungen. Eine automatische Qualitätsprüfung mit Hilfe graphischer Darstellungen ist daher zunächst nicht möglich.[380] Neben der graphischen Visualisierung bietet die deskriptive Statistik Möglichkeiten an, Daten und deren Datenverteilung durch Messzahlen über deren charakteristische Eigenschaften zu beschreiben. Durch statistische Messzahlen, von denen einige besonders wichtige im folgenden genannt werden, lassen sich die Daten dann in komprimierter Form quantitativ darstellen.[381]

4.2.2.1.1.1 Univariate Methoden Sei S eine Datenmenge aus n Datentupeln, mit den Attributwerten $a_1, \ldots, a_n$ eines interessierenden Attributes A. Für jeden aufgetretenen Attributwert a_i in der Menge S bezeichne $h(a_i)$ die absolute Häufigkeit und $f(a_i) = \frac{h(a_i)}{n}$ die relative Häufigkeit des Attributwertes a_i. Eine Approxi-

[380] Auf die Beschreibung der zahlreichen graphischen Darstellungsformen wird daher im folgenden verzichtet und lediglich auf die Literatur verwiesen. Siehe hierzu beispielsweise die Ausführungen in Fahrmeier et al. (1997), S. 32-46 oder Bohley (1996), S. 73ff.

[381] Vgl. im folgenden hierzu z. B. Fahrmeier et al. (1997), S. 29ff; Bohley (1996), S. 117ff.

mation der Häufigkeitsverteilung lässt sich durch die Dichtefunktion

$$f(x) \text{ mit } f(x) \geq 0 \text{ und } \int f(x)dx = 1$$

beschreiben.[382] Die Fläche, die von der Kurve über einem bestimmten Intervall $[a,b]$ begrenzt wird, ist dann als prozentualer Anteil der Attributsausprägungen, die in dieses Intervall fallen, zu interpretieren. Zur Beschreibung des Zentrums einer Verteilung wird als Lagemasse für numerische Wertebereiche das arithmetisches Mittel $\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$ definiert. Zur Beschreibung der Datenverteilung um das arithmetische Mittel werden die allgemein gebräuchliche Streuungsmasse definiert:[383]

- Varianz: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \overline{a})^2$,
- Standardabweichung: $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - \overline{a})^2}$

Weichen die Daten extrem vom Mittel ab, tragen die Abweichungen durch das Quadrat stark zur Summe bei, wodurch die Varianz sehr gross wird. Die Varianz reagiert daher sensibel auf Ausreisser in den Daten und stellt neben dem arithmetischen Mittel ein geeignetes Mass zur Beschreibung von Dateneigenschaften dar.

4.2.2.1.1.2 Multivariate Methoden Typischerweise ist bei praktischen Fragestellungen nicht nur die Verteilung eines einzigen Attributs interessant sondern insbesondere die Beziehungen zwischen Attributen. Damit lassen sich dann im Grunde Abweichungen einzelner Werte von der allgemeinen Beziehungsstruktur erkennen. Im Beispiel, das in Abbildung 4.7 dargestellt ist, wird die frühzeitige „Kapitalauszahlung am 01.01.1998“ ohne Kapitalbewilligung deutlich. Ein weiteres Beispiel ist die verspätete Kapitalauszahlung am 01.01.2002, bei der die Kapitalbewilligung bereits verfallen ist. Solche zeitlichen Abhängigkeiten zwischen Attributen findet sich häufig in betrieblichen Datenbeständen, die damit die zeitliche Folge eines Geschäftsprozesses dokumentieren.

[382] Vgl. Fahrmeier et al. (1997), S. 85; Bohley (1996), S. 347ff.

[383] Vgl. Fahrmeier et al. (1997), S. 67; Bohley (1996), S. 154ff. Die empirische Varianz und Standardabweichung wird aus konkreten Daten berechnet und ist daher von der Varianz und Standardabweichung von Zufallsvariablen zu unterscheiden.

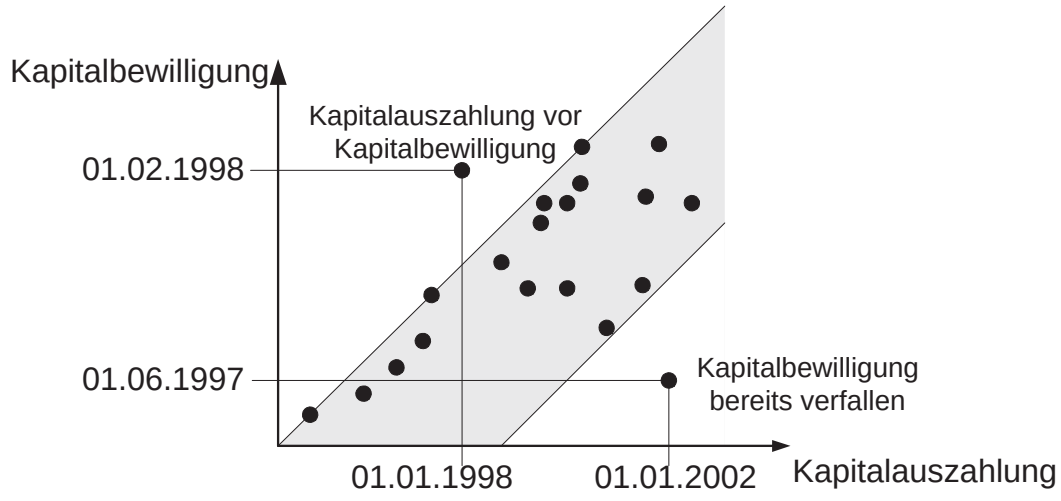


Abbildung 4.7: Beispiel für bivariate Analysen (In Anlehnung an Seidl (2001), S. 20)

Methoden, die Abhängigkeiten zwischen mehreren Attributen betrachten, werden als multivariate Methoden bezeichnet.[384] Im folgenden sollen exemplarisch lediglich Beziehungen zwischen zwei Attributen, sogenannte bivariate Methoden, untersucht werden. Zur Beurteilung des Zusammenhangs zweier Attribute ist die Betrachtung der Verteilung eines der beiden Attribute für die verschiedenen Ausprägungen des anderen Attributs besonders interessant. Diese stellen dann sogenannte bedingte Verteilungen dar, die zwar eine Beurteilung des Zusammenhangs erlauben, jedoch ohne dessen Stärke zu quantifizieren. Der Kontingenzkoeffizient drückt dagegen die Stärke des Zusammenhangs aus, indem er die Differenz zwischen der Verteilung bei unabhängigen Attributen in Beziehung zu der tatsächlichen Verteilung setzt. Dieses Abhängigkeitsmass ist definiert als χ^2-Koeffizient durch:[385]

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \frac{h_{i\cdot} \cdot h_{\cdot j}}{n})}{\frac{h_{i\cdot} \cdot h_{\cdot j}}{n}} \text{ und } \chi^2 \in [0, \infty)$$

Aus Normierungsgründen wird im allgemeinen der Kontingenzkoeffizient ver-

[384] Vgl. Sachs (1999), S. 576; Fahrmeier et al. (1997), S. 109ff.

[385] Wobei h_{ij} die absolute Häufigkeit und $h_{i\cdot}$,$h_{\cdot j}$ die Randhäufigkeiten der Kontingenztabelle bezeichnen; k, m bezeichnen die Anzahl der Wertausprägungen der jeweiligen Attribute; vgl. Fahrmeier et al. (1997), S. 123.

wendet:[386]

$$K = \sqrt{\frac{\chi^2}{n+\chi^2}} \text{ und } K \in \left[0,\ \sqrt{\frac{M-1}{M}}\right] \text{ mit } M = min\{k,m\}$$

Bei stetigen Wertebereichen verwendet man überlichweise zur Erfassung der Stärke des Zusammenhangs den empirischen Korrelationskoeffizienten, der den linearen Zusammenhang zwischen zwei Attributen misst. Für Attribute A und B mit stetigen Wertebereiche und Werten a_i und b_i sowie den arithmetischen Mitteln $\overline{a}$ und $\overline{b}$ ist der empirische Korrelationskoeffizient r_{AB} wie folgt definiert:[387]

$$r_{AB} = \frac{\sum_{i=1}^{n}(a_i-\overline{a})(b_i-\overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i-\overline{a})^2 \cdot \sum_{i=1}^{n}(b_i-\overline{b})^2}} \text{ und } r_{AB} \in [-1,1]$$

Allerdings ist bei der Interpretation des Korrelationskoeffizienten zu beachten, dass dieser keinen Aufschluss über die Richtung des Zusammenhangs erlaubt. Weiter ist zu beachten, dass der Korrelationskoeffizient lediglich für lineare Zusammenhänge zwischen zwei Attributen geeignet ist. Bei nicht linearen Abhängigkeiten heben sich die Summanden weitgehend auf und ein Wert nahe bei Null resultiert.

Zur Untersuchung gerichteter Zusammenhänge bietet sich die Regression als statistische Methode an.[388] Im einfachsten Fall, einer linearen Abhängigkeit zwischen zwei Attributen mit stetigen Wertebereichen, versucht man deren Beziehung durch eine Gerade der Form $f(A) = \alpha + \beta A$ zu beschreiben. Für die Datenpaare (a_i, b_i) mit $i = 1,\ldots,n$ gilt dann die lineare empirische Beziehung $b_i = \alpha + \beta a_i + \varepsilon_i$ wobei ε_i gerade den durch die Geradenanpassung bedingten Fehler wiedergibt. Dieses Modell wird als lineare Einfachregression bezeichnet,[389] das allerdings nicht immer zur Beschreibung eines Zusammenhangs zwischen zwei Attributen adäquat ist. Zur Analyse von Wachstumsverläufen oder Sättigungskurven werden daher

[386] n bezeichnet die Anzahl der zugrundeliegenden Datentupel und k,m die Anzahl der Wertausprägungen der jeweiligen Attribute; vgl. Fahrmeier et al. (1997), S. 125.

[387] Vgl. Fahrmeier et al. (1997), S. 136.

[388] Vgl. zu Regressionsmodelle z. B. Fahrmeir, Hamerle und Tutz (1996), S. 152ff. oder Bohley (1996), S. 205ff.

[389] Vgl. Fahrmeier et al. (1997), S. 158.

auch nichtlineare Regressionen gebildet, wobei der Zusammenhang zwischen den Attributen dann durch eine nicht lineare Funktion beschrieben wird.

Bislang wurden übliche Beschreibungen von Datenverteilungen anhand des arithmetischen Mittels, der Varianz und Standardabweichung sowie die Beschreibung von Beziehungen zwischen Attributen dargestellt. Im folgenden sollen die Möglichkeiten des Data Mining zur Datenanalyse sowie zur Beschreibung von Beziehungen und Strukturen in den Daten diskutiert werden.

4.2.2.1.2 Data Mining zur Musterbeschreibung Zur Ermittlung von Strukturen und Zusammenhängen in den Daten können, neben den Möglichkeiten der deskriptiven Statistik, Techniken des Data Mining genutzt werden.[390] Zwar sind die im Data Mininig eingesetzten Methoden aus den Bereichen der Statistik und der künstlichen Intelligenz seit längerem bekannt, jedoch wird der Begriff in den letzten Jahren in den Bereichen der Betriebswirtschaft, der Datenbanken und des Informationsmanagements gerne als Schlagwort verwendet. Im diesem Zusammenhang werden auch vielfach Begriffe wie neuronale Netze, Entscheidungsbäume, Regelinduktion, k-nächste-Nachbarn, Diskriminanzanalyse, Assoziationsregeln und Clusteranalysen sowie die Begriffe lineare Methoden, Fuzzy Logic, genetische Algorithmen und Knowledge Discovery genannt. Meist ist damit der Gesamtablauf der Erkenntnisgewinnung aus Datenbanken (Knwoledge Discovery in Databases) gemeint, bei dem Data Mining genaugenommen lediglich ein Teilschritt darstellt.[391] Allgemein stellt Data Mining ein Oberbegriff für Methoden und Techniken dar, die bislang unbekannte Zusammenhänge in den Datenbeständen eines Unternehmens aufdecken.[392] Der methodische Kernaspekt liegt bei Verfahren, die selbständig Annahmen generieren, diese prüfen und relevante Ergebnisse zur Datenbeschreibung extrahieren. Daher wird Data Mining häufig auch als Datenmustererkennung bezeichnet.[393] Data Mining umfasst die Anwendung spezieller Algorithmen zum Aufspüren von Datenmustern und wird beispielsweise im Rahmen von OLAP-Systemen

[390] Vgl. Witten und Frank (2000), S. 8.

[391] Vgl. Holthuis (1999), S. 58.

[392] Vgl. im folgenden z. B. Ester und Sander (2000), S. 4; Holthuis (1999), S. 58; Witten und Frank (2000), S. 61-81.

[393] Vgl. Mertens, Bissantz, Hagedorn und Schultz (1994), S. 739.

- zur Bonitätsanalysen bei der Kreditvergabe,
- bei Portfolioanalysen, Kreditanalysen und Risikoanalysen,
- beim Aufdecken von Kreditkartenmissbräuchen,
- bei Devisenkursprognosen sowie
- beim Erkennen von Kundenabwanderungstendenzen

eingesetzt.[394]

Während deduktives Lernen neue Erkenntnisse aus logischen Schlussfolgerungen gewinnt, ist Data Mining dem induktiven Lernen zugeordnet.[395] Induktives Lernen leitet aus grossen Datenbeständen neue Zusammenhänge und Erkenntnisse ab. Data Mining hat zunächst die Beschreibung einer Datenmenge durch Muster, die in ihr gefunden werden können, zum Ziel. Diese werden dann zur Vorhersage unbekannter oder zukünftiger Werte interessierender Attribute einer Datenmenge genutzt.[396] Oftmals liegen bereits unbestätigte Hypothesen oder empirische Beobachtungen vor, die durch den Einsatz von Data Mining gezielt verifiziert, erweitert oder auch widerlegt werden sollen. Aufgrund der Zielsetzung von Data Mining sind die Verfahren und Techniken insbesondere zur Ermittlung von komplexen Zusammenhängen in grossen Datenbeständen geeignet,[397] da

- sie gerade für grosse Datenbestände entwickelt wurden,
- die Isolierung von unregelmässigen und nicht offensichtlichen Datenmustern bezwecken sowie
- meist keine Annahmen über die Datenverteilung benötigen.

Die Anwendung des Data Mining auf Aspekte der Datenqualität wird bereits in einigen Veröffentlichungen unter dem Begriff des *(Data) Quality Mining* beschrieben.[398] Im Unterschied zum Data Mining werden beim Quality Mining die ex-

[394] Vgl. Mertens und Wieczorrek (1999), S. 213; Holthuis (1999), S. 60.
[395] Vgl. Mertens und Wieczorrek (1999), S. 212.
[396] Vgl. Schinzer et al. (1999), S. 104
[397] Vgl. Dasu und Johnson (1999), S. 89.
[398] Vgl. Soler und Yankelevich (2001), S. 163; Grimmer und Hinrichs (2001), S. 223ff.

plizierten Zusammenhänge genutzt, um Aussagen über die Datenqualität zu gewinnen. Im Rahmen der Arbeit sind dabei vor allem Techniken interessant, die leicht verständliche, strukturierte Datenbeschreibungen erzeugen.[399] Diese bilden dann eine Grundlage zur Diskussion mit den Fachexperten und zur Erstellung von Integritätsbedingungen.

Im folgenden werden daher einige aus dem Kontext des Data Mining bekannte Verfahren und Techniken erläutert, die bereits zur Analyse der Datenqualität eingesetzt werden.[400] Im Vordergrund steht dabei die Gewinnung von leicht zu repräsentierenden Beschreibungen über die Datenzusammenhänge. Dabei sind die Verfahren der Segmentierung, der Klassifizierung und der Assoziierung zu nennen.[401] Jedes dieser Verfahren bedient sich wiederum verschiedener Techniken, um zu einem entsprechenden Ergebnis zu kommen. Die Zusammenhänge zwischen den Techniken und den Verfahren des Data Mining sind in Abbildung 4.8 dargestellt und sollen im folgenden in bezug auf die Ermittlung der Datenqualität erörtert werden.

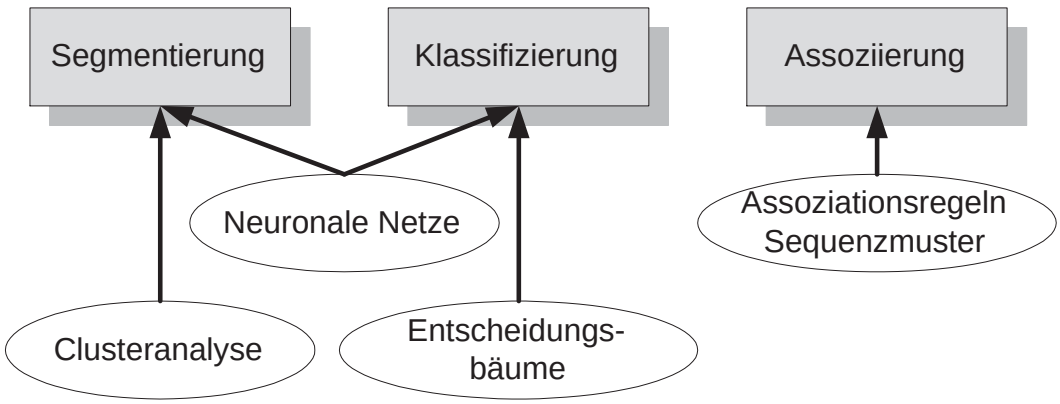


Abbildung 4.8: Verfahren und Techniken des Data Mining (In Anlehnung an Schinzer et al. (1999), S. 107)

[399] Techniken, wie beispielsweise neuronale Netze, die keine expliziten Beschreibungen erzeugen, sollen hier nicht berücksichtigt werden; vgl. zu neuronale Netze beispielsweise Haykin (1999).

[400] Vgl. z. B. de Fries, Seidl und Windheuser (1999), S. 515f.; Dasu und Johnson (1999), S. 90ff.; Soler und Yankelevich (2001), S. 163ff.; Grimmer und Hinrichs (2001), S. 225ff.

[401] Vgl. hierzu z. B. Ester und Sander (2000), S. 45ff.; Witten und Frank (2000), S. 61-81; Schinzer et al. (1999), S. 104-123.

4.2.2.1.2.1 Segmentierung Die Segmentierung, teilweise auch als Clusterung bezeichnet, teilt einen Datenbestand in Gruppen relativ homogener Datensätze ein.[402] Dabei werden Einheiten zu Gruppen zusammengefasst, die in einer gewissen Anzahl interessierender Eigenschaften weitgehend übereinstimmen. Ergebnis der Segmentierung ist eine aus wenigen Gruppen bestehende Zusammenfassung des Inhalts einer Datenbank. Die Gruppen können dann über charakteristische Gruppeneigenschaften, sogenannte Profile, beschrieben werden. Neben neuronalen Netzen ist die Clusteranalyse die wichtigste Technik zur Segmentierung. Dabei soll eine möglichst grosse Homogenität innerhalb und eine möglichst grosse Heterogenität zwischen den Gruppen erreicht werden. Die Ähnlichkeit von Daten wird durch Vergleich von zugehörigen Attributwerten ermittelt. Als Ähnlichkeitsmass eignen sich im wesentlichen Korrelationskoeffizienten oder Distanzmasse. Üblicherweise werden Distanzmasse eingesetzt, die eine Distanz zwischen Datenobjekten ermitteln. Im Gegensatz zur Klassifikation lassen sich allerdings keine eindeutigen „Wenn-Dann-Regeln“ formulieren. Vielmehr werden häufig auftretende Muster erkannt und zu Segmenten zusammengefasst.

Die Segmentierung findet in bezug auf *Datenqualität* ihre Anwendung insbesondere zur Ermittlung von Datenausreissern und Unregelmässigkeiten.[403] Datenausreisser bei der Segmentierung sind Daten, die sich keinem „sinnvollen“ Segment zuordnen lassen. Sinnvolle Segmente sind solche, die eine grosse Anzahl von Datensätzen aufweisen und häufig auftretende Muster im Datenbestand repräsentieren. Abweichungen von diesen Mustern sind mögliche Fehlerkandidaten. Segmente mit einer kleinen Anzahl von Datensätzen beinhalten dagegen solche Muster, die nur selten auftreten und daher mögliche Fehlerkandidaten darstellen.

4.2.2.1.2.2 Klassifizierung Aufgabe der Klassifizierung ist es, Elemente mit teilweise unbekannten Eigenschaften Klassen zuzuteilen.[404] Aufgrund der Klassenzugehörigkeit und den damit verbundenen Eigenschaften können dann Vor-

[402] Vgl. hierzu z. B. Witten und Frank (2000), S. 81f.; Ester und Sander (2000), S. 45ff.; Schinzer et al. (1999), S. 104f.

[403] In Dasu und Johnson (1999), S. 90ff. findet sich ein Beispiel zur Erkennung von Unregelmässigkeiten im Datenbestand auf Basis eines zur Segmentierung von Datenbeständen entwickelten Algorithmus.

[404] Vgl. hierzu z. B. Witten und Frank (2000), S. 62-67; Ester und Sander (2000), S. 107ff.; Schinzer et al. (1999), S. 106f. u. 108f.

hersagen über unbekannte Werte gemacht werden. Grundlage der Klassifizierung ist ein induktives und überwachtes Lernen der Klassifizierungsentscheidung in der Form, dass die wesentlichen Eigenschaften der Klassen durch Techniken selbständig anhand von sogenannten Trainingsdaten ermittelt werden. Durch die Trainingsdaten, die bereits mit einer Klassenzugehörigkeit versehen sind, wird ein Modell zur Beschreibung der Klassenzuordnung entwickelt und durch Strukturen und Regeln abgebildet. Dabei wird eine Funktion erstellt, die zukünftige Daten aufgrund derer Attributwerte auf eine Klasse abbildet. Eng mit der Klassifikation zusammenhängend sind die statistischen, multivariaten Regressionsverfahren, die allerdings bei besonders komplexen nichtlinearen Zusammenhängen weitaus schlechtere Ergebnisse liefern.[405] Schwächen zeigt die Klassifikation bei Daten aus Zeitreihen und der Vorhersage kontinuierlicher, numerischer Attribute.[406]

Zur Abbildung des Klassifikationsmodells haben sich Entscheidungsbäume und Entscheidungsregeln bewährt.[407] Zusammenhänge zwischen Attributen und Klassen lassen sich durch die graphische Darstellung der Klassifikation anhand einer Baumstruktur leicht erläutern. Wird der Baum allerdings komplex, bietet sich die Umwandlung in natürlichsprachliche Entscheidungsregeln an, die aufgrund ihrer Struktur häufig Sachverhalte kürzer darstellen können.[408]

Ein Entscheidungsbaum besteht aus einem Wurzelknoten, weiteren Entscheidungsknoten und Blättern. Die Entscheidungsknoten bedingen die Auswertung eines bestimmten Attributs. Im einfachsten Fall erfolgt die Auswertung in Form eines Vergleiches mit einer Konstanten. Denkbar sind auch Auswertungen zweier Attribute oder die Anwendung einer Funktion auf Attributwerte. Jeder aus den Knoten abgehende Ast steht für die möglichen Werte dieses Attributs. Die Daten werden klassifiziert, indem sie entsprechend des getesteten Attributwertes an verschiedene Unterbäume weitergeleitet werden und sich dort das Vorgehen mit einem anderen Attribut wiederholt. Durch den Unterscheidungsprozess an jedem

[405] Vgl. Schinzer et al. (1999), S. 106.

[406] Vgl. Berry und Linoff (1997), S. 284.

[407] Vgl. im folgenden Witten und Frank (2000), S. 62ff. und Schinzer et al. (1999), S. 109-112.

[408] Entscheidungsbäume können in Regelmengen transformiert werden, wobei im allgemeinen die abgeleitete Regelbasis nicht minimal ist. Die Umkehrung, d. h. die Umwandlung einer allgemeinen Regelbasis in einen Entscheidungsbaum ist dagegen aufwendiger. Häufig werden Regeln, aufgrund ihrer hohen Flexibilität hinsichtlich Änderungen, gegenüber Entscheidungsbäumen bevorzugt.

Knoten werden die zu klassifizierenden Daten solange disjunkt aufgeteilt, bis alle Daten in den Blättern des Baumes verteilt sind. Die Blätter geben dann für ihre jeweiligen Datensätze die Zuordnung zu einer Klasse an. Der Weg vom Wurzelknoten bis zum Endknoten beschreibt die Klassifikation, nach der die Daten gruppiert werden.

Eine beliebte Alternative zu Entscheidungsbäumen sind Klassifikationsregeln der allgemeinen Form: *Bedingung* $\rightarrow$ *Folgerung*.[409] Ähnlich den Auswertungen in den Knoten von Entscheidungsbäumen erfolgt die Anwendung einer Regel durch Auswertung der in der Regelbedingung genannten Attributen. Die Folgerung beschreibt dann die entsprechend zugeordnete Klasse. Häufig werden Bedingungselemente mit einem logischen *UND* verknüpft, wobei auch allgemein logische Ausdrücke möglich sind.[410] Generell gilt, dass die Klassifikationsgüte mit der Anzahl der in die Analyse einfliessenden Attribute steigt. Allerdings sinkt mit wachsender Anzahl der Regeln die Übersichtlichkeit und damit das Verständnis für das Ergebnis. Eine Erweiterung der Klassifikationsregeln besteht darin, Ausnahmen zu deklarieren.[411] Anstatt die gesamte Regelmenge neu zu definieren, werden Modifikationen an dieser vorgenommen. Es werden für bereits existierende Regeln Ausnahmen formuliert, zu deren Begründung Fachexperten einzubeziehen sind.

Das folgende Beispiel soll die Leistungsfähigkeit der Klassifikation für Analysen der Datenqualität erläutern.[412] Sei der in Tabelle 4.4 vorliegende Datenbestand mit der in Tabelle 4.5 aufgeführten Regelmenge zu prüfen. Die Regelmenge ist zusammen mit dem Fachexperten aufgrund von Qualitätsreferenzdaten und bisheriger Erfahrungen zu erstellen und laufend anzupassen. Es zeigt sich, dass ein aktuell vorliegende Datensatz $< 54, A, \ldots, 1 >$ der Form $< Var_1, Var_2, \ldots, Var_n >$ weder vereinbar mit Regel 4 noch mit Regel 5 (oder irgendeiner anderen Regel) ist und daher einen potentiellen Fehlerkandidaten darstellt.

