Datenqualitätsmanagement

Verfasser: M. Helfert, C. Herrmann, B. Strauch

Lehrstuhl: Prof. Dr. R. Winter

Bericht Nr.: BE HSG/CC DW2/02
Datum: September 2001

Universität St. Gallen –
Hochschule für Wirtschafts-, Rechts- und
Sozialwissenschaften (HSG)

Institut für Wirtschaftsinformatik
Müller-Friedberg-Strasse 8
CH-9000 St. Gallen
Tel.: + 41 (0) 71 224 2934
Fax: + 41 (0) 71 224 2189

Kompetenzzentrum Data Warehousing 2 (CC DW2)

In grösseren Unternehmen existieren eine Vielzahl verschiedener und häufig sehr heterogener Informationssysteme. Neue Problemstellungen und sich dynamisch verändernde Geschäftsmodelle machen es jedoch erforderlich, dass vorhandene Datenquellen auch integriert, das heisst unabhängig von ihrem operativen Einsatzzweck genutzt werden können. Die Zielvorstellungen reichen von einer einheitlichen Kundensicht bis hin zu schnell verfügbaren Führungsinformationen.

Während im technischen Bereich des Data Warehousing in den vergangenen Jahren erhebliche Fortschritte erzielt worden sind, fehlt es noch immer an gesicherten Erkenntnissen in Bereichen wie:

- Applikationsintegration,
- Entwicklung von Strategien zur operativen Nutzung des Data Warehouse,
- Integration des Operational Data Stores (ODS) ins Data Warehouse,
- Metadatenmanagement,
- Datenqualitätsmanagement und
- Datenschutz und –sicherheit beim Data Warehousing.

Im Rahmen des Kompetenzzentrums Data Warehousing 2 (CC DW2) entwickelt das Institut für Wirtschaftsinformatik der Universität St. Gallen (IWI-HSG) zusammen mit namhaften Partnerunternehmen Methoden und Referenzlösungen für die genannten Bereiche. Die folgenden Unternehmen sind Partner im CC DW2:

- ARAG Lebensversicherungs-AG (DE)
- Credit Suisse (CH)
- Mummert + Partner (DE)
- Rentenanstalt/Swiss Life (CH)
- SwissRe (CH)
- UBS AG (CH)
- Winterthur Versicherungen (CH)
- W & W AG (DE)

Durch die entstehenden dedizierten Problemlösungen kann das Data Warehouse als wichtige Komponente langfristig und in wirtschaftlicher Weise in das betriebliche Informationsmanagement integriert werden.

Der vorliegende Arbeitsbericht behandelt das Thema des *Datenqualitätsmanagements in Data-Warehouse-Systemen*.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation und Zielsetzung

Ein wichtiger Erfolgsfaktor für die Nutzung von Data-Warehouse-Systemen[1] ist die Qualität der Daten, denn betriebliche Entscheidungsprozesse basieren auf den zur Verfügung gestellten Informationen. Nach einer Studie aus dem Jahr 2000 wird die Sicherstellung der Datenqualität bei nahezu allen Unternehmen als problematisch eingeschätzt [vgl. Helfert 2000a, S. 13]. Auch die Partnerunternehmen des Kompetenzzentrums Data Warehousing 2 (CC DW2) sehen Handlungsbedarf im Bereich der Datenqualität. Daher wurde das Datenqualitätsmanagement (DQM) zu einem der Themenschwerpunkte des Kompetenzzentrums erklärt.

Eine bewährte und weit verbreitete Methode zur Verbesserung mangelhafter Datenqualität ist das nachträgliche Bereinigen der Daten im Rahmen eines Data-Cleansing-Prozesses. Dieser ist jedoch sehr kostenintensiv und versagt gänzlich, falls Daten nicht vorhanden, widersprüchlich oder falsch sind. Hieraus ergibt sich die Notwendigkeit für ein umfassendes und generell anwendbares Datenqualitätskonzept, welches die geforderte Datenqualität proaktiv sicherstellt, sodass die nachträgliche Datenbereinigung nahezu überflüssig wird.