[409] Vgl. Witten und Frank (2000), S. 63-67.

[410] Vgl. Witten und Frank (2000), S. 72. In Abhängigkeit der Bedingung können Regeln in propositionale und relationale Regeln unterteilt werden. Propositionale Regeln berücksichtigen ein Attribut und vergleichen dessen Wert mit einer Konstanten. Regeln die Beziehungen zwischen Attributen ausdrücken, werden als relational bezeichnet.

[411] Vgl. Witten und Frank (2000), S. 69-71.

[412] Vgl. im folgenden auch Seidl (2001), S. 25-28

Var_1	Var_2	...	Var_n
50	A		0
51	A		0
77	B		0
98	C		0
54	A		1
74	B		0
45	A		1
94	C		0
54	B		1
⋮	⋮	⋮	⋮

Tabelle 4.4: Daten für ein Klassifikationsbeispiel

$$
\begin{array}{lll}
\text{rule 1} & (Var_1 = 7x) & \Rightarrow \quad (Var_2 = B) \\
\text{rule 2} & (Var_1 = 9x) & \Rightarrow \quad (Var_2 = C) \\
\text{rule 3} & (Var_1 = 4x) AND (Var_n = 1) & \Rightarrow \quad (Var_2 = A) \\
\text{rule 4} & (Var_1 = 5x) AND (Var_n = 0) & \Rightarrow \quad (Var_2 = A) \\
\text{rule 5} & (Var_1 = 5x) AND (Var_n = 1) & \Rightarrow \quad (Var_2 = B) \\
\vdots & \vdots & \vdots \quad \vdots
\end{array}
$$

Tabelle 4.5: Regelmenge für ein Klassifikationsbeispiel

Segmentierung und Klassifizierung finden zunächst zur Qualitätsanalyse der Datenqualität ihre Anwendung. Diese Analysen können zum Aufdecken von Fehlerkandidaten und zur Aufstellung neuer und Prüfung bisheriger Integritätsbedingungen genutzt werden. Sie bieten eine geeignete Diskussionsgrundlage für Gespräche zwischen Datenqualitätsverantwortlichen und Fachexperten. Weiter liefern sie anhand einer Trainingsmenge ein Modell zur Aufteilung des Datenbestands in relativ homogene Klassen bzw. Segmente. Diese können dann durch charakteristische Eigenschaften, den sogenannten Profilen, beschrieben werden. Neue Datenbestände werden dann erneut durch das Segmentierungs- bzw. Klas-

sifikationsmodell unterteilt und deren Profile bestimmt. Unter der Annahme, dass beide Datenbestände ähnlich sind, müssten die Profile weitgehend identisch sein. Weichen die Profile signifikant voneinander ab, ist davon auszugehen, dass die Datenbestände in ihren charakteristischen Eigenschaften unterschiedlich sind. Es ist daher zu prüfen, ob ein Datenqualitätsproblem, eine einmalige Ausnahme oder eine grundlegende Veränderung der charakteristischen Eigenschaften vorliegt.

4.2.2.1.2.3 Assoziierung Neben der Klassifikation und Segmentierung können mit Hilfe der Assoziierung Datenmuster von zusammenhängenden Attributwerten entdeckt[413] und so für die Datenqualitätsanalyse genutzt werden. Die Muster werden im allgemeinen in Form von Assoziationsregeln ausgedrückt. Assoziationsregeln unterscheiden sich prinzipiell von Klassifikationsregeln, indem sie beliebige Attribute und Attributkombinationen vorhersagen und nicht ausschliesslich eine Klassenzugehörigkeit bestimmen. Ausserdem sind Assoziationsregeln im Unterschied zu Klassifikationsregeln nicht als zusammenhängende Regelbasis vorgesehen. Jede Assoziationsregel drückt ein bestimmtes Datenmuster im Datenbestand aus.

Assoziationsregeln beschreiben häufig auftretende und starke Zusammenhänge zwischen Attributwerten a, b, c durch Regeln, wie beispielsweise der Form a AND $b \Rightarrow c$. Um das Assoziationsproblem zu beschreiben, betrachtet man eine Datenmenge D von Transaktionen t. Jede Transaktion besteht aus einer Menge von Elementen, genannt Items. Die in Tabelle 4.6 exemplarisch dargestellte Relation könnte dann wie in Tabelle 4.7 als Menge von Transaktionen beschrieben werden. Die Auftretenshäufigkeit der einzelnen Items und ihre Zuordnung zu den Transaktionen wird in Tabelle 4.8 zusammengefasst.

Eine allgemeine Assoziationsregel $X \Rightarrow Y$ bedeutet, dass das Auftreten der Itemmenge X zum Auftreten der Itemmenge Y führt. Itemmenge X wird als Regelbedingung und Itemmenge Y als Regelfolge bezeichnet. Eine Transaktion erfüllt eine Assoziationsregel $X \Rightarrow Y$, wenn alle Items, die in der Regel vorkommen auch in der Transaktion auftauchen. Es gilt also $(X \cup Y) \subseteq t$. Für einzelne Assoziations-

[413] Vgl. hierzu z. B. Witten und Frank (2000), S. 67-69; Ester und Sander (2000), S. 159ff.; Schinzer et al. (1999), S. 106 u. S. 117-123.

Kunden_ID	Produkt	Datum	...
1	A	15.10.01	
1	B	26.10.01	
1	C	28.10.01	
2	B	16.10.01	
2	C	29.10.01	
3	A	04.10.01	
4	A	08.10.01	
4	C	14.10.01	
4	B	23.10.01	

Tabelle 4.6: Beispielrelation für Assoziationsregeln

Kunden_ID	Produkt
1	A,B,C
2	B,C
3	A
4	A,C,B

Tabelle 4.7: Datenmenge von Transaktionen

regeln sind insbesondere Support und Konfidenz interessant.[414] Der Support einer Assoziationsregel ist derjenige Anteil aller Transaktionen, der die Regel erfüllt, sprich[415]

$$support(X \Rightarrow Y) = \frac{|\{t \in D | (X \cup Y) \subseteq t\}|}{|D|}$$

$|D|$ bezeichnet dabei die Anzahl aller Transaktionen t in D. Konfidenz einer Assoziationsregel bezeichnet den Anteil der Transaktionen, welche X enthalten und

[414] Beim Data Mining werden durch Festlegen eines Mindestsupports Regeln ausgeschlossen, die nur für einen kleinen Bruchteil des Datenbestandes gelten. Analog können nicht sichere Regeln durch Angabe einer Mindestkonfidenz ausgeblendet werden.

[415] Vgl. Schinzer et al. (1999), S. 1119.

Item	Auftretenshäufigkeit	Transaktionen mit diesem Produkt
A	75%	1,3,4
B	75%	1,2,4
C	50%	2,4

Tabelle 4.8: Auftretenshäufigkeit der Items

für welche die Assoziationsregel gilt. Zur Berechnung wird die Anzahl der regelerfüllenden Transaktionen durch die Anzahl aller Transaktionen, die X enthalten, geteilt:[416]

$$konfidenz(X \Rightarrow Y) = \frac{|\{t \in D|(X \cup Y) \subseteq t\}|}{|t \in D|X \subseteq t\}}$$

In Tabelle 4.9 sind für sämtliche Kombinationen zweier Attributwerte deren Support und Konfidenz angegeben.

Regel	Support	Konfidenz
$A \Rightarrow B$	50%	66%
$B \Rightarrow A$	50%	66%
$A \Rightarrow C$	50%	66%
$C \Rightarrow A$	50%	66%
$B \Rightarrow C$	50%	100%
$C \Rightarrow B$	50%	100%

Tabelle 4.9: Support und Konfidenz

Ein Beispiel zur Verwendung von Assoziationsregeln für die Analyse der Datenqualität ist QUASAR.[417] Mit Hilfe allgemein üblicher Algorithmen werden hier Assoziationsregeln in Datenbeständen entdeckt, welche dann zur weiteren Untersuchung des Datenbestandes dienen. Die Konfidenz einer Assoziationsregel ist in zweifacher Hinsicht interessant. Einerseits legt die Mindestkonfidenz den Wert

[416] Vgl. Schinzer et al. (1999), S. 120.

[417] Vgl. Quality Assessment using Association Rules in Soler und Yankelevich (2001), S. 165-170.

fest, ab welcher eine Regel als gültig anzusehen ist. Regeln mit hoher Konfidenz stellen mögliche Integritätsbedingungen für den Datenbestand dar. Potentielle Fehlerkandidaten sind Daten, die zu diesen Regeln nicht konsistent sind. Andererseits sind Regeln mit einer geringen Konfidenz durch wenige Transaktionen gestützt. Diesen Transaktionen könnten Datenausreisser zugrunde liegen und sollten daher näher untersucht werden. Bezüglich des Supports sind solche Regeln interessant, die einen geringen Wert aufweisen, denn sie treten lediglich in einer geringen Anzahl von Transaktionen auf. Eine analoge Betrachtung bietet sich auch für Sequenzmuster an, welche die zeitliche Ordnung der Items in Transaktionen berücksichtigen.[418]

4.2.2.1.2.4 Einsatzgebiet des Data Mining Wie die obigen Ausführungen gezeigt haben, kann Data Mining zunächst zur Qualitätsanalyse des Datenbestandes und der Suche nach potentiellen Fehlerkandidaten eingesetzt werden. Weiter können durch den Einsatz von Techniken des Data Mining vermutete, aber nicht genau spezifizierte Regeln oder Muster aufgedeckt und beschrieben werden. Diese können dann zur Ableitung und Beschreibung von Integritätsbedingungen für den Datenbestand genutzt werden. Neben der Qualitätsanalyse erhöht der Einsatz strukturierter Datenanalysen die Vollständigkeit und Korrektheit von Plausibilitätsprüfungen.[419] Schwächen in den Integritätsbedingungen können aufgedeckt und durch konkrete Regeln und Muster unterstützt und ergänzt werden. Zur Ableitung fachbezogener Regeln sind neben dem Datenqualitätsanalysten Fachexperten einzubeziehen.

Die Techniken des Data Mining sind im allgemein effizient bei der Analyse grosser Datenbestände und erkennen bereits relativ geringe Variationen in den Datenverteilungen. Daher sind sie nicht nur auf wenige Attribute oder auf aggregierte Daten anwendbar, sondern eignen sich auch auf detaillierte, datensatzbezogene Analysen. Ein Anwendungsgebiet liegt daher in der Analyse der Data-Warehouse-Datenbasis aber auch in der Analyse operativer Datenbestände. Entscheidungsregeln, Entscheidungsbäume und Assoziationsregeln sind im Gegensatz zur Seg-

[418] Vgl. zu Sequenzmuster u. a. Schinzer et al. (1999), S. 121f.
[419] Vgl. de Fries et al. (1999), S. 516.

mentierung wegen ihrer transparenten Darstellungsform besonders zur Kommunikation mit den Fachexperten geeignet. Allerdings sinkt die Transparenz mit der Komplexität der Entscheidungsbäume bzw. der Regeln.

4.2.2.1.3 Berücksichtigung von Datenschwankungen Bislang wurden anhand der deskriptiven Statistik und des Data Mining charakteristische Eigenschaften der Datenbestände, Beziehungen zwischen Attributen sowie Muster in den Daten erläutert. Im folgenden wird die Möglichkeit zur Berücksichtigung der nicht qualitätsrelevanten Schwankungen untersucht. Hierzu eignen sich insbesondere stochastische Modelle, die zufallsabhängige Vorgänge beschreiben.[420] Stochastische Modelle fassen wichtige Aspekte einer ungewissen Wirklichkeit in transparente, mathematische Strukturen. Gedanklich wird dabei von einem Zufallsexperiment ausgegangen. Die den Zufallsvariablen zugrundeliegende Wahrscheinlichkeitsverteilung wird durch Parameter charakterisiert, die allerdings im allgemeinen unbekannt sind. Daher schätzt man sie meist aus empirischen Beobachtungen. Aufgrund der Probleme beim Schätzen der Parameter beschränkt man sich häufig auf einfache, übersichtliche Modelle. Diese stellen allerdings nur eine Approximation des tatsächlichen Modells dar. Teilweise lässt sich auch eine Annäherung an eine bekannte Verteilung, wie beispielsweise die Normalverteilung, durch Transformation erreichen.

4.2.2.1.3.1 Komplexitätsreduktion Eine allgemeine Beschreibung einer Datenverteilung kann in Form einer multivariaten Verteilungsfunktion angegeben werden.[421] Allerdings ist das Erfassen der Zusammenhangsstruktur von mehr als zwei Zufallsvariablen nicht einfach, so dass man häufig auf einfachere Modelle mit reduzierter Komplexität zurückgreift.[422] Zunächst kann die Zahl der zu untersuchenden Attribute auf wenige reduziert werden. Es werden dann lediglich

[420] Vgl. hierzu Sachs (1999), S. 194 u. S. 419f. sowie Weiss (1987). Stochastische Modelle bilden eine zentrale Grundlage des statistischen Testens, welcher auch die Grundlage der statistischen Prozesskontrolle bildet. Mit Hilfe statistischer Test können so Kontrollgrenzen um den Sollwert eines Prozesses ermittelt werden; vgl. hierzu z. B. Sachs (1999), S. 343f. und Fahrmeier et al. (1997), S. 387ff.

[421] Ausführliche Darstellungen multivariater Zufallsvariablen finden sich in der Literatur zur multivariaten Statistik, beispielsweise in Fahrmeir et al. (1996).

[422] Vgl. Milek et al. (2001), S. 193.

univariate und bivariate Verteilungsfunktionen betrachtet. Allerdings führt diese Reduktion zu Informationsverlusten über Beziehungen zwischen Attributen, die dann zu unvollständigen Aussagen über die Datenverteilung führen.[423] So lassen sich bestimmte Zusammenhänge bei univariater Betrachtung nicht erkennen, wenngleich sie unter multivariater Betrachtung relevant sind.

Eine weitere Möglichkeit zur Komplexitätsreduktion bietet sich durch Verwendung von aggregierten Daten hinsichtlich der Zeit und / oder des Zustandraumes an.[424] Während Aggregationen über den Zustandsraum Ausprägungen von Attributen zusammenfassen, berücksichtigen Aggregationen hinsichtlich der Zeit unterschiedliche Ausprägungen im Zeitablauf eines Attributes. Durch die Aggregation wird der Informationsgehalt erhöht und somit das Modell vereinfacht. Allerdings gleichen sich kleinere Schwankungen in den Daten meist aus, die dann im Aggregat nicht mehr erkannt werden. Zur Aggregation bieten sich dabei beispielsweise die Summe, die Anzahl, der Mittelwert, die Varianz, das Minimum oder Maximum über die Datenwerte des Aggregationsraumes an. So könnte an Stelle der multivariaten Analyse auf Detaildaten beispielsweise

- die durchschnittliche Anzahl der Konten pro Kundensegment, Produktkategorie oder Zeiteinheit sowie
- die Summe der Transaktionen pro Kundensegment, Produkt oder Zeiteinheit

betrachtet werden.[425]

4.2.2.1.3.2 Plausibilitätsintervalle Obgleich eine Komplexitätsreduktion durch Aggregation oder Reduktion der betrachteten Attribute vorgenommen wird, ist die Ermittlung einer guten Approximation der Datenverteilungen anhand eines stochastischen Modells zur Durchführung statistischer Testvorgänge notwendig.[426] Allerdings hat es sich gezeigt, dass sich auch einfache Aussagen

[423] Vgl. Wang (1999), S. 64f.
[424] Vgl. Milek et al. (2001), S. 194.
[425] Vgl. Milek et al. (2001), S. 194 oder auch Dasu et al. (2000), S. 191ff.
[426] Vgl. Sachs (1999), S. 343f.

im Sinne von Plausibilitätsintervallen ermitteln lassen. Diese Plausibilitätsintervalle können aufgrund von Erfahrungen und historischen Auswertungen mit dem Fachanwender festgelegt und sukzessive angepasst werden. Hierzu werden zunächst charakteristische Eigenschaften der Qualitätsreferenzdaten ausgewählt, mit dem Fachexperten diskutiert und realistische, auf Erfahrungen beruhende Plausibilitätsintervalle angegeben. Diese werden dann dynamisch überprüft und deren Bandbreite erweitert oder verringert. Einige exemplarische Angaben sollen dies im folgenden verdeutlichen.

Für Ausprägungen eines Attributs A können Plausibilitätsintervalle über deren obere und untere Grenzen ihrer relativen Häufigkeiten $f(a_i)$ angegeben werden. Für das Attribut Geschlecht könnte dies folgendermassen vorgenommen sein:

$$\begin{aligned} 0,55 &\leq f(\mathit{männlich}) \leq 0,65 \\ 0,35 &\leq f(\mathit{weiblich}) \leq 0,45 \end{aligned}$$

Für stetige Wertebereiche bietet sich eine Unterteilung in Gruppen an, so dass analog zu diskreten Wertebereichen Angaben über deren Gruppenverteilung möglich sind. So könnte der Wertebereich des Attributs Umsatz in vier Gruppen $\{a < 1000; 1000 \leq a \leq 10000; 10000 \leq a \leq 1000000; a > 1000000\}$ unterteilt und Angaben über deren Verteilung gemacht werden:

$$\begin{aligned} 0,03 &\leq f(a < 1000) \leq 0,05 \\ 0,1 &\leq f(1000 \leq a \leq 10000) \leq 0,2 \\ 0,7 &\leq f(10000 \leq a \leq 1000000) \leq 0,8 \\ 0,27 &\leq f(10000 \leq a \leq 1000000) \leq 0,6 \end{aligned}$$

Weiter kann der Wertebereich auf die als wahrscheinlich angenommenen Wertausprägungen durch Angabe von Minimal- und Maximalwerten begrenzt werden. Mit Hilfe von Plausibilitätsintervallen für das arithmetische Mittel, die Varianz bzw. deren Standardabweichung und den Korrelationskoeffizienten können ähnliche Angaben spezifiziert werden. Für das Attribut Geburtsdatum kann z. B. festgelegt sein, dass das arithmetische Mittel zwischen „1.1.1970“ und „1.1.1980“ mit einer Standardabweichung von maximal 10 Jahren sinnvoll ist. Es können aber auch Angaben in Abhängigkeit des arithmetischen Mittels gemacht werden. Ein

Beispiel wäre die Angabe, dass Werte des Attributs Geburtsdatum nicht mehr und nicht weniger als ± 10 Jahre vom arithmetischen Mittel abweichen dürfen.

4.2.2.1.3.3 Regressionsmodelle Wie oben dargestellt, lassen sich funktionale Zusammenhänge zwischen zwei Attributen A, B mit Hilfe einer approximativen Beziehung der Form $B = f(A) + \varepsilon$ abbilden. Dabei ist f eine deterministische Regressionsfunktion und ε ein Fehler, der durch das Attribut A alleine nicht erklärbar ist.[427] Im Sinne eines stochastischen Modells wird angenommen, dass der Fehler ε eine Zufallsvariable mit bestimmten Eigenschaften ist.[428] Dadurch entspricht $B = f(x) + \varepsilon$ ebenfalls einer Zufallsvariablen. Am bekanntesten ist das stochastische Modell der linearen Einfachregression, bei der eine lineare Regressionsfunktion $f(A) = \alpha + \beta A$ verwendet wird.[429] Unter Berücksichtigung des Fehlers ε geht dann die empirische Beziehung der Attribute über in das stochastische Grundmodell der linearen Regression:[430]

$$b_i = \alpha + \beta a_i + \varepsilon_i \text{ mit } E(\varepsilon_i) = 0 \text{ und } i = 1, \ldots, n.^{431}$$

Im allgemeinen liegen dem Modell weitere Annahmen zugrunde, wie beispielsweise die Unabhängigkeit und identische Verteilung der Zufallsvariablen ε_i mit $i = 1, \ldots, n$.[432] Wie oben lassen sich auch hier in Abstimmung mit dem Fachexperten für die Zufallsvariable ε_i Plausibilitätsintervalle angeben, in denen sich mit hoher Wahrscheinlichkeit die Werteausprägungen finden. In Abbildung 4.9 findet sich die visualisierte Interpretation, die gewissermassen ein Plausibilitätsband um die Gerade bildet, in welcher sich die Datenwerte mit hoher Wahrscheinlichkeit befinden.

[427] Vgl. Sachs (1999), S. 500ff.

[428] Vgl. Fahrmeier et al. (1997), S. 459.

[429] Vgl. Fahrmeier et al. (1997), S. 460ff.

[430] Vgl. Fahrmeier et al. (1997), S. 461; Anderson, Popp, Schaffranek, Steinmetz und Stenger (1997), S. 236f.; Bohley (1996), S. 653f.

[431] $E(\varepsilon_i)$ bezeichnet dabei den Erwartungswert der Verteilung von ε_i

[432] Vgl. Fahrmeier et al. (1997), S. 461f. Insbesondere besitzen dann alle ε_i gleich grosse Varianz. Sie wird oft dadurch verletzt, dass die Varianzen der ε_i mit steigenden Attributwerten ebenfalls zunehmen. Sind die Varianzen ungleich ist das Modell der linearen Regression nur in geeignet modifizierter Form anwendbar.

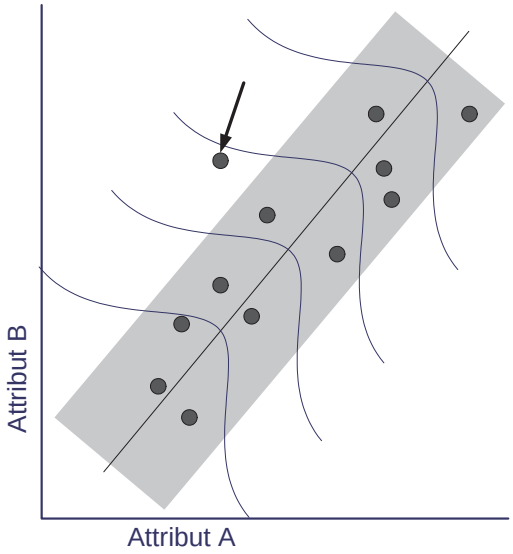


Abbildung 4.9: Plausibilitätsintervall für die lineare Regression (Eigene Darstellung)

4.2.2.1.3.4 Zeitreihenmodelle Besonders interessante Abhängigkeiten sind solche, die in Zusammenhang mit der Zeit entstehen, sogenannte Zeitreihen. Allgemein wird unter einer Zeitreihe eine zeitlich geordnete Folge a_t mit $t \in T$ von Beobachtungen einer Grösse verstanden.[433] Insbesondere bei ökonomischen Prozessen finden sich Zeitreihen, die langfristigen Veränderungen oder einer sich zyklisch wiederholenden Saisonsfigur unterliegen.[434] Es existieren daher zahlreiche ökonomische Zeitreihenmodelle, die diese Gegebenheiten berücksichtigen. Eine typische Approximation einer Zeitreihe wird durch verschiedene lineare und nichtlineare Regressionsmodelle unternommen, die ergänzend zum Grundmodell meist folgende Komponenten aufweisen:[435]

- Die Trendkomponente, die eine langfristige systematische Veränderung des mittleren Niveaus der Zeitreihe berücksichtigt.
- Eine Konjunkturkomponente, die eine mehrjährige, nicht notwendig regelmässige Schwankung darstellt.

[433] Vgl. Schlittgen und Streitberg (1999), S. 1.
[434] Vgl. Schlittgen und Streitberg (1999), S. 9.
[435] Vgl. Schlittgen und Streitberg (1999), S. 9; Stier (2001), S. 8f.

- Die Saisonskomponente, als eine jahreszeitlich bedingte Schwankungskomponente, die sich relativ unverändert jedes Jahr wiederholt.

Den allgemeinen Regressionsmodellen liegen allerdings zentrale Voraussetzungen zugrunde, die für Zeitreihen oft unrealistisch sind. Aus diesem Grund werden die oben beschriebenen linearen Regressionsmodelle für Zeitreihen weiter modifiziert. Es existieren für zahlreiche Anwendungsfälle geeignete Modelle.[436] Allgemein werden diese Modelle zur Vorhersage zukünftiger Werte genutzt, die allerdings auch im Sinne eines Filters für Marktschwankungen und insbesondere saisonale Zyklen anwendbar sind.

Anhand eines Beispiels soll diese Anwendung verdeutlicht werden.[437] Ausgehend von einer umfangreichen Datenbasis, die Daten über den Telefonverkehr in detaillierter Form enthält, wird ein Zeitreihenmodell erstellt. Hierzu wurden die Daten einer Analyse unterzogen und eine Abhängigkeit zwischen der Zeit und der Anzahl der Anrufe erfasst. Die Abhängigkeit wurde dann in einem stochastischen Zeitreihenmodell zur Schätzung der Anzahl der Anrufe abgebildet:[438]

$$V'(t) = V_B(1+D_W(wkd(t)))(1+D_M(day(t)))(1+D_Y(mon(t)))L(t)+\varepsilon \text{ für } t > t_0$$

Dabei ist V' eine Zufallsvariable, welche die erwartete Anzahl von Anrufen für den Zeitpunkt t abbildet. Die Ausgangsbasis V_B bezieht sich auf das tägliche Telefonaufkommen zum Startzeitpunkt t_0. Die Zeitkomponenten D_W, D_M, D_Y repräsentieren die wöchentlichen, monatlichen und jährlichen Zyklen, wobei die Funktionen $wkd(), day(), mon()$ die Wochentage, Monatstage und den Monat von t ermitteln. $L(t)$ ist eine Funktion zur Berücksichtigung der Trendkomponente. Im vorliegenden Beispiel wurde zur Berechnung der Trendkomponente die Funktion $L(t) = (1+\eta)^{\Delta(t_0,t)}$ verwendet, welche die Zunahme des Telefonaufkommens zwischen den Zeitpunkten t_0 und t ausgedrückt. Die monatliche Zunahme des

[436] Vgl. z. B. in Stier (2001); Schlittgen und Streitberg (1999); Bohley (1996), S. 255ff.; Harvey (1995).

[437] Vgl. hierzu das Beispiel aus einem Unternehmen in der Telekommunikationsbranche in Busatto (2000), S. 132-134.

[438] Die Anzahl der Anrufe ist nur ein Beispiel für Abhängigkeiten zwischen Attributen. So wird auch ein Zusammenhang zur Anrufdauer und der Rechnungssumme vermutet. Das in Busatto (2000), S. 132, genannte Modell wurde hier um die Fehlerkomponente ε erweitert.

Telefonaufkommens wird durch η berücksichtigt. Werden nun noch spezielle Ereignisse zu Zeitpunkten t angenommen, setzt sich die Gleichung zur Ermittlung einer prognostizierten Anzahl von Anrufen wie folgt zusammen:

$$V(t) = V_E(t) + w(t)V'(t)$$

Der Gewichtungsparameter $w(t)$ beschreibt dabei den Einfluss der Zyklen und des Trends zum Zeitpunkt t in bezug auf das spezielle Ereignis. Das vorliegende Beispiel zeigt den Einsatz von Zeitreihenmodellen im Sinne eines Filters im Bereich der Datenqualität. So konnten auf Grundlage der prognostizierten Werte nicht glaubhaft erscheinende Daten entdeckt und deren Ursachen in den operativen Systemen (beispielsweise inkorrekte und unvollständige Datenerfassung) ermittelt werden.

4.2.2.1.4 Berücksichtigung mengenmässiger Aspekte Ein wichtiger Aspekt, der hier erläutert werden soll, bezieht sich auf das Datenvolumen in Form der Anzahl von Datentupeln einer Relation oder eines Transferprozesses.[439] Bei einer mengenmässigen Betrachtung der Datenbestände lässt sich feststellen, dass sich das Datenvolumen in gewissen Grenzen gleichmässig entwickelt. Eine Analyse der Datenvolumen im Zeitablauf, die in Abbildung 4.10 dargestellt ist, zeigt dies für ausgewählte Relationen innerhalb der Transferprozesse.

Anhand der graphischen Darstellung lassen sich zunächst relativ kleine Schwankungen der Datenvolumen erkennen. Einige Ausprägungen weichen allerdings signifikant von dieser Verteilung ab und deuten so auf potentielle Fehlerkandidaten hin. Häufig führen doppelt oder nicht ausgeführte Transferprozesse zu extrem grossen Schwankungen in den Datenvolumen. Da in der Literatur zahlreiche Zeitreihenanalysen existieren,[440] ist eine Analyse der Entwicklung des Datenvolumens über die Zeit relativ einfach möglich. Insbesondere lassen sich mit Hilfe von Zeitreihenmodellen saisonale Zyklen oder Trendverläufe erkennen und abbilden. Anhand dieser Analysen lassen sich dann ebenfalls Bedingungen und Plausibilitätsintervalle für die Entwicklung der Datenvolumina in Transferprozessen

[439] Vgl. hierzu auch die Aussagen in Abschnitt 3.3.

[440] Vgl. z. B. in Stier (2001); Schlittgen und Streitberg (1999); Bohley (1996), S. 255ff.; Harvey (1995).

und Relationen angeben. Zur Analyse der Datenvolumen und zur Ableitung von Plausibilitätsintervallen sind wiederum Fachexperten einzubeziehen.

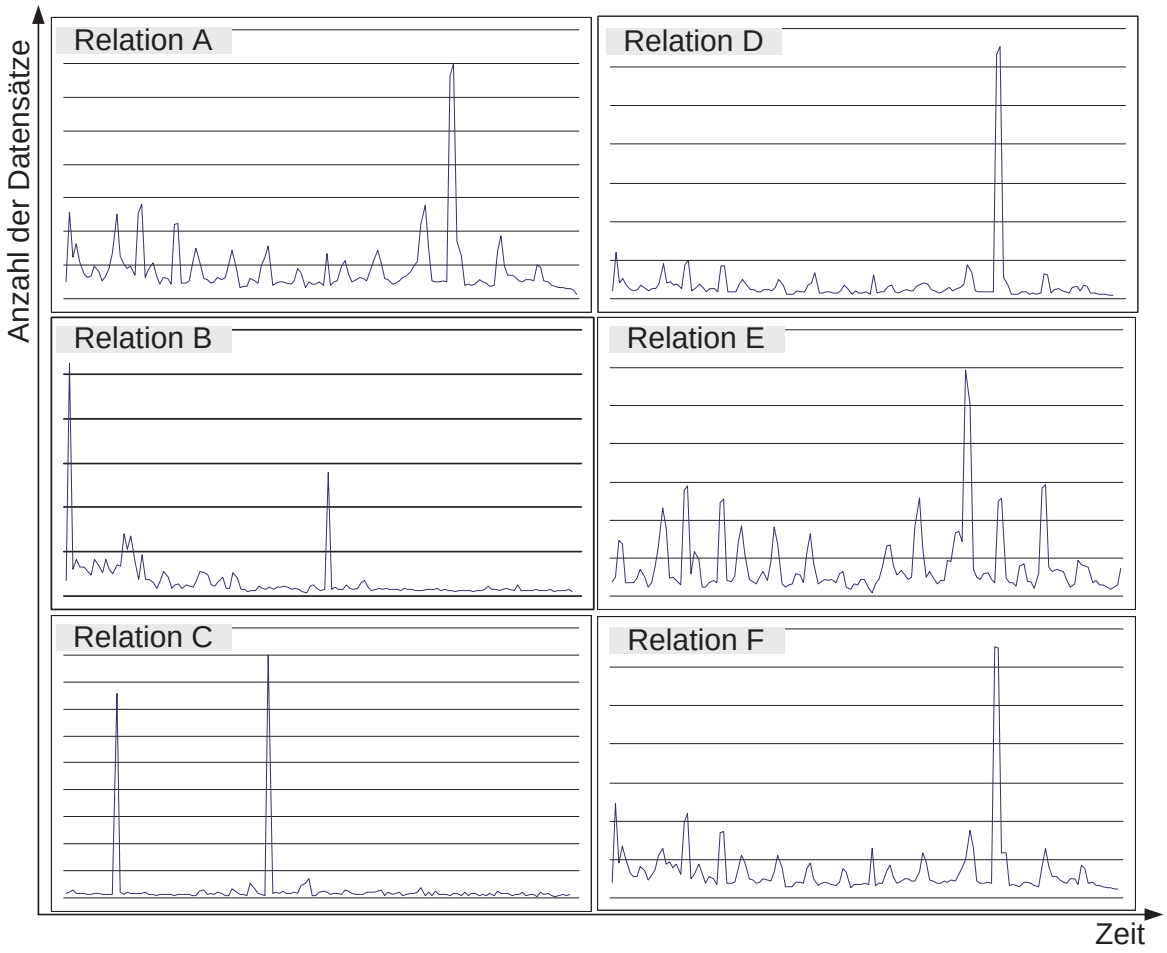


Abbildung 4.10: Analyse der Datenvolumen (Eigene Darstellung)

4.2.2.2 Aktualität und zeitliche Konsistenz

Neben den inhaltlichen Aspekten wird die Datenqualität der Datenwerte auch durch deren zeitlichen Bezug beeinflusst. Diese Eigenschaft bezieht sich auf die Aktualität der Datenwerte und die Konsistenz des Datenbestandes in zeitlicher Hinsicht. Aktualität bezieht sich auf die Repräsentation des derzeitigen Realweltzustands im Datenbestand. Konsistenz des Datenbestandes bezieht sich auf die Eigenschaft, dass alle auf einen bestimmten Zeitpunkt oder Zeitraum bezogenen Daten den entsprechenden Realweltzustand abbilden.

Die Aktualität und zeitliche Konsistenz der Datenwerte ist mit der Erfassung des zeitlichen Bezugs im Datenbankschema verbunden. Aufgrund der beabsichtigten langfristigen Datenspeicherung und der Datenanalysen über längere Zeiträume sind Zeitelemente bei der Modellierung der Datenbankschemata in einem Data-Warehouse-System immer vorhanden.[441] Mit Hilfe dieser Zeitelemente wird der Zeitpunkt oder der Zeitraum erfasst, auf welchen sich die jeweiligen Daten in der Realwelt beziehen.[442] Der Zeitpunkt oder Zeitraum sei als *Gültigkeitszeit* der Daten bezeichnet.[443] Der Zeitbezug weist dabei die vom Benutzer geforderte Granularität auf, welche aufgrund der Zielsetzung analyseorientierter Daten meist im Tages-, Monats- oder Jahresbereich liegt.[444]

Von der Gültigkeitszeit der Datenwerte ist der *Transferzeitpunkt*, sprich der Zeitpunkt des Datenimports, zu unterscheiden.[445] Dabei wird die Granularität der Transferzeiten nur in Ausnahmefällen identisch mit der Granularität der Gültigkeitszeiten sein. Im allgemeinen kann davon ausgegangen werden, dass die Aktualisierungshäufigkeit und somit die Dichte der Transferzeitpunkte eher grösser als die Granularität der Gültigkeitszeiten ist. Die Aktualisierungshäufigkeit hängt von der Änderungshäufigkeit der jeweiligen operativen Systeme ab. Ist die Granularität der Transferzeiten im Vergleich zur Änderungsfrequenz zu gering, werden Änderungen verzögert in die Datenbestände übernommen. Für den Zeitraum zwischen der Datenänderung und dem Datentransfer sind die Daten im Vergleich zur Realwelt dann nicht aktuell. Eine vollkommene Synchronisation zwischen Realwelt und den Datenbeständen kann allerdings aufgrund technischer und organisatorischer Restriktionen nicht erreicht werden. Zur Speicherung der Transferzeitpunkte reicht im allgemeinen ein Zeitstempel aus, der den Zeitpunkt der letzten erfolgreichen Durchführung abbildet. Bei der Gültigkeitszeit sind dagegen meist Zeiträume zu berücksichtigen.

Die Aktualität in den operativen Systeme und die Aktualität in der zentralen Data-Warehouse-Datenbasis sind zu unterscheiden. In den operativen Systemen sind

[441] Vgl. Holthuis (1999), S. 136; Stock (2001), S. 81.
[442] Vgl. Eicker (2001), S. 73.
[443] Vgl. Stock (2001), S. 111.
[444] Vgl. Stock (2001), S. 126.
[445] Vgl. Stock (2001), S. 126f.

meist keine Zeitelemente abgebildet und der Datenbestand bezieht sich im Sinne eines Schnappschusses auf den Zeitpunkt des Datenzugriffs.[446] Die zeitpunktgenaue Betrachtung aus den operativen Systemen wird dann in eine, durch Transferzeitpunkte festgelegte, zeitraumbezogene Betrachtungsweise transformiert. Jeder Datentransfer bezieht von den operativen Systemen Schnappschüsse aus dem Unternehmensgeschehen und bildet diese in die Data-Warehouse-Datenbasis ab. Die Schnappschüsse sind dann, bezogen auf die Data-Warehouse-Datenbasis, bis zum nächsten Transferzeitpunkt gültig und als aktuell einzustufen. Die Aktualität der Daten der Data-Warehouse-Datenbasis hängt somit von den Transferprozessen ab. Der Zusammenhang zwischen Änderungen in der Realwelt, der Aktualität, der Transferzeit und der Gültigkeitszeit von Datenwerten ist in Abbildung 4.11 zusammengefasst.

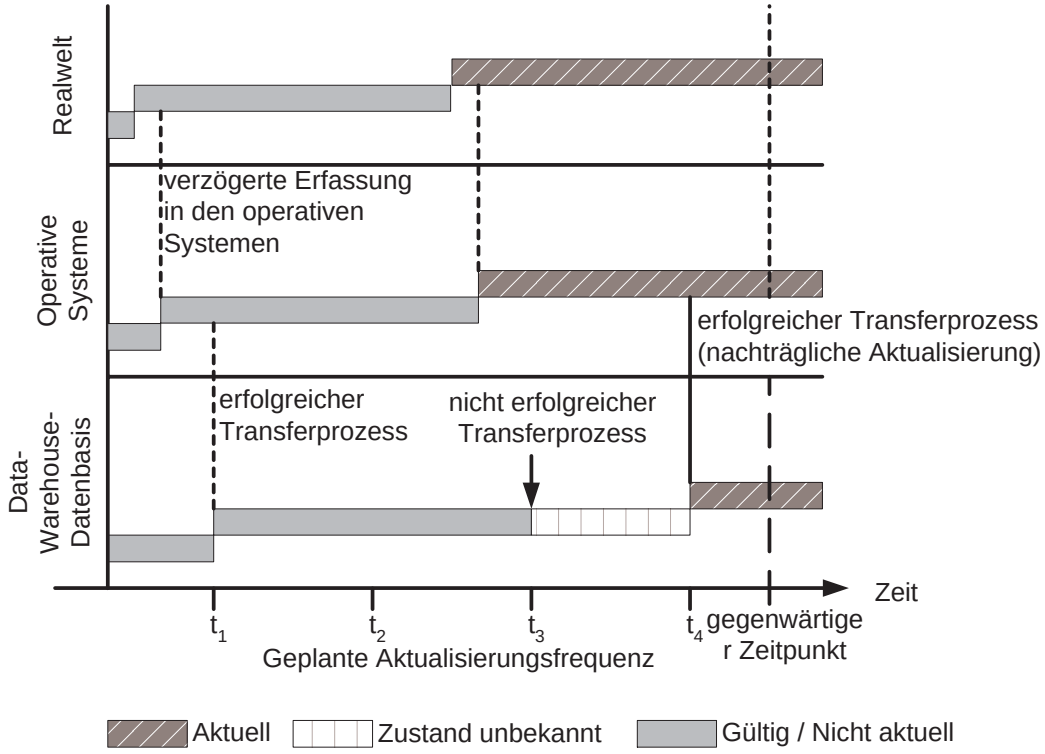


Abbildung 4.11: Zeitlicher Zusammenhang der Datenwerte (Eigene Darstellung)

Aufgrund dieser Zusammenhänge hängt die Aktualität der Daten eines Data-Warehouse-Systems einerseits von den Datenerfassungsprozessen in den opera-

[446] Vgl. Stock (2001), S. 81; Eicker (2001), S. 73.

tiven Systemen aber andererseits auch von den Datentransferprozessen ab. Durch die Erfassungsfrequenz determinieren die Datenerfassungsprozesse die Aktualität und zeitliche Konsistenz der jeweiligen operativen Datenbestände. Hier soll die Prüfung der Aktualität operativer Datenbestände allerdings ausgeklammert und auf operativer Ebene ein aktueller Datenbestand angenommen werden.[447]

Aufgabe der in Abschnitt 2.4.2.3 erläuterten Transferprozesse ist es, die Daten zu extrahieren und einen konsistenten, zentralen Datenbestand abzubilden. Aufgrund des fehlenden Zeitbezugs in den operativen Systemen, ist dieser im Rahmen der Transferprozesse zu explizieren und im Datenbestand festzuhalten. Im allgemeinen werden die Transferprozesse durch einen Ausführungsplan initiiert, wobei die Transferzeiten häufig, jedoch nicht notwendigerweise, in periodischen Zyklen stattfinden. Die Spezifikation der Transferzeiten betrifft den Aspekte der Designqualität. Die Ausführungsqualität wird dagegen durch die tatsächliche Ausführung der Transferprozesse beeinflusst und steht im Fokus der weiteren Betrachtung. In der Praxis werden festgelegte Transferprozesse teilweise nicht gestartet oder abgebrochen und aufgrund zahlreicher Ursachen nicht vollständig durchgeführt. Neben unvollständigen Daten sind zeitlich inkonsistente und nicht aktuelle Datenbestände die Folge. Zur Sicherstellung eines aktuellen und zeitlich konsistenten Datenbestandes gilt es, den Ausführungsplan möglichst konform einzuhalten und bei Verzögerungen regelnd einzugreifen. Zur Durchführung der Transferprozesse ist die Festlegung eines Zeitfensters üblich. Sofern möglich, können Transferprozesse innerhalb des Zeitfensters erneut gestartet werden. Nach Abschluss des Zeitfensters sollten dann alle notwendigen Daten extrahiert und ein neuer sowohl zeitlich konsistenter als auch inhaltlich widerspruchsfreier Datenbestand vorliegen.

Zunächst ist als Indikator für die Aktualität der Datenwerte der letzte erfolgreiche Transferzeitpunkt denkbar.[448] Allerdings ist diese Angabe insofern nicht vollständig, als sie die Beziehung zwischen geplanten Ausführungszeitpunkte und den tatsächlich ausgeführten Transferprozessen nicht berücksichtigt.[449] Aufgrund

[447] Die Aktualität der Datenwerte lässt sich beispielsweise durch empirische, meist auf Stichprobenbasis durchgeführte, Prüfungen testen. Siehe hierzu auch Helfert et al. (2001), S. 35.

[448] Vgl. hierzu Helfert et al. (2001), S. 13.

[449] Im vorliegenden Fallbeispiel werden generell alle Transferprozesse täglich ausgeführt, so dass die

dieser Beziehungen finden sich drei Aktualitätszustände:

(a) Zunächst kann ein Datenwert im Vergleich zu den operativen Systemen aktuell sein. Er ist beim letzten erfolgreichen Transferprozess berücksichtigt worden und stimmte (nach Abschluss des Transferprozesses) mit dem Datenwert des operativen Systems überein.

(b) Die Aktualität und Gültigkeit eines Datenwertes ist nicht bekannt, wenn Transferprozesse geplant, aber nicht erfolgreich abgeschlossen wurden.

(c) Abschliessend ist ein Datenwert nicht mehr aktuell, falls seine Gültigkeitszeit in der Vergangenheit liegt.

Angaben über die Gültigkeitszeit und Aktualität von Datenwerten können auf Ebene von Datenbanken, Relationen oder Datentupeln, aber auch einzelnen Werten erfolgen. Dabei ist der entsprechende Speicherbedarf im Gegensatz zur Granularität der Angaben abzuwägen. Durch die Gültigkeitszeit ist es dann für jeden Datenwert im Data-Warehouse-System möglich, dessen Gültigkeit zu einem Zeitpunkt t zu bestimmten. Für einen gegenwärtigen Zeitpunkt t' ist es zudem möglich, die aktuellen Datenwerte zu ermitteln. Es lassen sich dann für bestimmte Zeitpunkte oder Zeiträume deren zugehörige Daten selektieren und anhand oben erläuterter Integritätsbedingungen auf Glaubwürdigkeit prüfen.

4.2.3 Auswertung der Datenqualität

Aufbauend auf den Ergebnissen in Kapitel 3 wurde das Konzept des proaktiven Datenqualitätsmanagements erarbeitet. Dieses wurde im Rahmen der Projektarbeit diskutiert und durch projektspezifische Anforderungen konkretisiert. Nach der Betrachtung der Problemfelder und projektspezifischen Anforderungen an die Datenqualität und den daraus abgeleiteten Folgerungen wurden anschliessend Möglichkeiten zur Spezifikation von Bedingungen für die Datenqualitätsmerkmale Glaubwürdigkeit und des zeitlichen Bezugs erörtert. Hier wurden neben den allgemein üblichen Integritätsbedingungen weitere Möglichkeiten dargestellt.

Angabe der letzten erfolgreichen Prozessdurchführung als Indikator für Aktualität durchaus aussagekräftig ist.

Anhand der deskriptiven Statistik und den Techniken des Data Mining wurde auf die Qualitätsanalyse und die Beschreibung von charakteristischen Eigenschaften in den Daten eingegangen. Es wurden Möglichkeiten untersucht, die nicht qualitätsrelevanten Schwankungen abzubilden. Hierbei ist die Aggregation von Daten und die Bildung von Plausibilitätsintervallen auf Basis stochastischer Modelle zu nennen. Die Betrachtung der Datenvolumen wurde als eine weitere, wichtige Eigenschaft von Datenbeständen und den Transferprozessen erläutert. Die Aktualität und zeitliche Konsistenz von Datenbeständen wurden dann als andere bedeutende Datenqualitätsmerkmale betrachtet. Zusammenfassend lässt sich das in Abbildung 4.12 dargestellte Ergebnis zur Prüfung der Datenqualiät festhalten.

Im folgenden Abschnitt soll abschliessend eine Möglichkeit zur Auswertung der Datenqualität und deren Repräsentation in Kennzahlen vorgeschlagen werden. Durch Kennzahlen auf detaillierter Ebene können dann entsprechende Qualitätsvorgaben für die Transferprozesse und Datenbestände angegeben und mittels hierarchischer Regelkreise die Datenqualität sichergestellt werden. Zusätzlich können mit Hilfe der Qualitätsprüfungen Qualitätsaussagen ermittelt und den Endanwendern Informationen über die Datenqualität zur Verfügung gestellt werden. Damit eine direkte Interpretation dieser Aussagen durch den Datenverwender möglich wird, sind sie zu verständlichen Kennzahlen zu aggregieren. Auf höchster Aggregationsstufe wäre eine Aussage anhand von drei Zuständen für einzelne Berichte wünschenswert:

- Die Daten sind verwendbar (z. B. Kennzeichnung grün).
- Die Daten sind eingeschränkt verwendbar (z. B. Kennzeichnung gelb).
- Die Daten sind nicht zu verwenden (z. B. Kennzeichnung rot).

Zur weitgehenden Automatisierung der Qualitätsprüfung ist die Integration des Qualitätssystems in die Metadatenverwaltung notwendig (vgl. Abbildung 4.4). Beim Über- oder Unterschreiten bestimmter Qualitätsgrenzwerte sowie beim Auftreten bestimmter Ereignisse, können über Benachrichtigungsregeln dann entsprechende Personen oder Personengruppen informiert werden (z. B. der Datenqualitätsverantwortliche). Sie können die Datenqualitätsprobleme analysieren und geeignete Massnahmen einleiten.

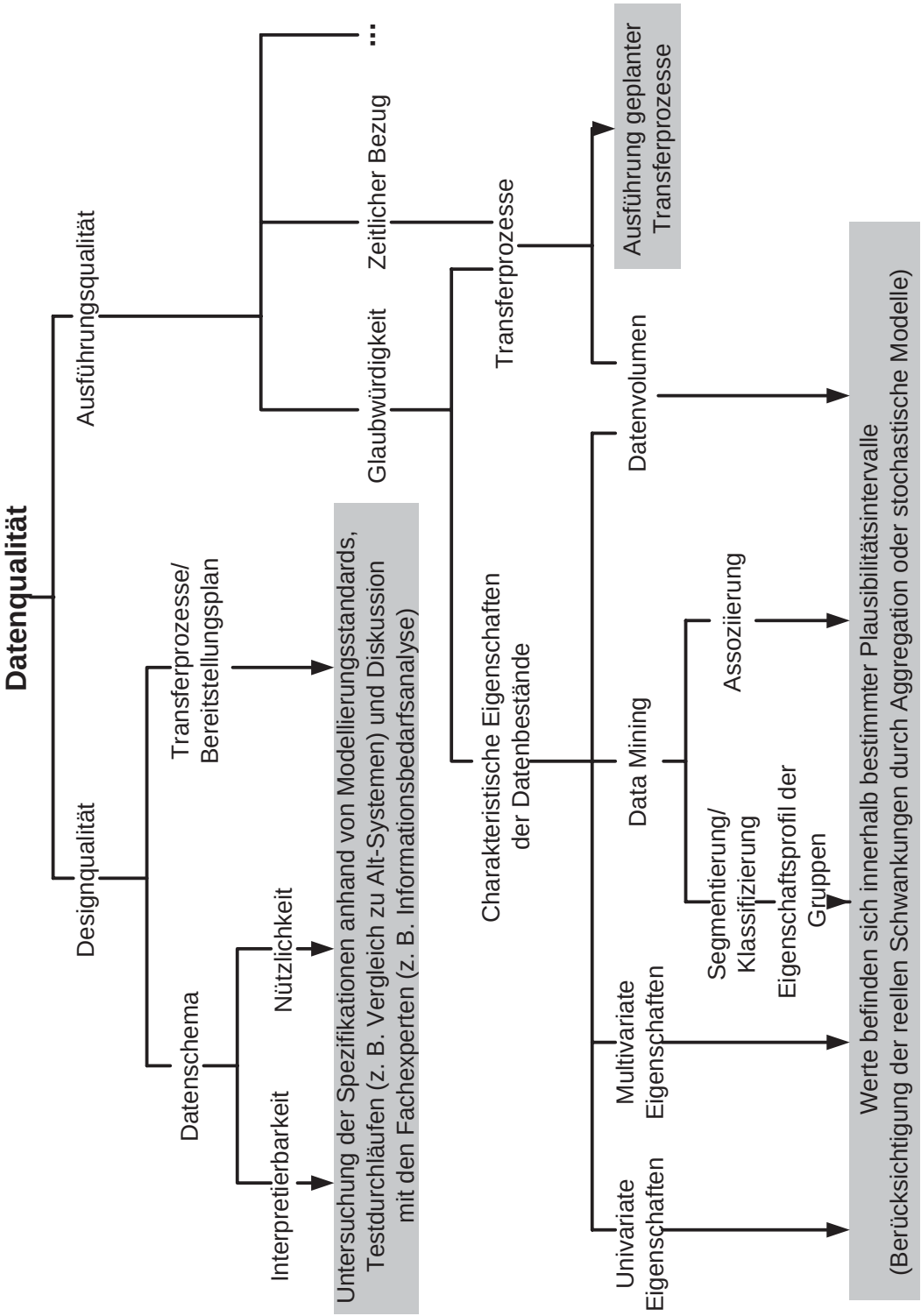


Abbildung 4.12: Datenqualitätsmessung (Eigene Darstellung)

Eine Menge von Regeln repräsentiert einzelne Qualitätsforderungen. Aufbauend auf den Regeln können dann Kennzahlen gebildet, Sollwerte vorgegeben und der aktuelle Qualitätszustand ermittelt werden. Basis zur Regelspezifikation können die in Abschnitt 2.3.1 dargestellten Integritätsregeln bilden. Diese sind relativ einfach in SQL abzubilden und umzusetzen. Ihre einfache Syntax ermöglicht eine schnelle Implementierung auf verschiedenen Plattformen, eine einfache Speicherung und eine Möglichkeit zur flexiblen Erweiterung. Mehrere Regeln werden dann zu einer Regelmenge zusammengefasst. Aufgrund der Notwendigkeit anwendungsspezifisches Wissen zu berücksichtigen, sind bei der Bildung und Validierung der Regelmenge Fachexperten, Datenanalysten und Datenqualitätsverantwortliche miteinzubeziehen. Eine unvollständige oder inkorrekte Regelmenge führt zu mangelhaften und falschen Qualitätsaussagen. Ist die Regelmenge ausreichend korrekt, sollte der Eingriff eines Qualitätsanalysten bzw. eines Datenqualitätsverantwortlichen nur in Ausnahmesituationen bei Über- oder Unterschreiten bestimmter Grenzwerte notwendig sein. In Tabelle 4.10 sind einige Beispiele der hier diskutierten Möglichkeiten aufgelistet.