Neben den bereits oben erwähnten Kosten für das nachträgliche Bereinigen hat schlechte Datenqualität noch weitere negative Konsequenzen zur Folge:

- **Geringe interne Akzeptanz:** Eine unzureichende Qualität der Daten kann z. B. zu einem Vertrauensverlust der Datennutzern bis hin zu einer generellen Verweigerungshaltung führen. Hierdurch kann der erwartete Nutzen des DWH nicht mehr erreicht werden und das DWH selbst wird infrage gestellt.
- **Schlechte Entscheidungsprozesse:** Entscheidungsprozesse basieren i. d. R. auf Analysen und Berichten, die mit Hilfe des DWH erstellt werden. Eine schlechte Datenqualität führt somit zu falschen Auswertungen und nicht aussagekräftigen Ergebnissen, sodass bspw. falsche Investitionsentscheidungen getroffen werden, falsche Kundensegmente ausgewählt werden oder eine mangelhafte Tarif- und Preiskalkulation durchgeführt wird.

[1] Unter einem Data-Warehouse-System werden alle Komponenten der Data-Warehouse-Architektur ausgehend von den operativen System über das Kern-Data-Warehouse bis hin zu den analytischen Informationssystemen zusammengefasst.

- **Unzureichende Unterstützung der operativen Geschäftsprozesse:** Auch in den operativen Geschäftsprozessen werden die Daten des DWH genutzt. Eine unzureichende Qualität dieser Daten kann u. a. Auswirkungen auf die Kundenzufriedenheit und damit auf das Unternehmensimage haben. Beispielsweise können Kunden durch falsch ausgestellte Rechnungen oder durch falsche Informationen zur Abwanderung veranlasst werden.
- **Zusatzaufwand:** Beispielsweise ist bei schlechter Datenqualität eine aufwendige Suche nach den richtigen Werten erforderlich oder durch Doppelerfassungen bzw. durch nachträgliches Erstellen von Analysen und Berichten entsteht zusätzlicher Aufwand.

Die bisher entwickelten Konzepte zur Sicherstellung der Datenqualität für Data-Warehouse-Systeme sind meist auf einer abstrakten Ebene angesiedelt [vgl. z. B. English 1999; Häussler 1998; Jarke et al. 2000]. Methoden zur Operationalisierung und Umsetzung der Verfahren sind i. d. R. nicht gegeben. Aufgrund dieser mangelnden Fundierung und des dringlichen Bedarfs nach anwendbaren Konzepten in der Praxis wird im Rahmen des Kompetenzzentrums Data Warehousing 2 ein Ansatz zur Sicherstellung der Datenqualität beim Data Warehousing entwickelt. Hierbei kann die geforderte Datenqualität mittels eines ganzheitlichen Datenqualitätsmanagements proaktiv sichergestellt werden. Ziel dieser Arbeit ist es, ein Vorgehen zur Implementierung eines Datenqualitätsmanagements zu entwickeln, anhand eines Beispiels zu konkretisieren und zu veranschaulichen sowie mögliche Probleme bei der Umsetzung aufzuzeigen.

1.2 Aufbau der Arbeit

Nach dem einleitenden Kapitel werden zunächst die definitorischen Grundlagen in Kapitel 1 gelegt, die zum weiteren Verständnis dieser Arbeit unabdingbar sind. Aus der allgemeinen Qualitätsdefinition wird zunächst der Datenqualitätsbegriff abgeleitet und dessen unterschiedliche Facetten werden beleuchtet. Des Weiteren werden die verschiedenen Teilbereiche des operativen Qualitätsmanagements näher erläutert und die wichtigsten Prinzipien des Total Quality Management dargestellt.

Nach der Schaffung der begrifflichen Ausgangsbasis wird in Kapitel 3 die aktuelle Unternehmenssituation bzgl. des Datenqualitätsmanagement betrachtet. Es werden kurz zwei konkrete Datenqualitätskonzepte aus der Praxis skizziert. Nachfolgend werden anhand einer empirischen Untersuchung Hindernisse für das Datenqualitätsmanagement und Massnahmen

zur Förderung eines solchen Konzepts aus Sicht der Unternehmenspraxis erörtert und priorisiert.

Im Anschluss daran wird in Kapitel 4 ein Datenqualitätskonzept für Data-Warehouse-Systeme entwickelt. Zunächst erfolgt eine konzeptionelle theoretische Darstellung eines ganzheitlichen Datenqualitätsmodells, welches im darauffolgenden Unterkapitel anhand eines konkreten Beispiels eingehend erläutert, detailliert und verfeinert wird. Abschliessend wird die Abarbeitungsreihenfolge der einzelnen, durchzuführenden Aufgaben anhand eines phasenorientierten Vorgehensmodells konkretisiert. Das Schlusskapitel fasst den Arbeitsbericht zusammen und zeigt den weiteren Forschungsbedarf auf.