Die so aus dem jeweiligen Anwendungszusammenhang erstellten Bedingungen können dann relativ leicht in einer an SQL orientierten Notation spezifiziert werden. Im Gegensatz zu den herkömmlichen Integritätsbedingungen, wie sie beispielsweise im Bereich der Integritätssicherung angewendet werden, wird hier allerdings deren Erfüllung nicht notwendigerweise verlangt. Die Regeln sind lediglich in bestimmten, festzulegenden Qualitätstoleranzen anhand von Sollvorgaben einzuhalten. Beispielsweise kann es in bestimmten Anwendungsfällen genügen, dass lediglich 95% der Attributwerte in einer Relation besetzt sind. Die entsprechende Qualitätsvorgabe könnte dann lauten:

```
Regel_x := Kunde.Telefonnummer NOT NULL
```

$$\frac{\texttt{COUNT(*) FROM Kunde WHERE Regel\_x = true}}{\texttt{COUNT(*) FROM Kunde}} \geq 0,95$$

Die Bedingungen werden dann zur Prüfung der Datenbestände und Transformationsprozesse verwendet, indem die Anzahl der Regelverletzungen ausgewertet werden. In diesem Zusammenhang sind die oben erwähnten Qualitätsprüfungen beim Transferprozess zu nennen. Eine Verletzung der hier festgelegten Integritäts-

Qualitätsmerkmal		Ansatz zur Qualitätsprüfung	Beispiel
Glaubwürdigkeit	Wertebereichs- und Attributbedingungen	Attributwert entspricht dem festgelegten Datentyp / Datenformat	`Geburtsdatum IS Date`
		Attributwert befindet sich in den angegeben Wertebereichen	`'01.01.1950'` $\leq$ `Geburtsdatum` $\leq$ `Today`
		Für Pflichtfelder sind keine Nullwerte eingetragen	`Geburtsdatum NOT NULL`
	Tupel und Relationen-bedingungen	Beziehungen zwischen einzelnen Tupeln innerhalb einer Relation	`(ProduktKat =` $4X$ `AND Bilanzschlüssel =` $1X$`)` $\Rightarrow$ `Guthaben` ≥ 10000
		Schlüsselwerte sind einmalig	`Kontonummer NOT NULL AND UNIQUE`
	Referentielle Integrität	Fremdschlüsselbeziehungen sind vollständig	`Konto.Kundennummer IN Kunde.Kundennummer`
	Sonstige Inte-gritätsbedingungen	Beziehungen zwischen Attributen und Strukturen in den Datenbeständen	Summe der Guthaben von System A entspricht altem Guthaben + Kontenbewegungen
			Durchschnittliches Kreditvolumen verhält sich „weitgehend“ linear zur Anzahl der Kunden (d. h. das Kreditvolumen entspricht einem prognostizierten Wert bzw. Plausibilitätsintervall)
		Datenvolumen	Anzahl der Kontenbewegungen befinden sich in einem bestimmten Plausibilitätsintervall (z. B. $100000 \leq$ `(COUNT (*) FROM KONTO WHERE TRANSAKTION_DATE = Today)` ≤ 110000)
Aktualität		Prüfung auf Ausführung geplanter Transferprozesse	Transferprozess T ist erfolgreich und vollständig zum geplanten Zeitpunkt t abgeschlossen (z. B. `GEPLANT_ZEITPUNKT = START_ZEITPUNKT AND STATUS = 'completed'`)

Tabelle 4.10: Beispiele zur Qualitätsspezifikation und -prüfung

bedingungen hat eine entsprechende Reaktion zur Folge. So werden bestimmte Operationen nicht ausgeführt oder ein Default-Wert eingefügt, was zu Fehler- oder Warnmeldungen führt (vgl. Tabelle 4.1). Diese Meldungen werden dann in einer Protokolldatei gespeichert. Durch Auswertung der Protokolle können dann Qualitätsaussagen über die Datenqualität der Liefersysteme gemacht werden. Qualitätsaussagen über die betroffenen Datenwerte in den Zieldaten lassen sich in Abhängigkeit der durchgeführten Folgeaktion ermitteln.

Durch Auswertung der transferprozessbezogenen Metadaten lassen sich Aussagen über die Aktualität der Daten gewinnen.[450] Für das Fallbeispiel sind hier die bereits in der Metadatenverwaltung vorgehaltenen Kontrollinformationen über die Transformationsprozesse zu nennen. Wird das in Abbildung 4.3 dargestellte Datenschema der Metadaten betrachtet, kommt insbesondere ein Vergleich von Attributen der Entität `SESSION_LOG` in Betracht. Hier können `GEPLANT_ZEITPUNKT`, `START_ZEITPUNKT`, `END_ZEITPUNKT` und `STATUS` verglichen und Aussagen über die tatsächliche Ausführung einzelner Transferprozesse im Vergleich zur geplanten Aktualisierungsfrequenz gewonnen werden. Weiter können die in POWERCENTER verwalteten Metadaten über das Attribut `PC_SESSION_ZEITPUNKT` identifiziert und detailliert einzelne Transformationsschritte analysiert werden.[451]

Nach einer Qualitätsprüfung anhand der Regelmenge können die nicht erfüllten Regeln zu den durch die Regel tangierten Daten in Beziehung gesetzt werden. Diese können je nach Granularität auf einzelne Datwerte, eine Tupelmenge, Relationen oder die Datenbank bezogen sein. Die Qualitätsangaben können dann eventuell durch weitere werkzeugunterstützte oder manuelle Qualitätsanalysen des Datenqualitätsverantwortlichen, durch Qualitätsaussagen der Datenverwender oder empirische Prüfungen auf Stichprobenbasis ergänzt werden. Als Ergebnis der Qualitätsprüfung können dann, wie in Abbildung 4.13 dargestellt, für beliebige Datenmengen die jeweils nicht erfüllten Bedingungen ermittelt werden. Für ausgewählte Datenmengen können dann Qualitätsvorgaben durch entsprechende

[450] Vgl. Abschnitt 4.2.2.2.

[451] In diesem Zusammenhang ist die von INFORMATICA angebotene Schnittstelle Metadata Exchange zu nennen, die eine relationale Sicht auf die Metadaten in POWERCENTER und somit Abfragen über SQL zulässt.

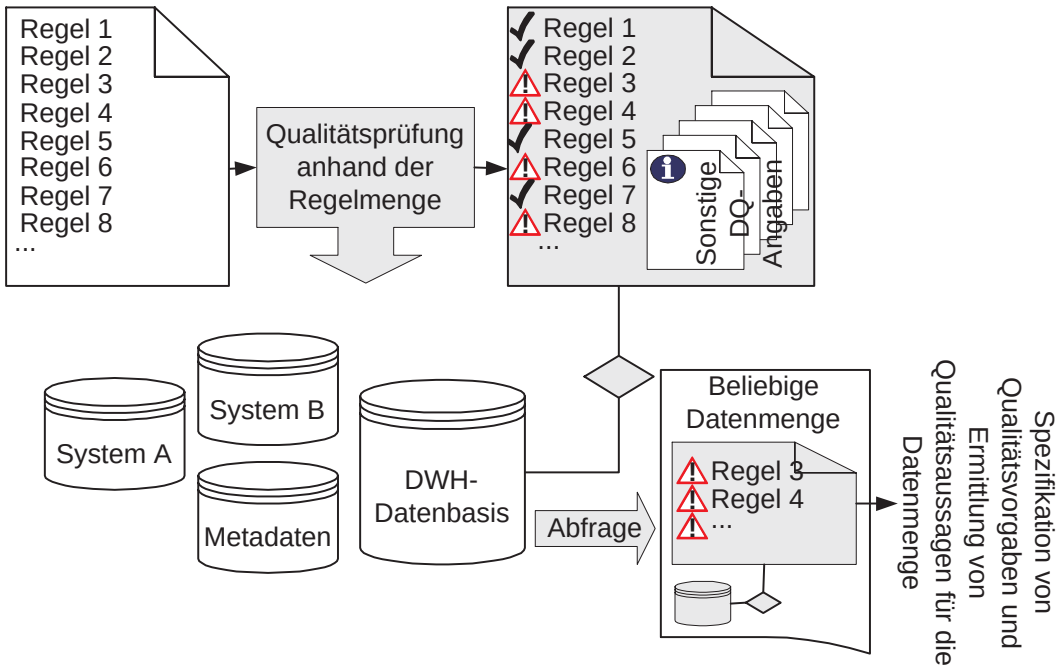


Abbildung 4.13: Ermittlung der nicht erfüllten Bedingungen für beliebige Datenmengen (Eigene Darstellung)

Kennzahlen festgelegt und geprüft werden. Einige Beispiele für Kennzahlen der Datenqualität sind in Tabelle 4.11 aufgeführt. Die nicht erfüllten Regeln geben dann Hinweise auf die Ursachen mangelnder Datenqualität und bieten so Anhaltspunkte zur Ermittlung von Qualitätssicherungsmassnahmen. Für Standardauswertungen lassen sich analog Qualitätsvorgaben spezifizieren und aggregierte Kennzahlen bilden. Die Qualitätsaussagen können dann dem Endanwender in aggregierter Form oder durch entsprechende Qualitätskennzeichnungen (z. B. Färbung) zur Verfügung gestellt werden.

Die hier dargestellten Qualitätskennzahlen stellen lediglich einige Möglichkeiten dar. In praktischen Anwendungsfällen sind noch eine Vielzahl von Kennzahlen auf Basis der Regeln denkbar.[452] Aufgrund des hohen Anwendungsbezugs sind die konkreten, anwendungsspezifischen Regeln und Kennzahlensysteme mit den Fachexperten, Datenqualitätsverantwortlichen und Datenanalysten zu erstellen.

[452] Vgl. Mandke und Nayar (1998), S. 237f.

Absolute Anzahl der verletzten Bedingungen	
Vollständigkeit	$\frac{\text{Gesamtzahl der Datentupel} - \text{Anzahl von Nullwerten}}{\text{Gesamtzahl der Datentupel}}$
Glaubwürdigkeit	$\frac{\text{Anzahl widerspruchsfreier Werte}}{\text{Gesamtzahl der Datentupel}}$
Widerspruchsfreiheit	$1 - \frac{\text{Gesamtzahl verletzten Bedingungen}}{\text{Anzahl spezifizierter Bedingungen}}$
Aktualität	$\frac{\text{Anzahl erfolgreicher Transferprozesse}}{\text{Anzahl geplanter Transferprozesse}}$
Datenqualitätsrate	$\frac{\text{Aktuelle Qualitätskennzahl}}{\text{Akzeptierte Qualitätskennzahl}}$

Tabelle 4.11: Exemplarische Qualitätskennzahlen für beliebige Datenmengen

Kapitel 5

Zusammenfassung und Ausblick

In diesem Kapitel sollen abschliessend die wichtigsten Ergebnisse und Erkenntnisse der Arbeit zusammengefasst und kritisch beleuchtet werden. Dabei sollen noch ungelöste Fragestellungen identifiziert und Anhaltspunkte für weitere Forschungen aufgezeigt werden. Neben Vorschlägen für das weitere Vorgehen im Rahmen der Fallstudie sollen auch allgemeine Forschungsfragen erörtert werden.

5.1 Zentrale Forschungsergebnisse

Einführend wurden in Kapitel 2 die konzeptionellen Grundlagen der Arbeit dargestellt, wobei in Abschnitt 2.4.2.5 zusammenfassend die zentralen Betrachtungsebenen eines Data-Warehouse-Systems erläutert wurden. Als eine Gesamtheit kann ein Data-Warehouse-System zunächst zur Bereitstellung analytischer Daten für Entscheidungsträger betrachtet werden. Durch Deduktion kann dann die Untersuchung auf einzelne Komponenten des Systems bezogen werden. Diese Systemkomponenten erfüllen durch Ausführen bestimmter Funktionen zusammen den Zweck des Gesamtsystems. Wesentliche Komponenten sind die Datenhaltungssysteme, die Transferkomponenten, die Endbenutzerwerkzeuge und die Metadatenverwaltung als integratives Element. Untersucht man die Komponenten weiter, besteht ein Data-Warehouse-System im Kern aus Datenbanken, Software-, Hardware- und Kommunikationskomponenten. Im Rahmen der Datenqualität ist die Betrachtung der Datenhaltungssysteme mit deren Datenschemata zentral. Ausgehend von dieser Betrachtungsweise können verschiedene Datenschemata auf unterschiedlichen Beschreibungs- und Architekturebenen erkannt

werden. Diese stehen durch die konzeptionellen und logischen Modelle, das Mapping sowie die Transferprozesse und die internen Modelle miteinander in Beziehung. Das System aus Modellen auf der konzeptionellen und logischen Ebene als auch deren physische Umsetzung ist konsistent zu halten. Inkonsistenzen resultieren in Datenqualitätsproblemen. Als Folge dieser Betrachtungsweise, ist das Data-Warehouse-System als ein Gegenstandsbereich zur Datenhaltung und Datenbereitstellung für analytische Daten ganzheitlich zu erfassen. Insbesondere ist der gesamte Datenfluss, von der Datenentstehung bis zur Datenverwendung zu berücksichtigen.

In Kapitel 3 wurden Fragestellungen bezüglich der Datenqualität in Data-Warehouse-Systemen erörtert. Vom allgemeinen Qualitätsbegriff ausgehend wurden Qualitätssichten und die Unterscheidung zwischen Design- und Ausführungsqualität erarbeitet. Eine Untersuchung von Ansätzen in der Literatur zeigte, dass bislang kein allgemein akzeptierter Datenqualitätsbegriff existiert. Allerdings finden sich bereits zahlreiche Datenqualitätskriterien und -listen für unterschiedliche Anwendungsbereiche. Weiter wurde die Problematik der Datenqualität in Data-Warehouse-Systeme bislang nicht ausreichend untersucht. Aufgrund dieses Defizits wurde zunächst eine empirische Untersuchung zur Problematik der Datenqualität in Data-Warehouse-Systemen durchgeführt, die insbesondere wichtige Probleme und relevante Datenqualitätskriterien erörterte.

Es zeigte sich, dass Datenqualität in den meisten Data-Warehouse-Systemen problematisch ist. Vor allem werden Probleme aufgrund inkorrekter, unvollständiger und inkonsistenter Daten genannt. Ursachen liegen teilweise in systemtechnischen Problemen aber insbesondere in organisatorischen Mängeln und in den operativen Systemen begründet. Zur Sicherstellung der Datenqualität finden sich sowohl technische Möglichkeiten in Form von Datenbereinigung als auch organisatorische Massnahmen, wenngleich diese bislang in der Praxis noch nicht zufriedenstellend sind. Es zeigte sich insbesondere, dass im Bereich der Qualitätsvorgaben und deren Überprüfung hoher Lösungsbedarf besteht. Werden die eingesetzten Lösungen weiter untersucht, finden sich bereits in einigen Unternehmen Ansätze für ein Datenqualitätsmanagement. Ein wichtiger Bestandteil nimmt dabei die Qualitätsprüfung ein, die üblicherweise bei der Datenanlieferung anhand von ein-

fachen Integritätsbedingungen, subjektiv durch den Endanwender und durch die Analyse des Datenbestandes vorgenommen wird. Neben dem Urteil des Datenverwenders werden dabei im wesentlichen Konsistenzbedingungen innerhalb und zwischen Datenbeständen untersucht, Prüfungen von Plausibilitäten vorgenommen sowie die Transferprozesse anhand von Systemprotokollen analysiert.

Werden die in der Praxis relevanten Eigenschaften für qualitativ hochwertige Daten untersucht, stellt sich der Begriff der Datenqualität als sehr umfassend dar. Er berücksichtigt die Aspekte der Datenmodellierung, der Datenwerte und des Gesamtsystems. Datenqualität ist eng mit Aspekten der Softwarequalität, der Qualität von Datenmodellen und der Anwenderzufriedenheit verbunden, wobei eine Abgrenzung bislang nicht eindeutig vorgenommen wird bzw. werden kann. Es zeigte sich weiter, dass die Unterscheidung zwischen Designqualität und Ausführungsqualität geeignet erscheint, die Datenqualität in Data-Warehouse-Systemen zu planen und zu lenken. Designqualität umfasst im wesentlichen die Aspekte des Datenschemas und der Funktionsanforderungen, während Ausführungsqualität auf die Datenwerte und die Funktionsausführung bezug nimmt. Als wichtige Qualitätskriterien der Datenwerte werde Konsistenz, Korrektheit, Vollständigkeit und Aktualität genannt. In bezug auf die Datenbeschreibung durch Datenschemata werden Interpretierbarkeit und Identifizierbarkeit als wichtige Qualitätskriterien gefordert.

Auf Basis dieser Ergebnisse wurde im zweiten Teil des Kapitels 3 ein Konzept für ein proaktives Datenqualitätsmanagement erarbeitet. Charakteristisch für dieses Konzept ist die ganzheitliche Betrachtungsweise im Sinne eines umfassenden Qualitätsmanagements. Es beinhaltet ein Qualitätsmanagementsystem, Methoden und Werkzeuge sowie die Integration der Datenqualitätsziele in das Zielsystem der Unternehmung. Es umfasst den Teil des Informationsmanagements, der die qualitativen Aspekte von Daten betrachtet. Kern des Datenqualitätsmanagements bilden die Qualitätsplanung und Qualitätslenkung auf operativer Ebene. Auf Grundlage des Prinzips der Regelung wurde ein hierarchisches Regelkreismodell für das operative Datenqualitätsmanagement erarbeitet. Ziel ist es, die operationalisierten, konkreten Qualitätsvorgaben durch die Qualitätslenkung mit Hilfe von Qualitätssicherungs- und Qualitätsverbesserungsmassnahmen zu erreichen und einzu-

halten. Aufbauend auf den Prinzipien des Konzepts und den Erkenntnissen der empirischen Untersuchung wurden dann zentrale Anforderungen an das proaktive Datenqualitätsmanagement abgeleitet. Diese wurden anschliessend mit ausgewählten Ansätzen in der Literatur verglichen. Obgleich Ansätze in der Literatur existieren, gibt es dennoch zahlreiche offene Fragestellungen. Insbesondere sind die Qualitätsplanung und die Qualitätsprüfung ein aktuelles Forschungsgebiet.

Im Rahmen der Forschungsarbeit im Kompetenzzentrum „Data Warehousing 2“ wurde das erarbeitete Konzept mit Partnerunternehmen diskutiert und innerhalb der Projektarbeit mit einer Schweizer Universalbank ein Ansatz für ein operatives Datenqualitätsmanagement zur Planung und Messung der Datenqualität erstellt (vgl. Kapitel 4). Kennzeichen des Ansatzes ist die Integration des Qualitätsmanagements in die Metadatenverwaltung. Es erlaubt so ein werkzeugunterstütztes und weitgehend automatisches Management der Datenqualität. Das Datenqualitätssystem besteht aus einer Regelmenge, einem Benachrichtigungssystem und den Qualitätsaussagen. Qualitätsaussagen werden durch automatische Prüfungen anhand der Regelmenge ermittelt und in einer Datenbank abgelegt. Diese werden dann durch ein Qualitätskennzahlensystem zu verdichteten Qualitätsaussagen aggregiert. Eine zentrale Rolle spielt der Datenqualitätsverantwortliche, der in Zusammenarbeit mit den Endanwendern bzw. Fachbereichsvertretern Qualitätsvorgaben erfasst und in eine Spezifikation überführt. Reklamationen über die Datenqualität und erkannte Datenqualitätsprobleme können über eine Erfassungskomponente festgehalten und dem Datenqualitätsverantwortlichen zugeführt werden. Beim Unter- oder Überschreiten bestimmter Grenzwerte für die Datenqualität sowie dem Auftreten festgelegter Ereignisse, werden ausgewählte Personen (z. B. der Datenqualitätsverantwortliche) durch das Benachrichtigungssystem in Kenntnis gesetzt. Diese können dann adäquate Massnahmen einleiten.

Ein Vorschlag für den Grobablauf der Qualitätsplanung und -lenkung ist zusammenfassend in Abbildung 5.1 dargestellt, deren Notation sich an ereignisgesteuerten Prozessketten orientiert.[453] Auf der linken Seite ist die dynamische Entwicklung der Qualitätspezifikation und deren Abbildung in einer Regelmenge dargestellt. Ausgehend von der dynamischen Änderung der Qualitätsforderungen wird

[453] Vgl. zur Notation ereignisgesteuerter Prozessketten z. B. Scheer (1998), S. 49ff.

die Spezifikation laufend angepasst und erweitert. Resultat der Qualitätsplanung sind konkrete Sollwertinformationen für die Data-Warehouse-Komponenten. Hier werden insbesondere die Anforderungen an die Datenbestände durch Datenschemata und Integritätsbedingungen sowie an die Transferprozesse durch zeitliche Vorgaben für Aktualisierungsfrequenzen festgelegt und deren Prüfung in einer Regelmenge abgelegt. Durch die Integration der Regelmenge in die Metadatenverwaltung können automatische Qualitätsprüfungen durchgeführt werden. Auf der rechten Seite ist die Qualitätsprüfung durch die Endanwender abgebildet. Ausgehend von subjektiv eingeschätzten Datenqualitätsmängeln werden diese erfasst und an den Datenqualitätsverantwortlichen weitergeleitet, der diese dann prüft. Identifizierte Datenqualitätsprobleme werden durch den Datenqualitätsverantwortlichen in Abstimmung mit den Fachexperten und den Data-Warehouse-Verantwortlichen auf deren Ursachen und Auswirkungen untersucht. Nach deren Relevanz für die Zielerreichung des Data-Warehouse-Systems werden dann Qualitätssicherungsmassnahmen identifiziert und eingeleitet. Zur Qualitätsplanung und -lenkung sind sowohl Fachexperten, Verantwortliche für die Data-Warehouse-Infrastruktur und -Architektur sowie die Datenqualitätsverantwortlichen einzubeziehen. Dieser grobe Vorschlag soll zusammenfassend den prinzipiellen Ablauf verdeutlichen und ist in weiteren Arbeiten zu konkretisieren.

Aufbauend auf diesen Ergebnissen wurden dann in Kapitel 4 Möglichkeiten zur Prüfung der Ausführungsqualität erörtert. Schwerpunkt bilden die Qualitätskriterien Glaubwürdigkeit, Aktualität und zeitliche Konsistenz. Die Prüfung der Glaubwürdigkeit basiert im wesentlichen auf dem Vergleich von verschiedenen Datenbeständen anhand zeitinvarianter Eigenschaften. Diese sind durch eine Vielzahl von Integritätsbedingungen in und zwischen den Datenbeständen konkretisiert. Zur Ermittlung und Beschreibung der charakteristischen Eigenschaften von Daten wurden Methoden der deskriptiven Statistik und des Data Mining untersucht. Allerdings unterliegen Daten in betrieblichen Systemen bestimmten Schwankungen in Form von individuellem Kundenverhalten, dem Marktverhalten oder saisonaler Entwicklungen. Für den Vergleich der Datenbestände im Zeitablauf sind diese in Form von Datenaggregationen, Plausibilitätsintervallen oder Modellen zu filtern. Aktualität und zeitliche Konsistenz werden durch die Erfas-

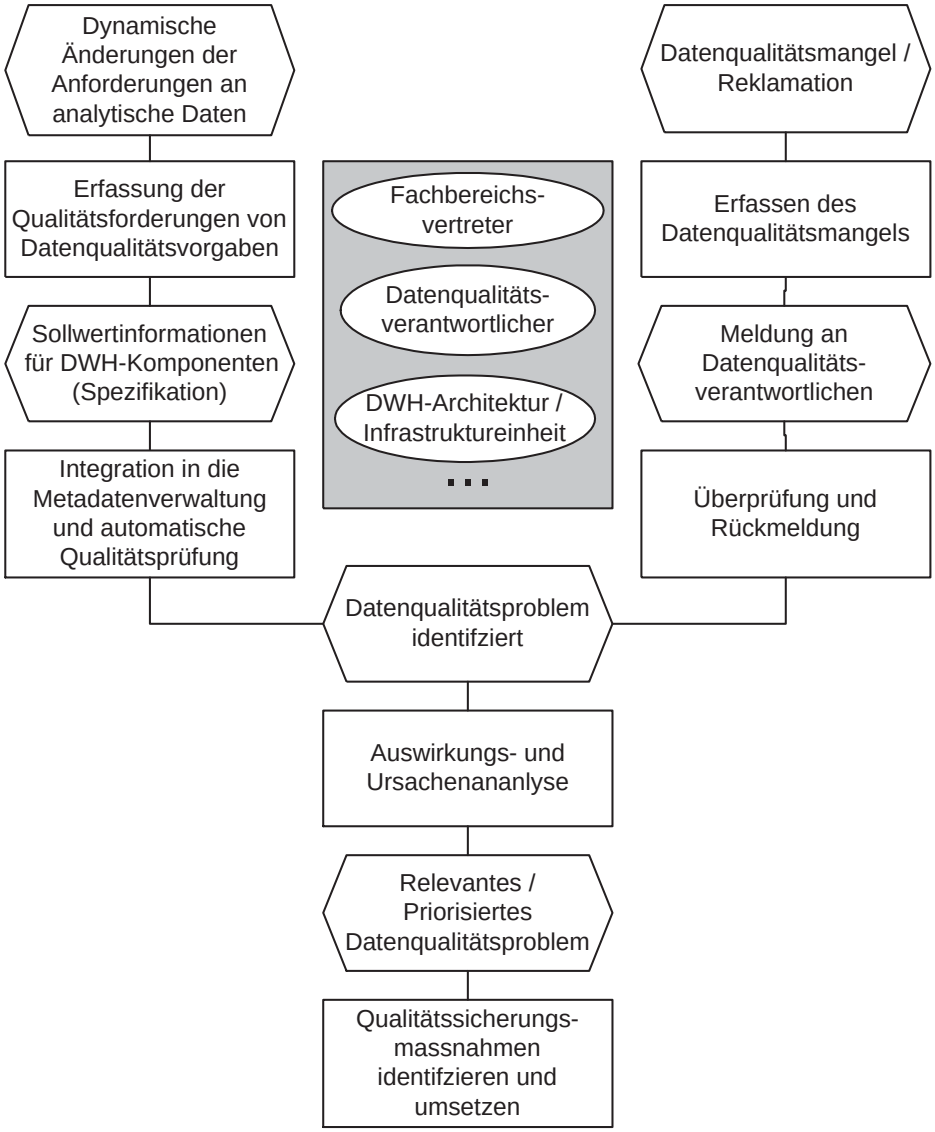


Abbildung 5.1: Grobablauf für die Datenqualitätsplanung und -lenkung (Eigene Darstellung)

sungsfrequenz in den operativen Systemen und die Aktualisierungsfrequenz der Transferprozesse beeinflusst. Für die Ausführungsqualität ist die Ausführung geplanter Erfassungs- und Transferprozesse wesentlich. Die Differenz zwischen geplanten Prozessen und der tatsächlichen Ausführung kann dann als Indikator für die Aktualität und zeitliche Konsistenz der Datenqualität herangezogen werden. Abschliessend wurden in Abschnitt 4.2.3 Möglichkeiten zur Auswertung der Datenqualität und Bildung von verständlichen Qualitätskennzahlen untersucht. Die

Ergebnisse der Qualitätsprüfung können dann den Endanwendern in Form von aggregierten Kennzahlen zur Verfügung gestellt werden und liefern wertvolle Informationen über die Qualität der verwendeten Daten. Weiter ist durch detaillierte Qualitätskennzahlen eine Ursachenanalyse mangelnder Datenqualität möglich.

5.2 Kritische Würdigung des Ansatzes

Im Verlauf der Arbeit zeigte es sich, dass Datenqualität ein zentrales Problem in Data-Warehouse-Systemen darstellt. Datenqualität umfasst zahlreiche Aspekte und ist bislang in Forschung und Praxis nicht ausreichend berücksichtigt. Wenngleich die Arbeit nicht alle Fragestellungen bezüglich der Datenqualität erörtern konnte, wurden wichtige Erkenntnisse für dieses Forschungsgebiet erarbeitet. Zusammenfassend sind

- das Konzept des proaktiven Datenqualitätsmanagements mit einer ganzheitlichen Betrachtungsweise,
- die Unterscheidung zwischen Design- und Ausführungsqualität,
- die Spezifikation und Prüfung der Datenqualität in Data-Warehouse-Systemen hinsichtlich der Qualitätskriterien Glaubwürdigkeit, Aktualität und zeitlicher Konsistenz sowie
- die Integration des Qualitätsmanagements in die Metadatenverwaltung

als zentrale Erkenntnisse zu nennen. Im Rahmen der empirischen Studie konnten Erkenntnisse über die derzeitigen Problemfelder im Bereich Datenqualität und deren Umgang in der Praxis gewonnen werden. Durch die intensive Projektarbeit im Kompetenzzentrum „Data Warehousing 2“, der Fallstudie und im Dialog mit weiteren Unternehmen konnte die Problematik tiefgreifend analysiert, die hier erarbeiten Konzepte reflektiert und auf deren praktische Umsetzungsfähigkeit geprüft werden. In Ergänzung zu dem in Kapitel 4 vorgestellten Ansatz für ein operatives Datenqualitätsmanagement sollen hier zwei weitere Beispiele genannt werden. Zunächst ist das bei der Firma THOMAS COOK AG eingesetzte

Datenqualitätsmanagement-Werkzeug aufzuführen, welches für die Analyse der Datenqualität benutzt wird. Anhand verschiedener Fehlertypen werden technische, inhaltliche und zeitliche Fehlerarten untersucht. Eine Regelmenge legt die notwendigen Prüfungen durch eine an SQL angelegte Spezifikation fest, welche in einer Datenbank verwaltet wird. Bislang finden die Qualitätsprüfungen täglich durch die Qualitätsverantwortlichen der jeweiligen Fachbereiche statt, wobei eine weitere Automatisierung bereits geplant ist. Ein Bildschirmausschnitt des eingesetzten Werkzeugs für das Datenqualitätsmanagement ist in Abbildung 5.2 dargestellt.

Als weiteres Beispiel sei das beim INFORMATIKZENTRUM DER SPARKASSENORGANISATION (SIZ) etablierte Datenqualitätsmanagement aufgeführt.[454] Neben der Analyse des Datenbestandes mit Hilfe von deskriptiver Statistik und Data Mining werden hier insbesondere organisatorische Massnahmen berücksichtigt. Der generelle Ablauf des Datenqualitätsmanagements ist in Abbildung 5.3 dargestellt. Ausgehend von Datenqualitätsmerkmalen werden Datenqualitätsziele identifiziert und durch konkrete Qualitätsvorgaben spezifiziert. Diese werden dann durch Datenqualitätsmetriken gemessen. Bei Datenqualitätsmängeln werden deren Ursachen analysiert und entsprechende Korrekturmassnahmen und präventive Qualitätssicherungsmassnahmen eingeleitet. Präventive Massnahmen berücksichtigen hier insbesondere den Bereich der Systementwicklung und Datenmodellierung durch Standards und Richtlinien. Weiter wurden idealtypische Rollen sowie organisatorische Massnahmen eingeführt. Einige der hier organisatorisch verankerten Rollen sind in Tabelle 5.1 aufgeführt. Durch die organisatorischen Massnahmen wurden klar geregelte Verantwortlichkeiten und Zielvorgaben geschaffen sowie eine regelmässige Qualitätskontrolle beabsichtigt.

Werden kommerzielle Werkzeuge im Bereich der Datenqualität untersucht, existieren neben Datenbereinigungswerkzeugen Werkzeuge zur Qualitätsanalyse des Datenbestandes.[455] Als Beispiel eines Analysewerkzeugs sei das von WIZSOFT angebotene Werkzeug WIZRULE genannt.[456] Es ermittelt anhand von Methoden der Statistik und des Data Mining Regeln für plausible Beziehungen zwischen At-

[454] Vgl. de Fries et al. (1999), S. 513ff.
[455] Eine Liste kommerzieller Werkzeuge findet sich beispielsweise in English (1999), S. 311ff.
[456] Vgl. Meidan (2001).

Rolle	**Rollenbeschreibung**
Chief Information Officer	Verantwortlich für den Aufbau einer Informations- bzw. Datenpolitik und für ein Kennzahlensystem zur Steuerung der Ressourcen.
Datenadministrator	Verantwortlich für die Abstimmung von Datenstrukturen und Modellierungsergebnissen sowie die Berücksichtigung aller benötigten Datenbegriffe im Datenmodell.
Datenbank-administrator	Verantwortlich für die Verwaltung der physischen Datenbanken und alle damit verbundenen Aufgaben, insbesondere die Umsetzung von konzeptionellen Vorgaben in eine physische Datenbank.
Dateneigentümer	Verantwortlich für die Qualität der Informationen, die er produziert.
Datenfachvertreter	Zuständig für die Validierung der Datendefinitionen inklusive der Festlegung der Wertebereiche und Geschäftsregeln.
Datenlieferant (extern)	Verantwortlich für die Qualität der Daten, die er bereitstellt.
Datenqualitäts-beauftragter	Stabsfunktion der Unternehmens- oder DV-Leitung, die mit der Überwachung und Verbesserung der Datenqualität betraut ist.
Informationsvermittler	Zuständig für die Datentransformation.
Prozessdesign-verantwortlicher	Verantwortlich für die Integrität der Geschäftsprozessdefinitionen oder einen Teil der Wertschöpfungskette.
Prozessverant-wortlicher	Verantwortlich für die Integrität der unterstellten Prozesse und für die Qualität aller darin erzeugten Informationen.
Allgemeiner Qualitätsbeauftragter	Verantwortlich für das Qualitätsmanagementsystem eines Unternehmens.

Tabelle 5.1: Zentrale Rollen im Datenqualitätsmanagement bei der SIZ (In Anlehnung an de Fries et al. (1999), S. 515)

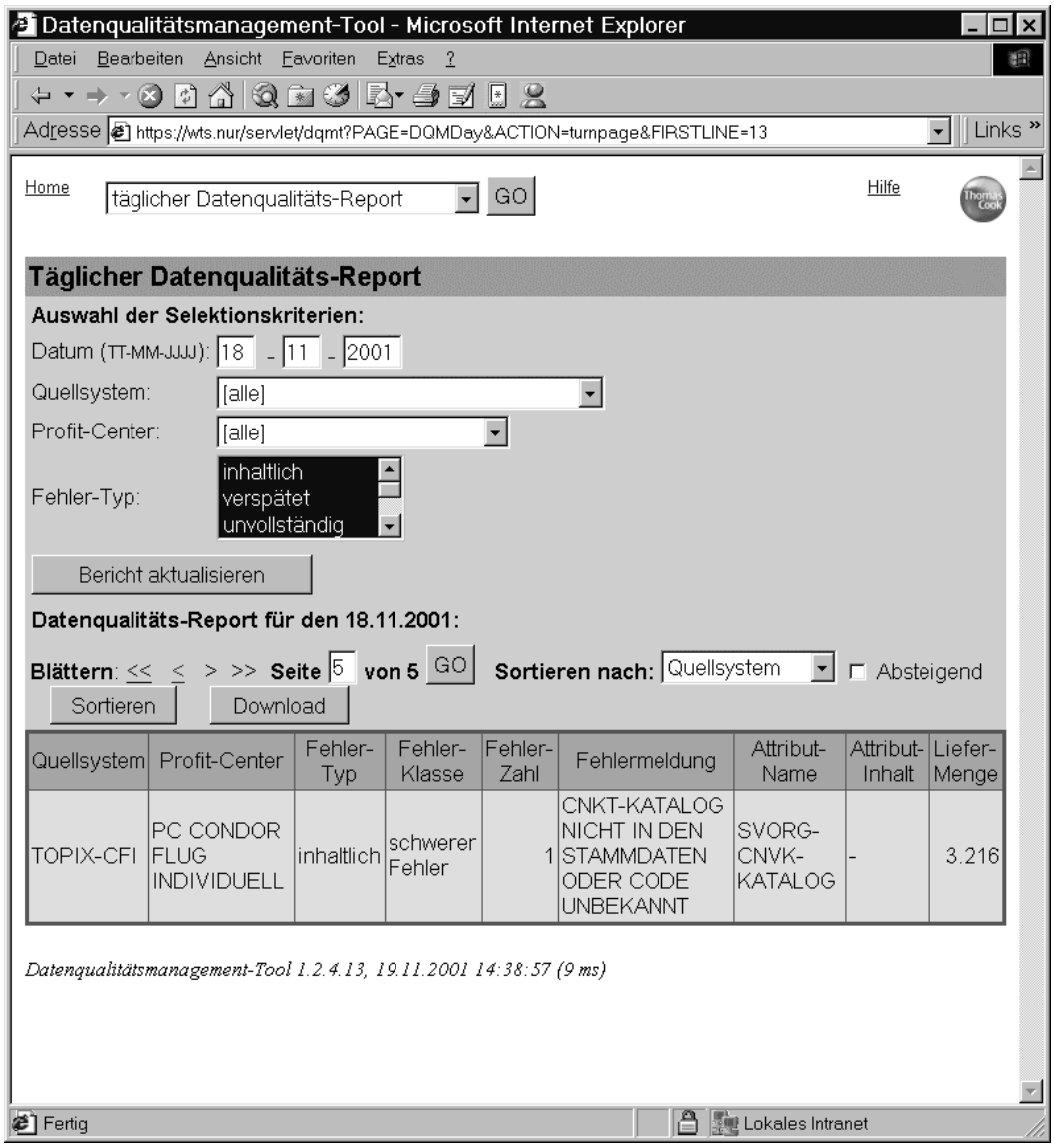


Abbildung 5.2: Werkzeug für das Datenqualitätsmanagement bei THOMAS COOK AG (Vgl. Thomas Cook AG, Konzerncontrolling (2001))

tributen. Auf Basis der erzeugten Regelmenge können dann mit Hilfe des Werkzeugs Datenbestände analysiert werden.

Diese Beispiele verdeutlichen exemplarisch die Umsetzung von Teilaspekten des hier erarbeiteten Konzeptes eines proaktiven Datenqualitätsmanagements in praktischen Anwendungsfällen. So können Datenbestände anhand von statistischen Methoden und Data Mining hinsichtlich der Datenqualität analysiert und beschrieben werden. Hiermit wird dann eine Qualitätsplanung und -lenkung möglich.

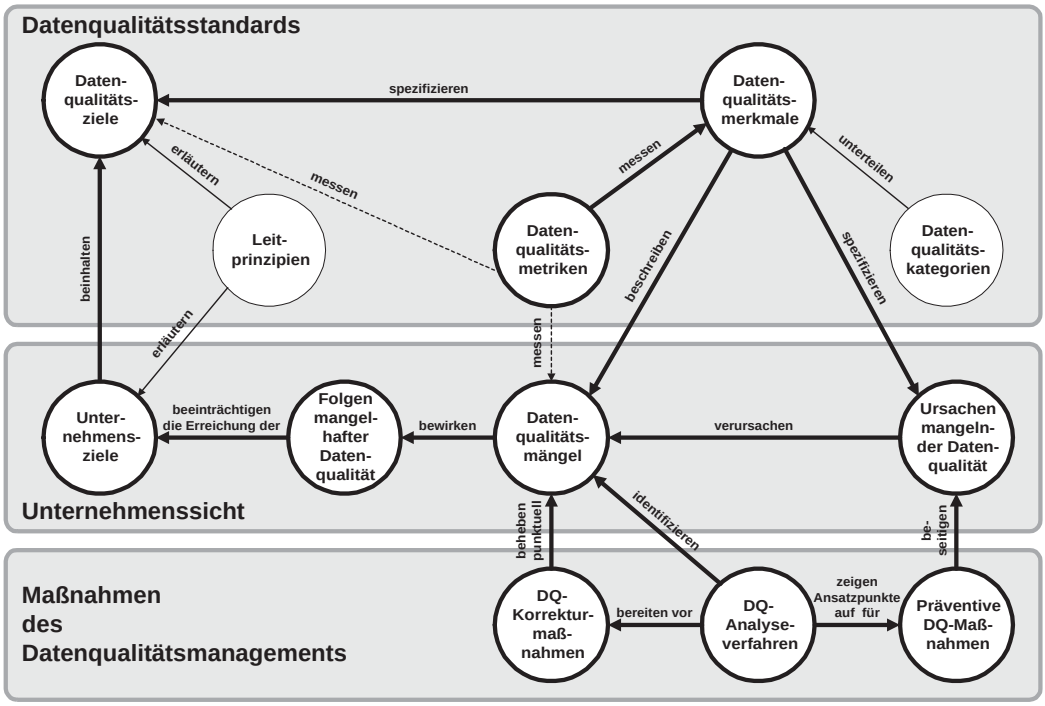


Abbildung 5.3: Datenqualitätsmanagement bei der SIZ (Vgl. de Fries et al. (1999), S. 514.)

Die ganzheitliche Betrachtung des Datenflusses sowie organisatorische und vorbeugende Massnahmen sind zu etablieren. Im Rahmen der zukünftigen Projektarbeit mit der Universalbank soll das vorgeschlagene Konzept weiter umgesetzt sowie zusätzliche anwendungsspezifische Regeln und Kennzahlen der Datenqualität ermittelt werden. Dabei ist die intensive Zusammenarbeit zwischen Fachexperten, Mitarbeitern auf der konzeptionellen Ebene und den Data-Warehouse-Entwicklern unabdingbar. In einem Folgeschritt soll die Betrachtung auf die operativen Vorsysteme erweitert und insbesondere organisatorische Strukturen zur Sicherstellung der Datenqualität etabliert werden. Die bislang vorwiegend auf die Data-Warehouse-Datenbasis spezifizierten Regeln sollen dann im Sinne einer ganzheitlichen Betrachtung festgelegt und auf die jeweiligen Komponenten verteilt werden.[457] Organisatorisch wird die Rolle von Datenqualitätsverantwortlichen angestrebt, welche im Sinne des Datenqualitätsmanagements Verantwortung

[457] Vgl. hierzu die Ausführungen zu unternehmensweiten Qualitätsregelkreisen in Abschnitt 3.4.3.2.

für die Datenqualität auf Fachbereichsebene übernehmen.

Wenngleich das Konzept des proaktiven Datenqualitätsmanagements für sinnvoll und notwendig erachtet wird, ist es dennoch nicht kritiklos. Insbesondere wird die ganzheitliche Umsetzung des Konzeptes als problematisch erachtet. Hier sind weniger technische Probleme als vielmehr organisatorische Hindernisse zu nennen. Technisch ist die Integration in die bisherige Architektur sowie die Beherrschung der Komplexität in bezug auf die Systemverwaltung zu nennen. Organisatorisch werden Widerstände auf Ebene der operativen Vorsysteme gegen Qualitätsverbesserungen im Hinblick auf Ziele der analytischen Systeme erwartet. Aussagen wie beispielsweise „Für uns ist die Qualität der Daten in Ordnung“ und „Wir können damit arbeiten“ sind häufig zu finden. Zur Verbesserung der Datenqualität sind neben konkreten Qualitätsvorgaben auch Anreizsysteme in den einzelnen Bereichen zu etablieren. Ohne die generelle Unterstützung der Unternehmensführung sind solche Änderungen allerdings kaum möglich. Die Verankerung einer umfassenden Datenqualitätskultur im Unternehmen ist eine der wichtigen Herausforderungen in naher Zukunft. Denn ohne diese Kultur wird die Bereitschaft zur Veränderung hinsichtlich der Datenqualitätsziele eher gering sein. Diese Erkenntnisse decken sich auch mit Erfahrungen im Kompetenzzentrum „Data Warehousing 2“. Im Rahmen eines Workshops wurde die Schaffung eines Datenqualitätsbewusstseins in Unternehmen als notwendig zur Sicherstellung der Datenqualität identifiziert.[458] Als häufige Hindernisse werden insbesondere die fehlende Managementunterstützung und Probleme bei der organisatorischen und technischen Realisierung genannt.

Ungeachtet dieser Herausforderungen ist die Umsetzung eines proaktiven Datenqualitätsmanagements notwendig. Zukünftig gilt es, nicht nur die Leistungsprozesse qualitativ auszuführen, sondern auch insbesondere die Qualität des Informationssystems und der Daten auf allen Ebenen zu gewährleisten. Die im Rahmen dieser Arbeit erzielten Ergebnisse bilden eine wichtige Basis im Bereich des Datenqualitätsmanagements und bieten konkrete Vorschläge zur Planung und Messung der Datenqualität in Data-Warehouse-Systemen sowie der Gestaltung eines metadatenbasierten Qualitätssystems.

[458] Vgl. Helfert et al. (2001), S. 14-17.

5.3 Weiterer Forschungsbedarf

Die Arbeit erörtert einige wesentliche Forschungsfragen im Bereich der Datenqualität. Jedoch ist das Forschungsgebiet sehr umfassend. Daher sollen im folgenden einige Ansatzpunkte für weitere Forschungsarbeiten aufgeführt werden. Weiterer Forschungsbedarf wird sowohl speziell im Bereich der Data-Warehouse-Systeme als auch im allgemeinen Datenqualitätsmanagement gesehen.

Im Rahmen der Arbeit wurde schwerpunktmässig die Ausführungsqualität mit den Qualitätskriterien Glaubwürdigkeit und zeitlicher Bezug betrachtet. Diese Betrachtungen sind auf weitere Qualitätskriterien und insbesondere die Designqualität auszuweiten. Zur Messung der Designqualität (Datenmodelle und Transferprozesse) sind Möglichkeiten zur Prüfung anhand von Modellierungsstandards und Entwurfsrichtlinien weiter zu konkretisieren. Im Bereich der Data-Warehouse-Systeme sind die hier erarbeiteten Möglichkeiten zur Spezifikation von Qualitätsvorgaben, der Datenanalyse und der Qualitätsprüfung zu erweitern. Insbesondere bieten die Verfahren der Statistik und des Data Mining weitere interessante und vielversprechende Potentiale zur Spezifikation und Messung der Ausführungsqualität. Die hier zur Qualitätsprüfung vorgeschlagenen Integritätsbedingungen können dann im Sinne einer Integritätssicherung zur Sicherstellung konsistenter Datenbestände im Data-Warehouse-System ganzheitlich erweitert werden. Technisch ist das Datenqualitätsmanagement weiter mit der Metadatenverwaltung zu integrieren und so ein weitgehend automatisches, ganzheitliches Datenqualitätsmanagement zu ermöglichen. Hier sind die Möglichkeiten im Common Warehouse Metamodel zu prüfen und dieses eventuell hinsichtlich der Berücksichtigung des Datenqualitätsmanagements zu ergänzen. In diesem Zusammenhang ist die Standardisierung der bislang häufig proprietären Metadatenmodelle mit Hilfe des CWM zu nennen. Durch diese kann eine durchgängige Modellierung der Metadaten auf allen Architekturebenen ermöglicht werden, auf der dann eine umfassende Betrachtung des Datenflusses unterstützt wird. So könnte der in Abschnitt 2.4.2.5 angesprochene Mechanismus zur Prüfung der Datenmodelle und des Data-Warehouse-Designs realisiert werden. Die Forschungsergebnisse sind dann in praktikable und kommerzielle Softwarelösungen umzusetzen.

Das hier für Data-Warehouse-Systeme vorgeschlagene Konzept ist in weiteren Fallbeispielen kritisch zu reflektieren und in einem iterativen Prozess weiterzuentwickeln. Hier bieten sich Anwendungen im Bereich der Banken aber auch anderen Branchen an. Insbesondere ist das Konzept hinsichtlich komplexerer Informationssysteme nicht nur für analytische Fragestellungen sondern auch zur Unterstützung operativer Geschäftsprozesse (z. B. Operational Data Store) zu erweitern. Aufgrund des ganzheitlichen Ansatzes ist die Übertragung des Konzepts auf andere Anwendungsbereiche nicht nur vorstellbar, sondern im Sinne eines umfassenden Datenqualitätsmanagements notwendig und weiter zu untersuchen. Potentiale zur Übertragung der hier gewonnenen Erkenntnis sind sicherlich in Anwendungen des Wissensmanagements und in unternehmensübergreifenden Informations- und Logistiksystemen möglich.

Ergänzend sind im Bereich des Datenqualitätsmanagements organisatorische und methodische Aspekte zu untersuchen. So sind bisherige Organisationskonzepte hinsichtlich der Berücksichtigung des Datenqualitätsmanagements zu prüfen und eventuell zu modifizieren. Neben den hier dargestellten Möglichkeiten zur Erfassung von Qualitätsforderungen und deren Konkretisierung in Qualitätsvorgaben sind standardisierte Methoden und Vorgehensweisen erforderlich. Dieses Forschungsgebiet ist, wie es sich im Verlauf der Arbeit gezeigt hat, bislang nicht ausreichend berücksichtigt. Beispielsweise ist die Informationsbedarfsanalyse hinsichtlich der Erfassung von Qualitätsforderungen zu erweitern. Zur Einführung und dynamischen Entwicklung des Datenqualitätsmanagements in einem Unternehmen wäre ein methodisches Vorgehen notwendig. Die Forschungsergebnisse könnten dann zusammenfassend, ähnlich der DIN ISO-Qualitätsnormen, zu einem einheitlichen Standard für das Datenqualitätsmanagement normiert werden.

Wie die Ausführungen in Kapitel 3 zeigen, ist die theoretische Basis und die Abgrenzung des Begriffes der Datenqualität bislang nicht ausreichend untersucht. Hier sind insbesondere Arbeiten zur Definition allgemeiner Datenqualitätskriterien und deren Beziehungsstruktur notwendig. Weiter ist die Beziehung zwischen Datenqualität, Datenmodellqualität, Softwarequalität und Anwenderzufriedenheit zu erforschen. Zusammenfassend besteht Forschungsbedarf im Bereich der Datenqualität sowohl zur Entwicklung einer theoretischen Basis, der Ausgestaltung

des Datenqualitätsmanagements durch Methoden und Organisationskonzepte sowie der Weiterentwicklung von Möglichkeiten zu Spezifikation und Prüfung der Datenqualität und deren technische Umsetzung.

Anhang A

Empirische Untersuchung

A.1 Fragebogen

Universität St. Gallen
Markus Helfert

markus.helfert@unisg.ch
Tel.: +41-(0)71 224 33 82
www.iwi.unisg.ch

Firma:	
Name:	
Telefonnr.:	
Email:	

Frage 1:

Welche Zwecke verfolgt das Data Warehouse-System in Ihrem Unternehmen?

(Mehrfachnennungen möglich)

- ❑ Langfristiges Planungs- und Entscheidungssystem (Strategieplanung)
- ❑ Analysesystem
- ❑ Berichts- und Kontrollsystem (Controlling)
- ❑ Unterstützung der wertorientierten Systeme (Buchhaltungssysteme, Zahlungssysteme)
- ❑ Unterstützung der mengenorientierten operativen Systeme (Operational Data Store; direkte Unterstützung der Geschäftsprozesse)

Frage 2:

Welche betriebswirtschaftlichen Funktionen werden von dem Data Warehouse-System vorwiegend unterstützt?

(Mehrfachnennungen möglich)

- ❑ Beschaffung / Einkauf
- ❑ Produktion
- ❑ Logistik
- ❑ Vertrieb / Marketing
- ❑ Produktionstechnik / -anlagen
- ❑ Personalplanung

Frage 3:

Was beschreibt den Fokus Ihrer Aufgaben im Bereich des Data Warehousing am Besten?

(Mehrfachnennungen möglich)

	Fachlicher Schwerpunkt		
	Spezifisches Fachbereichswissen	Konzeptionell / Design	Technisch / Infrastruktur
Endbenutzer / Datenanwender			
Data Marts / multidimensionale Modelle / OLAP			
Zentrale Data Warehouse-Datenbasis			
Transformationskomponente / ETL-Prozess			
Operative Vorsysteme / OLTP			

Universität St. Gallen
Markus Helfert

markus.helfert@unisg.ch
Tel.: +41-(0)71 224 33 82
www.iwi.unisg.ch

Frage 4:
Stellt die Sicherstellung der Datenqualität in Ihrem Unternehmen / Data Warehouse-System ein Problem dar?

- ○ Sehr grosses Problem
- ○ Grosses Problem
- ○ Kein besonderes Problem
- ○ Kein Problem

Frage 5:
Wie wird bei Ihnen die Sicherstellung der Datenqualität erreicht?
(Mehrfachnennungen möglich)

- ❑ Keine besondere Massnahmen zur Sicherstellung der Datenqualität
- ❑ Datenbereinigung (Werkzeugunterstützung)
- ❑ Datenqualitätsmanagement mit: (wenn ja: Wie?)
  - ❑ Organisatorischen Massnahmen (z. B. DQ-Verantwortlichen)
  - ❑ Dokumentierter DQ-Prozess
  - ❑ Festlegung von Datenqualitätszielen
  - ❑ Datenqualitätsmessung (Kennzahlen über die aktuelle Datenqualität)
  - ❑ Systematische Ursachenanalyse

Sonstiges: _____

Universität St. Gallen
Markus Helfert

markus.helfert@unisg.ch
Tel.: +41-(0)71 224 33 82
www.iwi.unisg.ch

Frage 6:
Wie werden in Ihrem Unternehmen Datenqualitäts-Probleme entdeckt?
(Mehrfachnennungen möglich)

- ❑ Durch fest etablierte und periodisch stattfindende Qualitätsüberprüfungen (organisatorisch geregelt)
- ❑ Bei der Datenmodellierung (Qualitätssicherung bei der Systementwicklung)
- ❑ Bei der Datenanlieferung bzw. Extraktion und der Transformation der Daten sowie dem anschliessenden Laden in die Data Warehouse-Datenbasis (Qualitätssicherung bei der Datenlieferung, ETL-Prozess)
- ❑ Durch Analyse des Data Warehouse-Datenbestandes (z. B. Statistik, Data Mining, Integritätsbedingungen, ...)
- ❑ Bei der Datenbereitstellung an die Endanwender (Datenanwendung)
- ❑ Keine besondere Prüfung

Sonstiges: _____

Frage 6.1: Was wird geprüft und wie wird es geprüft?

_____

_____

_____

Frage 6.2: Welches sind die wesentlichsten Probleme, die hierbei gefunden/genannt werden?

_____

_____

_____

Frage 6.3: Wie wird mit diesen Problemen umgegangen? Wie werden diese Probleme beseitigt?

_____

_____

_____

Frage 7:
Mit welchen Eigenschaften würden Sie qualitativ hochwertige Daten für Data Warehouse-Systeme beschreiben?
(Liste von Eigenschaften)

_____

_____

_____

_____

Frage 8:
Welche Aspekte sind zur Beschreibung des Begriff der "Datenqualität" in Data Warehouse-Systemen wichtig?
(Mehrfachnennungen möglich)

- ❑ Qualität der Datendefinition (Datenmodelle, Datenprozesse)
- ❑ Qualität der Dateninhalte (Datenwerte)
- ❑ Qualität der Datenbereitstellung (Gesamtes Data Warehouse-System)

Universität St. Gallen
Markus Helfert

markus.helfert@unisg.ch
Tel.: +41-(0)71 224 33 82
www.iwi.unisg.ch

Frage 9:

Wie wichtig sind die folgenen Eigenschaften zur Bestimmung der Datenqualität in Data Warehouse-Systemen?

		entscheidend	sehr wichtig	wichtig	kann vernachlässigt werden	unwichtig	gehört nicht zu Datenqualität
Semantik	Die Entitäten, Beziehungen und Attribute und deren Wertebereiche sind einheitlich, klar und genau beschreiben sowie dokumentiert						
Identifizierbarkeit	Einzelne Informationsobjekte können eindeutig identifiziert werden						
Synonyme	Beziehungen zwischen den Synonymen sind bekannt und dokumentiert						
Zeitlicher Bezug	Der zeitliche Bezug einzelner Informationsobjekte ist abgebildet						
Repräsentation fehlender Werte	Fehlende Werte (Null-Werte / Default-Werte) sind definiert und können abgebildet werden						
Vollständigkeit	Alle wesentlichen Entitäten, Beziehungen und Attribute sind erfasst						
Erforderlichkeit	Definition von Pflicht- und Kann-Felder						
Ganularität	Die Entitäten, Beziehungen und Attribute sind im notwendigen Detaillierungsgrad erfasst						
Präzision der Wertebereichsdefinitionen	Die definierten Wertebereiche repräsentieren die möglichen und sinnvollen Datenwerte						
Semantische Korrektheit	Die Daten stimmen inhaltlich mit der Datendefinition überein und sind empirisch korrekt						
Datenherkunft	Die Datenherkunft und die vorgenommenen Datentransformationen sind bekannt						
Vollständigkeit	Alle Daten sind gemäss Datenmodell erfasst						
Widerspruchsfreiheit (Daten)	Die Daten weisen innerhalb des Datenbestands und zu anderen Datenbeständen keine Widersprüche auf						
Widerspruchsfreiheit (Regeln)	Die Daten weisen keine Widersprüche zu allgemeingültigen Geschäftsregeln, Integritätsbedingungen und Wertebereichsdefinitionen auf						
Syntaktische Korrektheit	Die Daten stimmen mit der spezifizierten Syntax (Format) überein						
Zuverlässigkeit	Die Glaubwürdigkeit der Daten ist konstant (z. B. gleichbleibendes Datenvolumen)						
Aktualität	Die Datenwerte beziehen sich auf den aktuellen Zeitpunkt						
Zeitliche Konsistenz	Alle Datenwerte bzgl. eines Zeitpunktes sind gleichermassen aktuell						
Zeitliche Verfügbarkeit	Die Daten stehen rechtzeitig zur Verfügung						
Systemverfügbarkeit	Das Gesamtsystem ist verfügbar						
Transaktionsverfügbarkeit	Einzelne Transaktionen sind ausführbar sowie die Zugriffszeit ist akzeptabel und gleichbleibend						
Zugriffsrechte	Die benötigten Zugriffsreche sind ausreichend						
Zeitliche Bezug	Die Datenwerte beziehen sich auf den benötigten Zeitraum						
Relevanz	Die Datenwerte können auf einen relevanten Datenausschnitt beschränkt werden						
Nicht-Volatilität	Die Datenwerte sind permanent und können zu einem späteren Zeitpunkt wieder aufgerufen werden						

Gibt es noch weitere Eigenschaften, die besonders die Datenqualität eines Data Warehouse-Systems beschreiben?

A.2 Detailergebnisse

Verteilung der Anschreiben und Antworten nach Branche und Land:

		Absolut		Relativ	
		An-schreiben	**Antworten**	**An-schreiben**	**Antworten**
Branche	Baugewerbe	1	0	0,91%	0%
	Chemische Industrie	8	2	7,27%	8%
	Dienstleistung	10	0	9,09%	0%
	Energie- und Wasservers.	2	1	1,82%	5%
	Gesundheitswesen	2	0	1,82%	4%
	Handel	1	0	0,91%	0%
	Kredit- und Versicherungsgew.	44	16	40,00%	64%
	Nachrichtenübermitt-lung	5	0	4,55%	0%
	Öffentliche Verwaltung	7	2	6,36%	8%
	Verarb. Gewerbe (nicht ch. Ind.)	20	2	18,18%	8%
	Verkehr	10	2	9,09%	8%
Land	Schweiz	76	17	69%	68%
	Deutschland	29	6	26%	24%
	Österreich	5	2	4%	8%

Relevanz der Datenqualitätsmerkmale:

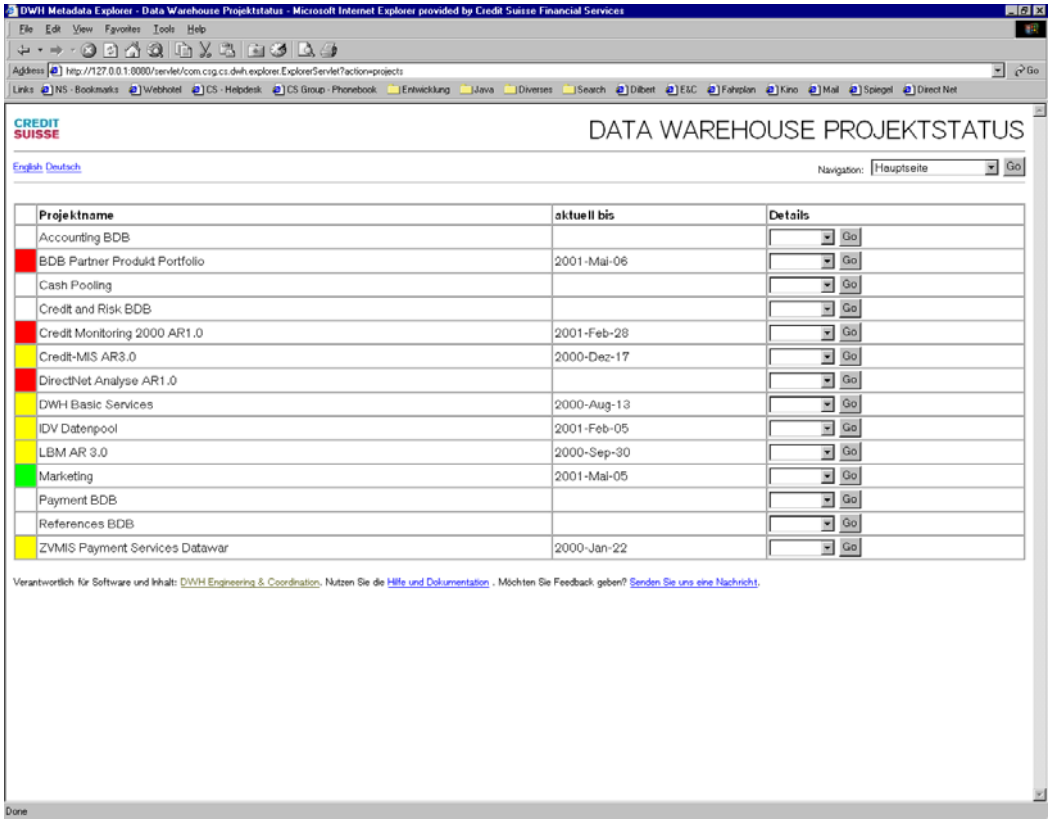
	Anzahl der Nennungen					
	e[a]	s[b]	w[c]	v[d]	u[e]	n[f]
Semantik	48	36	12	0	0	0
Identifizierbarkeit	44	40	8	0	0	4
Synonyme	16	16	44	8	0	0
Zeitlicher Bezug	20	24	36	4	0	8
Repräsentation fehlender Werte	8	28	52	4	0	0
Vollständigkeit (Datenmodell)	16	40	32	4	0	8
Erforderlichkeit	8	12	52	12	0	8
Granularität	20	32	36	0	0	12
Präzision der Wertebereichsdefinitionen	12	28	40	8	4	4
Semantische Korrektheit	52	20	24	0	0	0
Datenherkunft	44	24	20	8	0	0
Vollständigkeit (Datenwerte)	16	32	32	8	4	4
Widerspruchsfreiheit (Daten)	52	28	16	0	0	0
Widerspruchsfreiheit (Regeln)	36	40	16	4	0	0
Syntaktische Korrektheit	20	36	32	0	0	0
Zuverlässigkeit	24	28	32	0	4	8
Aktualität	12	16	40	8	4	8
Zeitliche Konsistenz	20	40	24	8	0	4
Zeitliche Verfügbarkeit	12	24	44	0	0	16
Systemverfügbarkeit	12	32	28	0	0	24
Transaktionsverfügbarkeit	0	16	44	4	4	28
Zugriffsrechte	4	16	24	8	4	40
Zeitlicher Bezug	20	24	36	4	0	8
Relevanz	4	20	24	8	8	12
Nicht-Volatilität	12	40	16	8	8	12

[a] Entscheidend
[b] Sehr Wichtig
[c] Wichtig
[d] Kann vernachlässigt werden
[e] Unwichtig
[f] gehört nicht zur Datenqualität

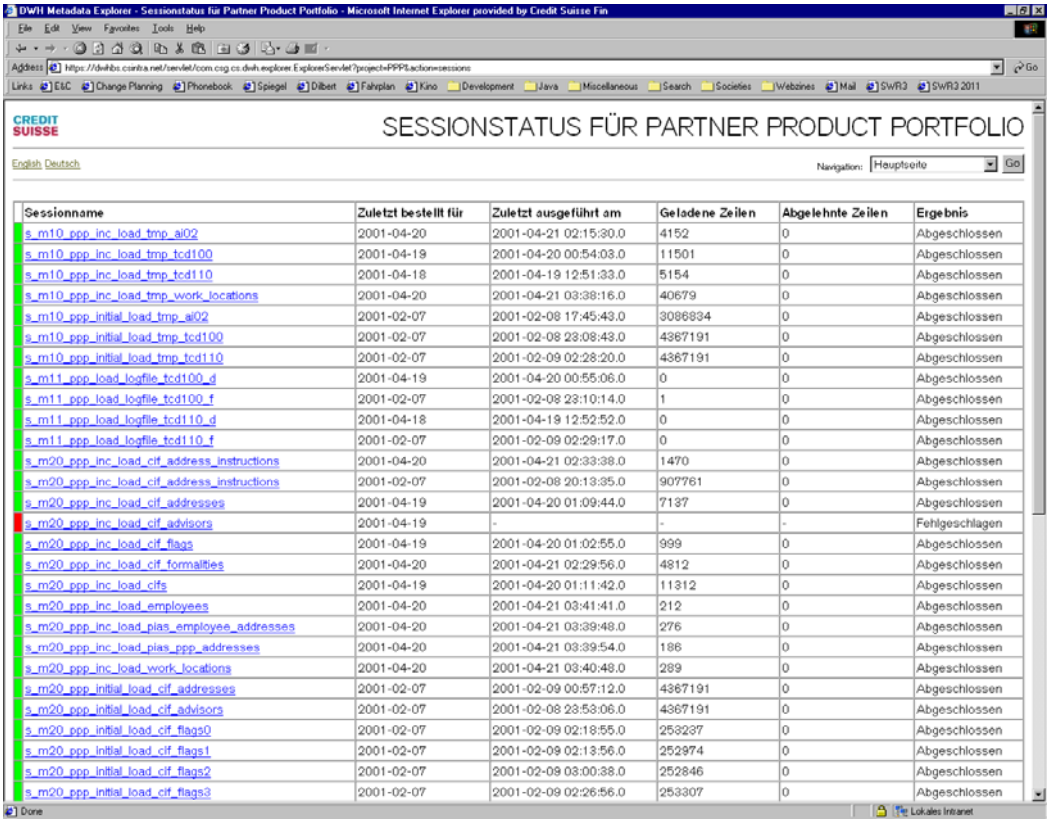
Anhang B

Fallstudie

Die folgenden Abbildungen zeigen Beispiele von bislang verfügbaren technischen Qualitätsangaben.[459]



[459] Vgl. hierzu Wegener (2001).



DWH Metadata Explorer - Sessionstatus für Partner Product Portfolio - Microsoft Internet Explorer provided by Credit Suisse Fin

CREDIT SUISSE

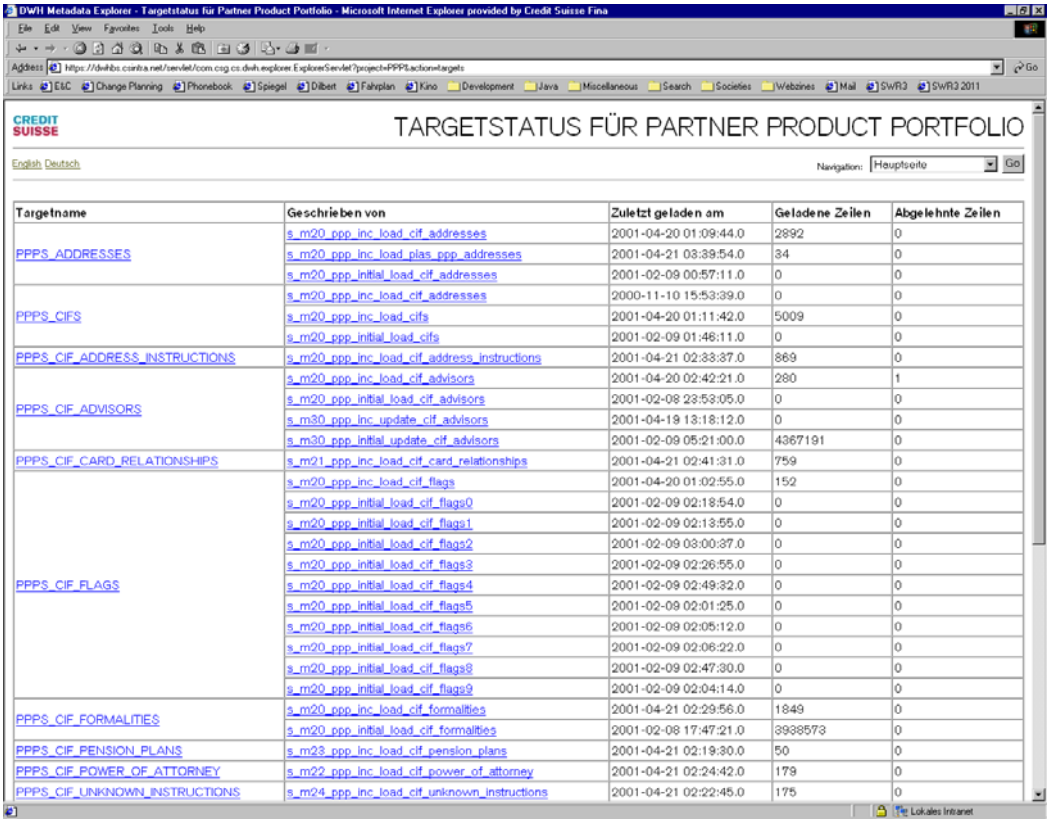
SESSIONSTATUS FÜR PARTNER PRODUCT PORTFOLIO

English Deutsch

Navigation: Hauptseite Go

Sessionname	Zuletzt bestellt für	Zuletzt ausgeführt am	Geladene Zeilen	Abgelehnte Zeilen	Ergebnis
s_m10_ppp_inc_load_tmp_ai02	2001-04-20	2001-04-21 02:15:30.0	4152	0	Abgeschlossen
s_m10_ppp_inc_load_tmp_tcd100	2001-04-19	2001-04-20 00:54:03.0	11501	0	Abgeschlossen
s_m10_ppp_inc_load_tmp_tcd110	2001-04-18	2001-04-19 12:51:33.0	5154	0	Abgeschlossen
s_m10_ppp_inc_load_tmp_work_locations	2001-04-20	2001-04-21 03:38:16.0	40679	0	Abgeschlossen
s_m10_ppp_initial_load_tmp_ai02	2001-02-07	2001-02-08 17:45:43.0	3086834	0	Abgeschlossen
s_m10_ppp_initial_load_tmp_tcd100	2001-02-07	2001-02-08 23:08:43.0	4367191	0	Abgeschlossen
s_m10_ppp_initial_load_tmp_tcd110	2001-02-07	2001-02-09 02:28:20.0	4367191	0	Abgeschlossen
s_m11_ppp_load_logfile_tcd100_d	2001-04-19	2001-04-20 00:55:06.0	0	0	Abgeschlossen
s_m11_ppp_load_logfile_tcd100_f	2001-02-07	2001-02-08 23:10:14.0	1	0	Abgeschlossen
s_m11_ppp_load_logfile_tcd110_d	2001-04-18	2001-04-19 12:52:52.0	0	0	Abgeschlossen
s_m11_ppp_load_logfile_tcd110_f	2001-02-07	2001-02-09 02:29:17.0	0	0	Abgeschlossen
s_m20_ppp_inc_load_cif_address_instructions	2001-04-20	2001-04-21 02:33:38.0	1470	0	Abgeschlossen
s_m20_ppp_inc_load_cif_address_instructions	2001-02-07	2001-02-08 20:13:35.0	907761	0	Abgeschlossen
s_m20_ppp_inc_load_cif_addresses	2001-04-19	2001-04-20 01:09:44.0	7137	0	Abgeschlossen
s_m20_ppp_inc_load_cif_advisors	2001-04-19	-	-	-	Fehlgeschlagen
s_m20_ppp_inc_load_cif_flags	2001-04-19	2001-04-20 01:02:55.0	999	0	Abgeschlossen
s_m20_ppp_inc_load_cif_formalities	2001-04-20	2001-04-21 02:29:56.0	4812	0	Abgeschlossen
s_m20_ppp_inc_load_cifs	2001-04-19	2001-04-20 01:11:42.0	11312	0	Abgeschlossen
s_m20_ppp_inc_load_employees	2001-04-20	2001-04-21 03:41:41.0	212	0	Abgeschlossen
s_m20_ppp_inc_load_pias_employee_addresses	2001-04-20	2001-04-21 03:39:48.0	276	0	Abgeschlossen
s_m20_ppp_inc_load_pias_ppp_addresses	2001-04-20	2001-04-21 03:39:54.0	186	0	Abgeschlossen
s_m20_ppp_inc_load_work_locations	2001-04-20	2001-04-21 03:40:48.0	289	0	Abgeschlossen
s_m20_ppp_initial_load_cif_addresses	2001-02-07	2001-02-09 00:57:12.0	4367191	0	Abgeschlossen
s_m20_ppp_initial_load_cif_advisors	2001-02-07	2001-02-08 23:53:06.0	4367191	0	Abgeschlossen
s_m20_ppp_initial_load_cif_flags0	2001-02-07	2001-02-09 02:18:55.0	253237	0	Abgeschlossen
s_m20_ppp_initial_load_cif_flags1	2001-02-07	2001-02-09 02:13:56.0	252974	0	Abgeschlossen
s_m20_ppp_initial_load_cif_flags2	2001-02-07	2001-02-09 03:00:38.0	252846	0	Abgeschlossen
s_m20_ppp_initial_load_cif_flags3	2001-02-07	2001-02-09 02:26:56.0	253307	0	Abgeschlossen

Done Lokales Intranet



DWH Metadata Explorer - Targetstatus für Partner Product Portfolio - Microsoft Internet Explorer provided by Credit Suisse Fina

CREDIT SUISSE

TARGETSTATUS FÜR PARTNER PRODUCT PORTFOLIO

English Deutsch

Navigation: Hauptseite Go

Targetname	Geschrieben von	Zuletzt geladen am	Geladene Zeilen	Abgelehnte Zeilen
PPPS_ADDRESSES	s_m20_ppp_inc_load_cif_addresses	2001-04-20 01:09:44.0	2892	0
	s_m20_ppp_inc_load_pias_ppp_addresses	2001-04-21 03:39:54.0	34	0
	s_m20_ppp_initial_load_cif_addresses	2001-02-09 00:57:11.0	0	0
PPPS_CIFS	s_m20_ppp_inc_load_cif_addresses	2000-11-10 15:53:39.0	0	0
	s_m20_ppp_inc_load_cifs	2001-04-20 01:11:42.0	5009	0
	s_m20_ppp_initial_load_cifs	2001-02-09 01:46:11.0	0	0
PPPS_CIF_ADDRESS_INSTRUCTIONS	s_m20_ppp_inc_load_cif_address_instructions	2001-04-21 02:33:37.0	869	0
PPPS_CIF_ADVISORS	s_m20_ppp_inc_load_cif_advisors	2001-04-20 02:42:21.0	280	1
	s_m20_ppp_initial_load_cif_advisors	2001-02-08 23:53:05.0	0	0
	s_m30_ppp_inc_update_cif_advisors	2001-04-19 13:18:12.0	0	0
	s_m30_ppp_initial_update_cif_advisors	2001-02-09 05:21:00.0	4367191	0
PPPS_CIF_CARD_RELATIONSHIPS	s_m21_ppp_inc_load_cif_card_relationships	2001-04-21 02:41:31.0	759	0
PPPS_CIF_FLAGS	s_m20_ppp_inc_load_cif_flags	2001-04-20 01:02:55.0	152	0
	s_m20_ppp_initial_load_cif_flags0	2001-02-09 02:18:54.0	0	0
	s_m20_ppp_initial_load_cif_flags1	2001-02-09 02:13:55.0	0	0
	s_m20_ppp_initial_load_cif_flags2	2001-02-09 03:00:37.0	0	0
	s_m20_ppp_initial_load_cif_flags3	2001-02-09 02:26:55.0	0	0
	s_m20_ppp_initial_load_cif_flags4	2001-02-09 02:49:32.0	0	0
	s_m20_ppp_initial_load_cif_flags5	2001-02-09 02:01:25.0	0	0
	s_m20_ppp_initial_load_cif_flags6	2001-02-09 02:05:12.0	0	0
	s_m20_ppp_initial_load_cif_flags7	2001-02-09 02:06:22.0	0	0
	s_m20_ppp_initial_load_cif_flags8	2001-02-09 02:47:30.0	0	0
	s_m20_ppp_initial_load_cif_flags9	2001-02-09 02:04:14.0	0	0
PPPS_CIF_FORMALITIES	s_m20_ppp_inc_load_cif_formalities	2001-04-21 02:29:56.0	1849	0
	s_m20_ppp_initial_load_cif_formalities	2001-02-08 17:47:21.0	3938573	0
PPPS_CIF_PENSION_PLANS	s_m23_ppp_inc_load_cif_pension_plans	2001-04-21 02:19:30.0	50	0
PPPS_CIF_POWER_OF_ATTORNEY	s_m22_ppp_inc_load_cif_power_of_attorney	2001-04-21 02:24:42.0	179	0
PPPS_CIF_UNKNOWN_INSTRUCTIONS	s_m24_ppp_inc_load_cif_unknown_instructions	2001-04-21 02:22:45.0	175	0

Lokales Intranet

Literaturverzeichnis

ALBRECHT, JENS ET AL. (2001): *Entwicklung*, in: Bauer, Andreas und Günzel, Holger (Hrsg.): *Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung*, S. 148–346, dpunkt-Verlag, Heidelberg 2001.

ANDERSON, OSKAR; POPP, WERNER; SCHAFFRANEK, MANFRED; STEINMETZ, DIETER und STENGER, HORST (1997): *Schätzen und Testen: eine Einführung in die Wahrscheinlichkeitsrechnung und schliessende Statistik*, 2. Aufl., Springer, Berlin et al. 1997.

ANSI/X3/SPARC STUDY GROUP ON DATA BASE MANAGEMENT SYSTEMS (1975): *Interim Report 75-02-08*, in: FDT (Bulletin of ACM-SIGMOD), 7. Jg., Nr. 2 (1975), S. 1–140.

ATTESLANDER, PETER (1995): *Methoden der empirischen Sozialforschung*, 8. Aufl., de Gruyter, Berlin et al. 1995.

AUGUSTIN, SIEGFRIED (1990): *Information als Wettbewerbsfaktor, Informationslogistik - Herausforderungen an das Management*, Verlag Industrielle Organisation und Verlag TÜV Rheinland, Zürich und Köln 1990.

BAETGE, JÖRG (1974): *Betreibswirtschaftliche Systemtheorie: regelungstheoretische Planungs-Überwachungsmodelle für Produktion, Lagerung und Absatz*, Westdt. Verlag, Opladen 1974.

BALLOU, DONALD P. und PAZER, HAROLD L. (1985): *Modeling Data and Process Quality in Multi-input, Multi-output Information Systems*, in: Management Science, 31. Jg., Nr. 2 (1985), S. 150–162.

BANGE, CARSTEN ET AL. (2001): *Architektur*, in: Bauer, Andreas und Günzel, Holger (Hrsg.): *Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung*, S. 1–147, dpunkt-Verlag, Heidelberg 2001.

BARKOW, G.; HESSE, W.; KITTLAUS, H.-B.; SCHESCHONK, G. und VON BRAUN, H. (1997): *Anwendungsmodell*, in: Schneider, Hans-Jochen (Hrsg.): *Lexikon Informatik und Datenverarbeitung*, 4. Aufl., S. 45, Oldenbourg, München und Wien 1997.

BARTRAM, JENS (1992): *Qualitäts-Informationssysteme für die Textilindustrie: Gestaltung am Beispiel gewebeherstellender Textilbetriebe mit automatisierten Produktionsabläufen*, Dissertation, Hochschule St. Gallen, St. Gallen, 1992.

BAUER, ANDREAS ET AL. (2001): *Anwendung*, in: Bauer, Andreas und Günzel, Holger (Hrsg.): *Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung*, S. 347–509, dpunkt-Verlag, Heidelberg 2001.

BECKER, JÖRG und VOSSEN, GOTTFRIED (1996): *Geschäftsprozessmodellierung und Workflow-Management: Eine Einführung*, in: Becker, Jörg und Vossen, Gottfried (Hrsg.): *Geschäftsprozessmodellierung und Workflow-Management: Modelle, Methoden, Werkzeuge*, S. 12–21, Internat. Thomson Publ., Bonn 1996.

BEIER, DIRK; GABRIEL, ROLAND und STREUBEL, FRAUKE (1997): *Ziele und Aufgaben des Informationsmanagements*, Arbeitsbericht 97-23, Ruhr-Universität Bochum - Lehrstuhl für Wirtschaftsinformatik, Bochum 1997.

BERRY, MICHAEL J. A. und LINOFF, GORDON (1997): *Data Mining Techniques: For Marketing, Sales, and Customer Support*, Wiley, New York et al. 1997.

BERTHEL, JÜRGEN (1992): *Informationsbedarf*, in: Frese, Erich (Hrsg.): *Handwörterbuch der Organisation (HWO)*, 3. Aufl., S. 872–886, Poeschel, Stuttgart 1992.

BIERI, BRUNO (1995): *Kybernetisches Produktions-Controlling mit Hilfe von Kennzahlen*, Dissertation Hochschule St. Gallen, Difo-Druck, Bamberg 1995.

BOBROWSKI, MONICA; MARRE, MARTINA und YANKELEVICH, DANIEL (1999): *A homogeneous framework to measure data quality*, in: Lee, Yang W. und Tayi, Giri Kumar (Hrsg.): *Proceedings of the 1999 Conference on Information Quality*, S. 115–124, Massachusetts Institute of Technology, Cambridge, MA 1999.

BODE, JÜRGEN (1997): *Der Informationsbegriff in der Betriebswirtschaftslehre*, in: Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, 49. Jg., Nr. 5 (1997), S. 449–468.

BOHLEY, PETER (1996): *Statistik: Einführendes Lehrbuch für Wirtschafts- und Sozialwissenschaftler*, 6. Aufl., Oldenbourg, München und Wien 1996.

BOTTA, VOLKMAR (1997): *Kennzahlensysteme als Führungsinstrumente: Planung, Steuerung und Kontrolle der Rentabilität im Unternehmen*, 5. Aufl., Schmidt, Berlin 1997.

BRACKETT, MICHAEL H. (1996): *The Data Warehouse Challenge - Taming Data Chaos*, Wiley, New York et al. 1996.

BUSATTO, RENATO (2000): *Using Time Series to Assess Data Quality in Telecommunications Data Warehouses*, in: Klein, Barbara D. und Rossin, Donald F. (Hrsg.): *Proceedings of the 2000 Conference on Information Quality*, S. 129–136, Massachusetts Institute of Technology, Cambridge, MA 2000.

BUSCH, ULRICH (1983): *Konzeption betrieblicher Informations- und Kommunikationssysteme (IKS)*, Erich Schmidt, Berlin 1983.

CADUFF, DIRK (1997): *Vorgehensweise für entwicklungsorientierte Assessments am Beispiel von Modellen des Total Quality Managements*, Dissertation Universtität St. Gallen, Difo-Druck, Bamberg 1997.

CHAMONI, PETER und GLUCHOWSKI, PETER (1998): *Analytische Informationssysteme - Einordnung und Überblick*, in: Chamoni, Peter und Gluchowski, Peter (Hrsg.): *Analytische Informationssysteme: Data Warehouse, On-line Analytical Processing, Data Mining*, S. 3–25, Springer, Berlin et al. 1998.

CHEN, PETER PIN-SHAN (1976): *The Entity-Relationship Model - Toward a Unified View of Data*, in: ACM Transactions on Database Systems, 1. Jg., Nr. 1 (1976), S. 9–36.

CLAUSEN, NILS (1998): *OLAP-Multidimensionale Datenbanken: Produkte, Markt, Funktionsweise und Implementierung*, Addison-Wesley, Reading et al. 1998.

CODD, EDGAR F. (1970): *A Relational Model of Data for Large Shared Data Banks*, in: Communications of the ACM, 13. Jg., Nr. 6 (1970), S. 377–387.

CODD, EDGAR F. (1972a): *Further Normalization of the Data Base Relational Model*, in: Rustin, Randall (Hrsg.): *Courant Computer Science Symposium 6, May 24-25, 1971 - Data Base Systems*, S. 33–64, Prentice Hall International, Englewood Cliffs 1972.

CODD, EDGAR F. (1972b): *Relational Completeness of Data Base Sublanguages*, in: Rustin, Randall (Hrsg.): *Courant Computer Science Symposium*

6, May 24-25, 1971 - Data Base Systems, S. 65–98, Prentice Hall International, Englewood Cliffs 1972.

CONRAD, WERNER (2000): *Qualitätsmanagement in Data Warehouse-Projekten - Methoden und Verfahren für die Projektpraxis*, in: Mucksch, Harry und Behme, Wolfgang (Hrsg.): *Das Data-Warehouse-Konzept - Datenmodelle - Anwendungen: mit Erfahrungsberichten*, 4. Aufl., S. 291–329, Gabler, Wiesbaden 2000.

DASU, TAMRAPARNI und JOHNSON, THEODORE (1999): *Hunting of the Snark - Findinig Data Glitches using Data Mining Methods*, in: Lee, Yang W. und Tayi, Giri Kumar (Hrsg.): *Proceedings of the 1999 Conference on Information Quality*, S. 89–98, Massachusetts Institute of Technology, Cambridge, MA 1999.

DASU, TAMRAPARNI; JOHNSON, THEODORE und KOUTSOFIOS, ELEFTHERIOS (2000): *Hunting Data Glitches in Massive Time Series Data*, in: Klein, Barbara D. und Rossin, Donald F. (Hrsg.): *Proceedings of the 2000 Conference on Information Quality*, S. 190–199, Massachusetts Institute of Technology, Cambridge, MA 2000.

DATE, CHRISTOPHER J. (1992): *Entity/Relationship Modeling and the Relational Model*, in: Date, Christopher J. und Darwen, Hugh (Hrsg.): *Relational database writings, 1989-1991*, Kap. 22, S. 357–364, Addison Wesley, Reading et al. 1992.

DATE, CHRISTOPHER J. (2000): *An introduction to database systems*, 7. Aufl., Addision-Wesley, Reading et al. 2000.

DE FRIES, DIETRICH; SEIDL, JÖRG und WINDHEUSER, ULRICH (1999): *Datenqualität - ein unterschätzter Erfolgsfaktor*, in: Betriebswirtschaftliche Blätter, 48. Jg., Nr. 11 (1999), S. 513–517.

DEDEKE, ADENEKAN (2000): *A Conceptual Framework for Developing Quality Measures for Information Systems*, in: Klein, Barbara D. und Rossin, Donald F. (Hrsg.): *Proceedings of 2000 Conference on Information Quality*, S. 126–128, Massachusetts Institute of Technology, Cambridge, MA 2000.

DELONE, WILLIAM H. und MCLEAN, EPHRAIM R. (1992): *Information System Success: The Quest for the Dependent Variable*, in: Information System Research, 3. Jg., Nr. 1 (1992), S. 60–95.

DEVLIN, BARRY (1997): *Data Warehouse: from architecture to implementation*, Addison Wesley, Reading et al. 1997.

DIN, DEUTSCHES INSTITUT FÜR NORMUNG E. V. (1995): *Qualitätsmanagement, Statistik, Zertifizierung: Begriffe aus DIN-Normen*, 2. Aufl., Beuth, Berlin, Wien und Zürich 1995.

EDELSTEIN, HERBERT A. (1997): *An Introduction to Data Warehousing*, in: Barquin, Ramon C. und Edelstein, Herbert A. (Hrsg.): *Planing and Designing the Data Warehouse*, S. 31–50, Prentice-Hall, Upper Saddle River, NJ 1997.

EICKER, STEFAN (2001): *Ein Überblick über die Umsetzung des Data-Warehouse-Konzeptes aus technischer Sicht*, in: Schütte, Reinhard; Rotthowe, Thomas und Holten, Roland (Hrsg.): *Data Warehouse Managementhandbuch: Konzepte, Software, Erfahrungen*, S. 65–79, Springer, Berlin et al. 2001.

ELMASRI, RAMEZ und NAVATHE, SHAMKANT B. (1994): *Fundamentals of Database Systems*, 2. Aufl., Addison-Wesley, Reading et al. 1994.

ENGLISH, LARRY P. (1999): *Improving Data Warehouse and Business Information Quality*, Wiley, New York et al. 1999.

EPPLER, MARTIN J. und WITTIG, DÖRTE (2000): *Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years*, in: Klein, Barbara D. und Rossin, Donald F. (Hrsg.): *Proceedings of the 2000 Conference on Information Quality*, S. 83–96, Massachusetts Institute of Technology, Cambridge, MA 2000.

ESTER, MARTIN und SANDER, JÖRG (2000): *Knowledge discovery in databases: Techniken und Anwendung*, Springer, Berlin et al. 2000.

FAHRMEIER, LUDWIG; KÜNSTLER, RITA; PIGEOT, IRIS und TUTZ, GERHARD (1997): *Statistik - Der Weg zur Datenanalyse*, Springer, Berlin et al. 1997.

FAHRMEIR, LUDWIG; HAMERLE, ALFRED und TUTZ, GERHARD (1996): *Multivariate statistische Verfahren*, 2. Aufl., de Gruyter, Berlin et al. 1996.

FEIGENBAUM, ARMAND V. (1961): *Total Quality Control*, 2. Aufl., McGraw-Hill, New York et al. 1961.

FERSTL, OTTO K. und SINZ, ELMAR J. (2001): *Grundlagen der Wirtschaftsinformatik Band 1*, Oldenbourg, München und Wien 2001.

FLADE-RUF, URSULA (1996): *Data Warehouse - nicht nur etwas für Grossunternehmen*, in: Uwe, Hannig (Hrsg.): *Data Warehouse und Managementinformationssysteme*, S. 25–31, Schäffer-Poeschel, Stuttgart 1996.

FLECHTNER, HANS-JOACHIM (1984): *Grundbegriffe der Kybernetik. Eine Einführung*, Deutscher Taschenbuch-Verlag, München 1984.

FOX, CHRISTOPHER J.; LEVITIN, ANANY V. und REDMAN, THOMAS C. (1994): *The Notion of Data and its Quality Dimensions*, in: Information Processing and Management, 30. Jg., Nr. 1 (1994), S. 9–19.

FRANK, ULRICH; KLEIN, STEFAN; KRCMAR, HELMUT und TEUBNER, ALEXANDER (1999): *Aktionsforschung in der WI - Einsatzpotentiale und Einsatzprobleme*, in: Schütte, Reinhard; Siedentopf, Jukka und Zelewski, Stephan (Hrsg.): *Wirtschaftsinformatik und Wissenschaftstheorie. Grundpositionen und Theoriekerne. Arbeitsbericht des Instituts für Produktion und Industrielles Informationsmanagement Nr. 4*, S. 71–90, Universität GH Essen, Essen 1999.

FRIES, STEFAN (1994): *Neuorientierung der Qualitätskostenrechnung in prozessorientierten TQM-Unternehmen: Entwurf eines ganzheitlichen Entwicklungsprozesses zur Auswahl von Prozessmessgrössen*, Dissertation Hochschule St. Gallen, Rosch-Buch, Hallstadt 1994.

GABRIEL, ROLAND und RÖHRS, HEINZ-PETER (1995): *Datenbanksysteme: konzeptionelle Datenmodellierung und Datenbankarchitekturen*, 2. Aufl., Springer, Berlin et al. 1995.

GALHARDAS, HELENA; FLORESCU, DANIELA; SHASHA, DENNIS und SIMON, ERIC (2000): *Declaratively Cleansing your Data using AJAX*, in: *Journées Bases de Données Avancées (BDA)*, Blois, France 2000.

GARVIN, DAVID A. (1984): *What does 'Product Quality' really mean?*, in: Sloan Management Review, 26. Jg., Nr. 1 (1984), S. 25–43.

GEIGER, WALTER (1994): *Qualitätslehre - Einführung - Systematik - Terminologie*, 2. Aufl., Vieweg, Braunschweig und Wiesbaden 1994.

GEIGER, WALTER (2001): *Qualität als Fachbegriff des QM*, in: Zollondz, Hans-Dieter (Hrsg.): *Lexikon Qualitätsmanagement: Handbuch des modernen Managements auf der Basis des Qualitätsmanagements*, S. 801–810, Oldenbourg, München und Wien 2001.

GERTZ, WINFRIED (1999): *Data-Warehouse: Ein Datenfriedhof ist keine Informationsquelle*, in: Computerwoche, o. Jg., Nr. 13 (1999), S. 49–50.

GILLIES, ALAN C. (1992): *Software Quality: Theory and management*, Chapman and Hall, London et al. 1992.

GRIMMER, UDO und HINRICHS, HOLGER (2001): *A Methodological Approach to Data Quality Management Supported by Data Mining*, in: Pierce, Elizabeth M. und Kaatz-Haas, Raissa (Hrsg.): *Proceedings of the Sixth International Conference on Information Quality*, S. 217–232, Massachusetts Institute of Technology, Cambridge, MA 2001.

GROCHLA, ERWIN (1975): *Betriebliche Planung und Informationssysteme*, Rowohlt, Reinbek bei Hamburg 1975.

GROTZ-MARTIN, SILVIA (1976): *Informations-Qualität und Informations-Akzeptanz in Entscheidungsprozessen - Theoretische Ansätze und ihre empirische Überprüfung*, Dissertation, Universität des Saarlandes, 1976.

GUNTRAM, ULRICH (1985): *Die Allgemeine Systemtheorie - Ein Überblick*, in: Zeitschrift für Betriebswirtschaft, 55. Jg., Nr. 3 (1985), S. 296–323.

HABERFELLNER, REINHARD (1975): *Die Unternehmung als dynamisches System. Der Prozesscharakter der Unternehmensaktivitäten*, Reihe Forschungsberichte für die Unternehmenspraxis des BWI ETH Zürich, Verl. Industrielle Organisation, Zürich 1975.

HANSEN, HANS ROBERT (1996): *Grundlagen betrieblicher Informationsverarbeitung*, Nr. 802 der Reihe UTB für Wissenschaft: Uni-Taschenbücher, Grundwissen der Ökonomik: Betriebswirtschaftslehre, 7. Aufl., Lucius und Lucius, Stuttgart 1996.

HARTUNG, JOACHIM; ELPELT, BÄRBEL und KLÖSENER, KALR-HEINZ (1998): *Statistik: Lehr- und Handbuch der angewandten Statistik*, Oldenbourg, München und Wien 1998.

HARVEY, ANDREW C. (1995): *Zeitreihenmodelle*, 2. Aufl., Oldenbourg, München und Wien 1995.

HAUKE, PETER (1984): *Informationsverarbeitungsprozesse und Informationsbewertung*, GBI-Verl., München 1984.

HÄUSSLER, CHRISTA (1998): *Datenqualität*, in: Wolfgang, Martin (Hrsg.): *Data Warehousing*, S. 75–89, Internat. Thomson Publ., Bonn 1998.

HAYKIN, SIMON (1999): *Neural networks - A comprehensive foundation*, 2. Aufl., Prentice-Hall, Upper Saddle River, NJ 1999.

HEINE, PETER (1999): *Unternehmensweite Datenintegration: modular-integrierte Datenlogistik in betrieblichen Informationssystemen,* zugl. Dissertation Universität Leipzig, Teubner, Stuttgart und Leipzig 1999.

HEINRICH, LUTZ J. (1992): *Informationsmanagement: Planung, Überwachung und Steuerung der Informations-Infrastruktur,* 4. Aufl., Oldenbourg, München und Wien 1992.

HELFERT, MARKUS (2000a): *Massnahmen und Konzepte zur Sicherung der Datenqualität,* in: Jung, Reinhard und Winter, Robert (Hrsg.): *Data-Warehousing-Strategie: Erfahrungen, Methoden, Visionen,* S. 61–77, Springer, Berlin et al. 2000.

HELFERT, MARKUS (2000b): *Eine empirische Untersuchung von Forschungsfragen beim Data Warehousing aus Sicht der Unternehmenspraxis,* Arbeitsbericht BE HSG/CC DWS/05, Institut für Wirtschaftsinformatik der Universität St. Gallen, St. Gallen 2000.

HELFERT, MARKUS und RADON, RENATE (2000): *An Approach for Information Quality measurement in Data Warehousing*, in: Klein, Barbara D. und Rossin, Donald F. (Hrsg.): *Proceedings of the 2000 Conference on Information Quality,* S. 109–125, Massachusetts Institute of Technology, Cambridge, MA 2000.

HELFERT, MARKUS und VON MAUR, EITEL (2001): *A Strategy for Managing Data Quality in Data Warehouse Systems*, in: Pierce, Elizabeth M. und Kaatz-Haas, Raissa (Hrsg.): *Proceedings of the Sixth International Conference on Information Quality,* S. 62–76, Massachusetts Institute of Technology, Cambridge, MA 2001.

HELFERT, MARKUS; HERRMANN, CLEMENS und STRAUCH, BERNHARD (2001): *Datenqualitätsmanagement,* Arbeitsbericht BE HSG/CC DW2/02, Institut für Wirtschaftsinformatik der Universität St. Gallen, St. Gallen 2001.

HENNEBÖLE, JÖRG (1995): *Executive Information Systems für Unternehmensführung und Controlling - Strategie, Konzeption, Realisierung*, Gabler, Wiesbaden 1995.

HEUER, ANDREAS und SAAKE, GUNTER (2000): *Datenbanken. Konzepte und Sprachen,* 2. Aufl., mitp-Verlag, Bonn 2000.

HINRICHS, HOLGER (2001): *Datenqualitätsmanagement in Data Warehouse-Umgebungen*, in: *Datenbanksysteme in Büro, Technik und Wissenschaft,*

9. GI-Fachtagung BTW 2001 Oldenburg, S. 187–206, Springer, Berlin et al. 2001.

HINRICHS, HOLGER und ADEN, THOMAS (2001): *An ISO 9001:2000 Compliant Quality Management System for Data Integration in Data Warehouse Systems*, in: Theodoratos, Dimitri; Hammer, Joachim; Jeusfeld, Manfred A. und Staudt, Martin (Hrsg.): *Proceedings of the Intenational Workshop on Design and Management of Data Warehouses (DMDW'2001)*, S. 1/1 – 1/12, Interlaken 2001.

HOLTEN, ROLAND (1999): *Entwicklung von Führungsinformationssystemen: ein methodenorientierter Ansazt*, zugl. Dissertation Universität Münster, Gabler, Wiesbaden 1999.

HOLTEN, ROLAND; KNACKSTEDT, RALF und BECKER, JÖRG (2001a): *Betriebswirtschaftliche Herausforderungen durch Data-Warehouse-Technologien*, in: Schütte, Reinhard; Rotthowe, Thomas und Holten, Roland (Hrsg.): *Data Warehouse Managementhandbuch: Konzepte, Software, Erfahrungen*, S. 41–64, Springer, Berlin et al. 2001.

HOLTEN, ROLAND; ROTTHOWE, THOMAS und SCHÜTTE, REINHARD (2001b): *Grundlagen, Einsatzbereiche, Modelle*, in: Schütte, Reinhard; Rotthowe, Thomas und Holten, Roland (Hrsg.): *Data Warehouse Managementhandbuch: Konzepte, Software, Erfahrungen*, S. 3–24, Springer, Berlin et al. 2001.

HOLTHUIS, JAN (1999): *Der Aufbau von Data Warehouse-Systemen: Konzeption - Datenmodellierung - Vorgehen*, zugl. Dissertation Universität Göttingen, 2. Aufl., Gabler, Wiesbaden 1999.

HUANG, KUAN-TSAE; LEE, YANG W. und WANG, RICHARD Y. (1999): *Quality Information and Knowledge*, Prentice Hall, Upper Saddle River, NJ 1999.

INMON, WILLIAM H. (1996): *Building the Data Warehouse*, 2. Aufl., Wiley, New York et al. 1996.

INMON, WILLIAM H.; ZACHMAN, JOHN A. und GEIGER, JONATHAN G. (1997): *Data Stores, Data Warehousing and the Zachman Framework: Managing Enterprise Knowledge*, McGraw-Hill, New York et al. 1997.

INMON, WILLIAM H.; IMHOFF, CLAUDIA und SOUSA, RYAN (1998): *Corporate Information Factory: A proven approach to integrating: data marts and data warehouses*, Wiley, New York et al. 1998.

IVES, BLAKE; OLSEN, MARGRETHE H. und BAROUDI, JACK J. (1983): *The measurement of user informations satisfaction*, in: Communications of the ACM, 26. Jg., Nr. 10 (1983), S. 785–793.

JANTSCH, ERICH (1994): *Systemtheorie*, in: Seiffert, Helmut und Radnitzky, Gerard (Hrsg.): *Handlexikon zur Wissenschaftstheorie*, 2. Aufl., S. 331–338, Ehrenwirth, München 1994.

JARKE, MATTHIAS und VASSILIOU, YANNIS (1997): *Foundations of Data Warehouse Quality - A Review of the DWQ Project*, in: Strong, Diane M. und Kahn, Beverly K. (Hrsg.): *Proceedings of the 1997 Conference on Information Quality*, S. 299–313, Massachusetts Institute of Technology, Cambridge, MA 1997.

JARKE, MATTHIAS; JEUSFELD, MANFRED A.; QUIX, CHRISTOPH und VASSILIADIS, PANOS (1999): *Architecture and Quality in Data Warehouses: An Extended Repository Approach*, in: Information Systems, 24. Jg., Nr. 3 (1999), S. 229–253.

JARKE, MATTHIAS; LENZERINI, MAURIZIO; VASSILIOU, YANNIS und VASSILIADIS, PANOS (2000): *Fundamentals of data warehouses*, Springer, Berlin et al. 2000.

JIN, BINLING und EMBURY, SUZANNE M. (2001): *Non-Intrusive Assessment of Organisational Data Quality*, in: Pierce, Elizabeth M. und Kaatz-Haas, Raissa (Hrsg.): *Proceedings of the Sixth International Conference on Information Quality*, S. 398–411, Massachusetts Institute of Technology, Cambridge, MA 2001.

JUNG, REINHARD und ROWOHL, FREDERIC (2000): *Vergleichende Analyse der Standardisierungsbestrebungen im Metadatenmanagement: Neue Metamodelle der MDC und OMG*, in: Jung, Reinhard und Winter, Robert (Hrsg.): *Data Warehousing 2000 - Methoden, Anwendungen, Strategien*, S. 113–133, Physica-Verlag, Heidelberg 2000.

JUNG, REINHARD und WINTER, ROBERT (2000): *Data Warehousing: Nutzungsaspekte, Referenzarchitektur und Vorgehensmodell*, in: Jung, Reinhard und Winter, Robert (Hrsg.): *Data-Warehousing-Strategie: Erfahrungen, Methoden, Visionen*, S. 3–20, Springer, Berlin et al. 2000.

JURAN, JOSEPH M. (1999): *How to think about Quality*, in: Juran, Joseph M. und Godfrey, A. Blanton (Hrsg.): *Juran's Quality Handbook*, 5. Aufl., Kap. 2, S. 1–18, McGraw Hill, New York et al. 1999.

KAHN, BEVERLY K.; STRONG, DIANE M. und WANG, RICHARD Y. (1997): *A Model for Delivering Quality Information as Product and Services*, in: Strong, Diane M. und Kahn, Beverly K. (Hrsg.): *Proceedings of the 1997 Conference on Information Quality*, S. 80–94, Massachusetts Institute of Technology, Cambridge, MA 1997.

KAMINSKY, FRANK (2000): *Das Metadaten-Gesteuerte Data Warehouse*, Vortrag auf dem 1. DWH-Forum St. Gallen am 26.02. 2000; Universität St. Gallen (http://www.dwh-forum.iwi.unisg.ch), 2000.

KANDZIA, PETER und KLEIN, HANS-JOACHIM (1993): *Theoretische Grundlagen relationaler Datenbanken*, Reihe Informatik; Bd. 79, BI-Wissenschaftsverlag, Mannheim et al. 1993.

KAPOSI, AGNES und MYERS, MARGARET (1994): *Systems, Models and Measures*, Springer, Berlin et al. 1994.

KEMPER, ALFONS und EICKLER, ANDRE (1996): *Datenbanksysteme: eine Einführung*, Oldenbourg, München und Wien 1996.

KEPPEL, BERND; MÜLLENBACH, STEFAN und WÖLKHAMMER, MARKUS (2001): *Vorgehensmodelle im Bereich Data Warehouse: Das Evolutionary Data Warehouse Engineering (EDE)*, in: Schütte, Reinhard; Rotthowe, Thomas und Holten, Roland (Hrsg.): *Data Warehouse Managementhandbuch: Konzepte, Software, Erfahrungen*, S. 81–105, Springer, Berlin et al. 2001.

KETTING (1999): *Geschichte des Qualitätsmanagements*, in: Masing, Walter (Hrsg.): *Handbuch Qualitätsmanagement*, S. 17–30, Hanser, München und Wien 1999.

KOPCSO, DAVID; PIPINO, LEO und RYBOLT, WILLIAM (2000): *The Assessment of Web Site Quality*, in: Klein, Barbara D. und Rossin, Donald F. (Hrsg.): *Proceedings of 2000 Conference on Information Quality*, S. 97–108, Massachusetts Institute of Technology, Cambridge, MA 2000.

KROMREY, HELMUT (1998): *Empirische Sozialforschung: Modelle und Methoden der Datenerhebung und Datenauswertung*, 8. Aufl., Leske und Budrich, Opladen 1998.

LAMNEK, SIEGFRIED (1995): *Qualitative Sozialforschung - Methodologie*, Bd. 1, 3. Aufl., Belz Psychologie Verlags Union, Weinheim 1995.

LAUDON, KENNETH C. (1986): *Data quality and due process in large interorganizational record systems*, in: Communication of the ACM, 29. Jg., Nr. 1 (1986), S. 4–11.

LEHMANN, PETER und ORTNER, ERICH (2000): *Entwurf einer Beschreibungskomponente für fachliche (Meta-)Daten aus einem Data Warehouse-Repository*, in: Jung, Reinhard und Winter, Robert (Hrsg.): *Data Warehousing 2000 - Methoden, Anwendungen, Strategien*, S. 365–393, Physica-Verlag, Heidelberg 2000.

LEVITIN, ANANY V. und REDMAN, THOMAS C. (1995): *Quality Dimensions of a Conceptual View*, in: Information Processing and Management, 31. Jg., Nr. 1 (1995), S. 81–88.

MAIER, RONALD und LEHNER, FRANZ (1995): *Daten, Informationen, Wissen*, in: Lehner, Franz; Maier, Roland und Hildebrand, Knut (Hrsg.): *Wirtschaftsinformatik*, S. 165–272, Hanser, München und Wien 1995.

MANDKE, VIJAY V. und NAYAR, MADHAVAN K. (1998): *Information Integrity Technology Product Structure*, in: Chengalur-Smith, InduShobha und Pipino, Leo L. (Hrsg.): *Proceedings of the 1998 Conference on Information Quality*, S. 232–246, Massachusetts Institute of Technology, Cambridge, MA 1998.

MEIDAN, ABRAHAM (2001): *WizRule - White Paper*, http://www.wizsoft.com (Download am 18.12. 2001), 2001.

MERTENS, PETER (1995): *Integrierte Informationsverarbeitung Band 1 - Administrations- und Dispositionssysteme in der Industrie*, 10. Aufl., Gabler, Wiesbaden 1995.

MERTENS, PETER und HOLZNER, JOCHEN (1992): *WI - State of the Art: Eine Gegenüberstellung von Integrationsansätzen der Wirtschaftsinformatik*, in: Wirtschaftsinformatik, 34. Jg., Nr. 1 (1992), S. 5–25.

MERTENS, PETER und WIECZORREK, HANS WILHELM (1999): *Data X Strategien: Data Warehouse, Data Mining und operationale Systeme für die Praxis*, Spinger, Berlin et al. 1999.

MERTENS, PETER; BISSANTZ, NICOLAS; HAGEDORN, JÜRGEN und SCHULTZ, JENS (1994): *Datenmustererkennung in der Ergenisrechnung mit Hilfe der Clusteranalyse*, in: Die Betriebswirtschaft, 54. Jg., Nr. 6 (1994), S. 739–753.

MERTENS, PETER; BODENDORF, FREIMUT; KÖNIG, WOLFGANG; PICOT, ARNOLD und SCHUMANN, MATTHIAS (2000): *Grundzüge der Wischaftsinformatik*, 6. Aufl., Springer, Berlin et al. 2000.

MEYER, CLAUS (1994): *Betriebswirtschaftliche Kennzahlen und Kennzahlensysteme*, 2. Aufl., Schäffer-Poeschel, Stuttgart 1994.

MEYER, MARKUS (2000): *Organisatorische Gestaltung des unternehmensweiten Data Warehousing: Konzeption der Rollen, Verantwortlichkeiten und Prozesse am Beispiel einer Schweizer Universalbank*, Dissertation Universtität St. Gallen, Difo-Druck, Bamberg 2000.

MEYER, MARKUS und WINTER, ROBERT (2000): *Organisation des unternehmensweiten Data Warehousing*, in: Jung, Reinhard und Winter, Robert (Hrsg.): *Data Warehousing 2000 - Methoden, Anwendungen, Strategien*, S. 309–331, Physica-Verlag, Heidelberg 2000.

MICROSOFT CORP. (1998): *OLE DB for OLAP - Version 1.0 Specification*, http://www.microsoft.com/Data/oledb/olap/spec/ (Download am 18.12. 2001), 1998.

MILEK, JANUSZ; REIGROTZKI, MARTIN; BOSCH, HOLGER und BLOCK, FRANK (2001): *Monitoring and Data Quality Control of Financial Databases from a Process Control Perspective*, in: Pierce, Elizabeth M. und Katz-Haas, Raissa (Hrsg.): *Proceedings of the Sixth International Conference on Information Quality*, S. 189–205, Massachusetts Institute of Technology, Cambridge, MA 2001.

MILLER, HOLMES (1996): *The multiple dimensions of information quality*, in: Information Systems Management, 13. Jg., Nr. 2 (1996), S. 79–82.

MOREY, RICHARD C. (1982): *Estimating and improving the quality of information in the MIS*, in: Communications of the ACM, 25. Jg., Nr. 5 (1982), S. 337–342.

MUCKSCH, HARRY (1998): *Das Data Warehouse als Datenbasis analytischer Informationssysteme - Architektur und Komponenten*, in: Chamoni, Peter und Gluchowski, Peter (Hrsg.): *Analytische Informationssysteme: Data Warehouse, On-line Analytical Processing, Data Mining*, S. 123–140, Springer, Berlin et al. 1998.

MUCKSCH, HARRY und BEHME, WOLFGANG (2000): *Das Data Warehouse-Konzept als Basis einer unternehmensweiten Informationslogistik*, in:

Mucksch, Harry und Behme, Wolfgang (Hrsg.): *Das Data-Warehouse-Konzept - Datenmodelle - Anwendungen: mit Erfahrungsberichten*, 4. Aufl., S. 3–80, Gabler, Wiesbaden 2000.

MÜLLER, JOCHEN (2000): *Transformation operativer Daten zur Nutzung im Data Warehouse*, zugl. Dissertation Universität Bochum, Gabler, Wiesbaden 2000.

MÜLLER-BÖLING, DETLEF und KLANDT, HEINZ (1996): *Methoden empirischer Wirtschafts- und Sozialforschung: eine Einführung mit wirtschaftswissenschaftlichem Schwerpunkt*, 3. Aufl., Förderkreis Gründungs-Forschung, Köln und Dortmund 1996.

MUTSCHELLER, ANDREAS MARTIN (1996): *Vorgehensmodell zur Entwicklung von Kennzahlen und Indikatoren für das Qualitätsmanagement*, Dissertation Universität St. Gallen, Difo-Druck, Bamberg 1996.

NAUMANN, FELIX und ROLKER, CLAUDIA (1999): *Do metadata models meet IQ requirements?*, in: Lee, Yang W. und Tayi, Giri Kumar (Hrsg.): *Proceedings of the 1999 Conference on Information Quality*, S. 99–114, Massachusetts Institute of Technology, Cambridge, MA 1999.

NAUMANN, FELIX und ROLKER, CLAUDIA (2000): *Assessment Methods for Information Quality Criteria*, in: Klein, Barbara D. und Rossin, Donald F. (Hrsg.): *Proceedings of the 2000 Conference on Information Quality*, S. 148–162, Massachusetts Institute of Technology, Cambridge, MA 2000.

NIESCHLAG, ROBERT; DICHTL, ERWIN und HÖRSCHGEN, HANS (1994): *Marketing*, 17. Aufl., Duncker und Humblot, Berlin 1994.

OBJECT MANAGEMENT GROUP, INC. (2001): *Common Warehouse Metamodell (CWM) Specification, Version 1.0 vom 2. Februar 2001*, http://www.omg.org/technolgy/cwm/index.htm (Download am 21.09. 2001).

OSANNA, HERBERT (2001): *SPC - Statistical Process Control*, in: Zollondz, Hans-Dieter (Hrsg.): *Lexikon Qualitätsmanagement: Handbuch des modernen Managements auf der Basis des Qualitätsmanagements*, S. 1101–1105, Oldenbourg, München und Wien 2001.

ÖSTERLE, HUBERT (1995): *Business Engineering: Prozess- und Systementwicklung Band 1: Entwurfstechniken*, Springer, Berlin et al. 1995.

ÖSTERLE, HUBERT; BRENNER, WALTER und HILBERS, KONRAD (1991): *Unternehmensführung und Informationssystem: Der Ansatz des St. Galler Informationssystem-Managements*, 2. Aufl., Teubner, Stuttgart 1991.

PETERHAUS, MARKUS (1995): *Informationsmanagement*, in: Lehner, Franz; Maier, Ronald und Hildebrand, Knut (Hrsg.): *Wirtschaftsinformatik: theoretische Grundlage*, S. 327–368, Hanser, München und Wien 1995.

PFEIFER, TILO (1996): *Praxishandbuch Qualitätsmanagement*, Hanser, München und Wien 1996.

PFEIFER, TILO (2001): *Qualitätsregelkreis*, in: Zollonds, Hans-Dieter (Hrsg.): *Lexikon Qualitätsmanagement: Handbuch des modernen Managements auf Basis des Qualitätsmanagements*, S. 998–1002, Oldenbourg, München und Wien 2001.

PICOT, ARNOLD und FRANK, EGON (1988): *Die Planung der Unternehmensressource Information - Teil 2*, in: WISU - Das Wirtschaftsstudium, 27. Jg., Nr. 11 (1988), S. 608–614.

PICOT, ARNOLD und REICHWALD, RALF (1991): *Informationswirtschaft*, in: Heinen, Edmund (Hrsg.): *Industriebetriebslehre: Entscheidungen im Industriebetrieb*, 9. Aufl., S. 241–390, Gabler, Wiesbaden 1991.

POE, VIDETTE (1998): *Building a Data Warehouse for Decision Support*, 2. Aufl., Prentice-Hall, Upper Saddle River, NJ 1998.

POKORNY, L. ROBERT (2000): *Assigning a Quality Measurement to Matching Records form Heterogeneous Legacy Databases: A Practical Experience*, in: Klein, Barbara D. und Rossin, Donald F. (Hrsg.): *Proceedings of 2000 Conference on Information Quality*, S. 70–75, Massachusetts Institute of Technology, Cambridge, MA 2000.

PROBST, GILBERT J. B. und RAUB, STEFFEN (1995): *Action Research - Ein Konzept angewandter Managementforschung*, in: Die Unternehmung, 49. Jg., Nr. 1 (1995), S. 3–19.

RAUH, OTTO und STICKEL, EBERHARD (1997): *Konzeptuelle Datenmodellierung*, Teubner, Stuttgart und Leipzig 1997.

RAUTENSTRAUCH, CLAUS (1997): *Effiziente Gestaltung von Arbeitsplatzsystemen*, Addison-Wesley, Bonn et al. 1997.

REDMAN, THOMAS C. (1996): *Data Quality for the information age*, Artech House, Norwood, MA 1996.

REICHMANN, THOMAS (1995): *Controlling mit Kennzahlen und Managementberichten: Grundlage einer systemgestützten Controlling-Konzeption*, 4. Aufl., Vahlen, München 1995.

RINNE, HORST und MITTAG, HANS-JOACHIM (1995): *Statistische Methoden der Qualitätssicherung*, 3. Aufl., Hanser, München und Wien 1995.

ROCKART, JOHN F. (1979): *Chief executives define their own data needs*, in: Harvard Business Review, 57. Jg., Nr. 2 (1979), S. 81–93.

ROSEMANN, MICHAEL (1996): *Komplexitätsmanagement in Prozessmodellen. Methodenspezifische Gestaltungsempfehlungen für die Informationsmodellierung*, zugl. Dissertation Universität Münster, Gabler, Wiesbaden 1996.

ROSEMANN, MICHAEL und ROTTHOWE, THOMAS (1995): *Der Lösungsbeitrag von Prozessmodellen bei der Einführung von SAP R/3 im Finanz- und Rechnungswesen.*, in: HMD - Theorie und Praxis der Wirtschaftsinformatik, 32. Jg., Nr. 182 (1995), S. 8–25.

RÜTTLER, MARTIN (1991): *Information als strategischer Produktionsfaktor*, Erich Schmidt Verlag, Berlin 1991.

SACHS, LOTHAR (1999): *Angewandte Statistik: Anwendung statistischer Methoden*, 9. Aufl., Springer, Berlin et al. 1999.

SCHEER, AUGUST-WILHELM (1998): *Wirtschaftsinformatik: Referenzmodelle für industrielle Geschäftsprozesse*, Studienausgabe, 2. Aufl., Springer, Berlin et al. 1998.

SCHELP, JOACHIM (2000): *Modellierung mehrdimensionaler Datenstrukturen analyseorientierter Informationssysteme*, zugl. Dissertation Universität Bochum, Gabler, Wiesbaden 2000.

SCHINZER, HEIKO D.; BANGE, CARSTEN und MERTENS, HOLGER (1999): *Data warehouse and Data mining - Marktführende Produkte im Vergleich*, 2. Aufl., Vahlen, München 1999.

SCHLITTGEN, RAINER und STREITBERG, BERND H. J. (1999): *Zeitreihenanalyse*, 8. Aufl., Oldenbourg, München und Wien 1999.

SCHMITZ, PAUL (1990): *Softwarequalitätssicherung*, in: Kurbel, Karl und Strunz, Horst (Hrsg.): *Handbuch Wirtschaftsinformatik*, S. 309–320, Schäffer-Poeschel, Stuttgart 1990.

SCHREIER, ULF (2001): *Data Dictionary*, in: Mertens, Peter et al. (Hrsg.): *Lexikon der Wirtschaftsinformatik*, 4. Aufl., S. 129–130, Springer, Berlin et al. 2001.

SCHÜTTE, REINHARD (1998): *Grundsätze ordnungsmässiger Referenzmodellierung: Konstruktion konfigurations- und anpassungsorientierter Modelle*, zugl. Dissertation Universität Münster, Gabler, Wiesbaden 1998.

SCHWARZ, STEFAN (2000): *Integriertes Metadatenmanagement - Ein Überblick*, in: Jung, Reinhard und Winter, Robert (Hrsg.): *Data Warehousing Strategie: Erfahrungen, Methoden, Visionen*, S. 101–116, Springer, Berlin et al. 2000.

SCHWEGMANN, ANSGAR (1999): *Objektorientierte Referenzmodellierung: theoretische Grundlagen und praktische Anwendung*, zugl. Dissertation Universität Münster, Gabler, Wiesbaden 1999.

SCHWINN, KLAUS; DIPPOLD, ROLF; RINGGENBERG, ANDRÉ und SCHNIDER, WALTER (1999): *Unternehmensweites Datenmanagement - Von der Datenbankadministration bis zum modernen Informationsmanagement*, 2. Aufl., Gabler, Wiesbaden 1999.

SEGHEZZI, HANS DIETER (1996): *Integriertes Qualitätsmanagment - das St. Galler Konzept*, Hanser, München und Wien 1996.

SEIDL, JÖRG (2001): *Business Data Intelligence - Mittels Data Mining die Qualität von Geschäftsdaten in den Griff bekommen*, T-System Konferenz - Erfolgsfaktor Informationsqualität, am 20. Mai 2001 in Frankfurt, 2001.

SEIFFERT, HELMUT (1992): *Einführung in die Wissenschaftstheorie, Band 3: Handlungstheorie, Modallogik, Ethik, Systemtheorie*, 2. Aufl., Beck, München 1992.

SEIFFERT, HELMUT (1994): *System*, in: Seiffert, Helmut und Radnitzky, Gerard (Hrsg.): *Handlexikon zur Wissenschaftstheorie*, S. 329–331, Ehrenwirth, München 1994.

SIEGWART, HANS (1998): *Kennzahlen für die Unternehmensführung*, 5. Aufl., Haupt, Bern et al. 1998.

SINZ, ELMAR J. (1997): *Architektur von Informationssystemen*, in: Rechnberg, Peter und Pomberger, Gustav (Hrsg.): *Informatik Handbuch*, S. 875–887, Hanser, München 1997.

SOEFFKY, MANFRED (1998): *Data Warehouse: Prozess- und Systemmanagement*, IT Research, Höhenkirchen 1998.

SOEFFKY, MANFRED (1999): *Operative Datenqualität nich überschätzen!*, in: CW-focus, o. Jg., Nr. 1 (1999), S. 8–10.

SOLER, SABRINA VÁZQUEZ und YANKELEVICH, DANIEL (2001): *Quality Mining: A Data Mining Method for Data Quality Evaluation*, in: Pierce, Elizabeth M. und Kaatz-Haas, Raissa (Hrsg.): *Proceedings of the Sixth International Conference on Information Quality*, S. 162–172, Massachusetts Institute of Technology, Cambridge, MA 2001.

STAHLKNECHT, PETER und HASENKAMP, ULRICH (1999): *Einführung in die Wirtschaftsinformatik*, 9. Aufl., Springer, Berlin et al. 1999.

STATISTISCHES BUNDESAMT (1999): *Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ 93)*, http://www.destatis.de/download/klassif/wz93int.pdf (Download am 18.12. 2001), Wiesbaden.

STIER, WIENFRIED (2001): *Methoden der Zeitreihenanalyse*, Springer, Berlin et al. 2001.

STOCK, STEFFEN (2001): *Modellierung zeitbezogener Daten im Data Warehouse*, Zugl. Dissertation Universität Duisburg, Gabler, Wiesbaden 2001.

STOREY, VEDA C. und WANG, RICHARD Y. (1998): *Modeling Quality Requirements in Conceptual Database Design*, in: Chengalur-Smith, InduShobha und Pipino, Leo L. (Hrsg.): *Proceedings of the 1998 Conference on Information Quality*, S. 64–87, Massachusetts Institute of Technology, Cambridge, MA 1998.

STRAHRINGER, SUSANNE (1996): *Zum Begriff des Metamodells*, in: Schriften zur Quantitativen Betriebswirtschaftslehre, o. Jg., Nr. 6 (1996).

STREUBEL, FRAUKE (1996): *Theoretische Fundierung eines ganzheitlichen Informationsmanagements*, Arbeitsbericht 96-21, Lehrstuhl für Wirtschaftsinformatik, Ruhr-Universität Bochum 1996.

THALHEIM, BERNHARD (1991): *Dependencies in Relational Databases*, Teubner, Stuttgart und Leipzig 1991.

THOMAS COOK AG, KONZERNCONTROLLING (2001): *DWH-Datenqualitätsmanagement-Tool*, Bildschirmausschnitt, Thomas Cook AG, Konzerncontrolling, Oberursel, 2001.

TOZER, GUY V. (1999): *Metadata management for information control and business success*, Artech House, Norwood, MA 1999.

ULRICH, HANS (1970): *Die Unternehmung als produktives soziales System*, 2. Aufl., Haupt, Bern und Stuttgart 1970.

ULRICH, HANS (1984): *Management*, Haupt, Bern und Stuttgart 1984.

VETTER, MAX (1994): *Informationssysteme in der Unternehmung - eine Einführung in die Datenmodellierung und Anwendungsentwicklung*, 2. Aufl., Teubner, Stuttgart 1994.

VETTER, MAX (1998): *Aufbau betrieblicher Informationssysteme mittels pseudo-objektorientierter konzeptioneller Datenmodellierung*, 8. Aufl., Teubner, Stuttgart 1998.

VOSSEN, GOTTFRIED (2000): *Datenmodelle, Datenbanksprachen und Datenbankmanagementsysteme*, 4. Aufl., Oldenbourg, München und Wien 2000.

WALLMÜLLER, ERNEST (1990): *Software-Qualitätssicherung in der Praxis*, Hanser, München und Wien 1990.

WALLMÜLLER, ERNEST (1995): *Ganzheitliches Qualitätsmanagement in der Informationsverarbeitung*, Hanser, München und Wien 1995.

WAND, YAIR und WANG, RICHARD Y. (1996): *Anchoring Data Quality Dimensions in Ontological Foundations*, in: Communications of the ACM, 39. Jg., Nr. 11 (1996), S. 86 – 95.

WANG, RICHARD Y. und STRONG, DIANE M. (1996): *Beyond Accuracy: What Data Quality Means to Data Consumers*, in: Journal of Management Information Systems, 12. Jg., Nr. 4 (1996), S. 5–33.

WANG, RICHARD Y.; KON, HENRY B. und MADNICK, STUARD E. (1993): *Data quality requirements analysis and modeling*, in: *Proceedings of the 9th International Conferece on Data Engineering (ICDE)*, S. 670 – 677, IEEE Computer Society, Wien 1993.

WANG, RICHARD Y.; REDDY, M. P. und KON, HENRY B. (1995a): *Toward quality data-An attribute-based approach*, in: Decision Support System, o. Jg., Nr. 13 (1995), S. 349–372.

WANG, RICHARD Y.; STOREY, VEDA C. und FIRTH, CHRISTOHPER P. (1995b): *A framework for analysis of data quality research*, in: IEEE Transactions on Knowledge and Data Engineering, 7. Jg., Nr. 4 (1995), S. 623–640.

WANG, RICHARD Y.; STRONG, DIANE M.; KAHN, BEVERLY K. und LEE, YANG W. (1999): *An information quality assessment methodology*, in: Lee, Yang W. und Tayi, Giri Kumar (Hrsg.): *Proceedings of the 1999 Conference on Information Quality*, S. 258–265, Massachusetts Institute of Technology, Cambridge, MA 1999.

WANG, RICHARD Y.; ZIAD, MOSTAPHA und LEE, YANG W. (2001): *Data Quality*, Kluwer Academic Publishers, Boston et al. 2001.

WANG, XUE Z. (1999): *Data Mining and Knowledge Discovery for Process Monitoring and Control*, Springer, Berlin et al. 1999.

WATSON, HUGH und HALEY, BARBARA J. (1998): *Data Warehousing: A Framework and Survey of Current Practices*, in: Chamoni, Peter und Gluchowski, Peter (Hrsg.): *Analytische Informationssysteme: Data Warehouse, On-line Analytical Processing, Data Mining*, S. 27–39, Springer, Berlin et al. 1998.

WEDEKIND, HARTMUT (2001): *Datenbanksystem*, in: Mertens, Peter et al. (Hrsg.): *Lexikon der Wirtschaftsinformatik*, 4. Aufl., S. 139–140, Springer, Berlin et al. 2001.

WEGENER, HANS (2000): *Erste Erfahrungen mit Komponenten, Metadaten und Wiederverwendung im Data Warehouse der Credit Suisse*, in: Flatscher, Rony G. und Turowskie, Klaus (Hrsg.): *Workshop komponentenorientierte betriebliche Anwendungssysteme*, S. 81–93, Wirtschaftsuniversität Wien, Wien 2000.

WEGENER, HANS (2001): *Datenqualitäts- und Metadata-Management: Eine sinnvolle Liaison?*, Vortrag auf dem 2. CC DW2 Workshop; Ermatingen, Schloss Wolfsberg am 6. und 7. Juni 2001, 2001.

WEIHS, CLAUS; JESSENBERGER, JUTTA und GRIZE, YVES-LAURENT (1999): *Statistische Methoden zur Qualitätssicherung und -optimierung in der Industrie*, Wiley-VCH, New York et al. 1999.

WEIKUM, GERHARD (1999): *Towards guaranteed quality and dependability of information systems*, in: Buchmann, Alejandro P. (Hrsg.): *Datenbanksysteme in Büro, Technik und Wissenschaft, 8. GI-Fachtagung, Freiburg im Breisgau, 1.-3. März 1999*, S. 379–409, Springer, Berlin et al. 1999.

WEISS, PETER (1987): *Stochastische Modelle für Anwender*, Teubner, Stuttgart 1987.

WELCH, J. D. (1997): *Updating the Data Warehouse*, in: Barquin, Ramon C. und Edlestein, Herbert A. (Hrsg.): *Building, Using and Managing the Data Warehouse*, S. 173–210, Prentice-Hall, Upper Saddle River, NJ 1997.

WIEKEN, JOHN-HARRY (1997): *Meta-Daten für Data Marts und Data Warehouse*, in: Mucksch, Harry und Behme, Wolfgang (Hrsg.): *Das Data Warehouse-Konzept. Architektur-Datenmodelle-Anwendungen*, 2. Aufl., Gabler, Wiesbaden 1997.

WINTER, ROBERT (1998): *Informationsableitung in betrieblichen Anwendungssystemen*, Vieweg, Braunschweig und Wiesbaden 1998.

WINTER, ROBERT (2000): *Zur Positionierung und Weiterentwicklung des Data Warehousing in der betrieblichen Applikationsarchitektur*, in: Jung, Reinhard und Winter, Robert (Hrsg.): *Data Warehousing Strategie: Erfahrungen, Methoden, Visionen*, S. 127–139, Springer, Berlin et al. 2000.

WISOM, BARBARA H. und WATSON, HUGH J. (2001): *An Empirical Investigation of the Factors Affecting Data Warehousing Sucess*, in: MIS Quarterly, 25. Jg., Nr. 1 (2001), S. 17–41.

WISSENSCHAFTLICHE KOMMISSION DER WIRTSCHAFTSINFORMATIK (1994): *Profil der Wirtschaftsinformatik*, in: Wirtschaftsinformatik, 36. Jg., Nr. 1 (1994), S. 80–81.

WITTEN, IAN H. und FRANK, EIBE (2000): *Data Mining - Praktische Werkzeuge und Techniken für das maschinelle Lernen*, Hanser, München und Wien 2000.

WITTMANN, WALDEMAR (1959): *Unternehmung und unvollkommene Information: unternehmerische Voraussicht - Ungewissheit und Planung*, Westdeutscher Verlag, Köln und Opladen 1959.

WOLF, PETER (1999): *Konzept eines TQM-basierten Regelkreismodells für ein Information Quality Management (IQM)*, zugl. Dissertation Universität Dortmund, Verl. Praxiswissen, Dortmund 1999.

YIN, ROBERT K. (1994): *Case study research: design and methods*, 2. Aufl., SAGE Publications, London et al. 1994.

ZACHMAN, JOHN A. (1987): *A Framework for Information Systems Architecture*, in: IBM Systems Journal, 26. Jg., Nr. 3 (1987), S. 454–470.

ZEHNDER, CARL A. (1998): *Informationssysteme und Datenbanken*, 6. Aufl., Teubner, Stuttgart 1998.