2 Definitorische Grundlagen

Das folgende Kapitel soll grundlegende Begriffe klären und die für diesen Arbeitsbericht gültigen Definitionen liefern. Zum einen soll der allgemeine Qualitätsbegriff und dessen unterschiedliche Ausprägungen diskutiert werden. Zum anderen soll aus diesem allgemeinen Verständnis heraus der Begriff der Datenqualität abgeleitet und konkretisiert werden. Des Weiteren sollen die Bestandteile des Qualitätsmanagements skizziert und der operative Teil eingehender erläutert werden. Abschliessend werden die einzelnen Prinzipien des Total Quality Management dargestellt.

2.1 Qualität

Als Folge der Entwicklung des Qualitätsbegriffs und der damit verbundenen Diskussion existiert eine Vielzahl von Definitions- und Interpretationsversuchen des komplexen und schwer zu beschreibenden Begriffs der Qualität. Ziel der Begriffsbestimmungen ist es, die Komplexität des Qualitätsphänomens zu reduzieren und zu operationalen Aussagen zu gelangen. Nach dem Systematisierungsansatz von Garvin lassen sich fünf Qualitäts-vorstellungen unterscheiden [vgl. Garvin 1998, S. 40].

- **Produktbezogener Ansatz:** Produkteigenschaften bestimmen bei der produktorientierten Qualitätsauffassung die Qualität, sodass Qualität präzise messbar und eine inhärente Eigenschaft des Produktes selbst darstellt. Danach spiegeln Qualitätsdifferenzen Unterschiede in der vorhandenen, beobachtbaren Quantität bestimmter Eigenschafts-ausprägungen wieder.
- **Anwenderbezogener Ansatz:** Beim anwenderbezogenen Ansatz liegt die Auffassung vor, dass Qualität durch den Produktbenutzer und nicht ausschliesslich durch das Produkt selbst festgelegt wird. Ein Produkt wird dann als qualitativ hochstehend angesehen, wenn es dem Zweck der Benutzung durch den Kunden während des Gebrauchs dient, wobei sich der Zweck der Benutzung nach den individuellen Bedürfnissen des Kunden richtet.
- **Prozessbezogener Ansatz:** Nach diesem Ansatz bedeutet Qualität die Einhaltung von Spezifikationen und die Abwesenheit von Fehlern. Im Mittelpunkt dieses Ansatzes stehen die auf Einhaltung der Spezifikation kontrollierte Produktionsprozesse. Jede Abweichung von der Spezifikation bedeutet Verringerung der Qualität.

- **Wertbezogener Ansatz:** Bei diesem Ansatz wird ein Bezug zwischen Preis und Qualität im Sinne von Nutzen hergestellt. Ein Produkt ist dann von hoher Qualität, wenn der zu entrichtende Preis und die empfangene Leistung in einem akzeptablen Verhältnis stehen.
- **Transzendenter Ansatz:** Der transzendente Ansatz kennzeichnet Qualität als angeborene Vortrefflichkeit, Einzigartigkeit oder Superlative, als ein Synonym für hohe Standards und Ansprüche. Qualität wird zu einer absoluten und universell erkennbaren Eigenschaft. Die diesem eher abstrakt philosophischen Verständnis folgende Auffassung, Qualität könne nicht exakt definiert werden, sondern sei nur erfahrbar [vgl. Garvin 1988, S. 41], ist jedoch für die weitere Betrachtungen nicht ausreichend und soll daher nicht weiter verfolgt werden.

Neben diesen beschreibenden Ansätzen gibt es eine Reihe von Versuchen, den Begriff der Qualität umfassend zu definieren [vgl. Wallmüller 1990, S. 8]. Exemplarisch sei die Definition nach DIN 55 350 genannt: „Qualität ist die Gesamtheit von Eigenschaften und Merkmalen eines Produktes oder einer Tätigkeit, die sich auf deren Eignung zur Erfüllung gegebener Erfordernisse bezieht."

2.2 Datenqualität

Es gilt nun aus dieser obigen allgemeinen Qualitätsdefinition den Begriff der Datenqualität abzuleiten. Wie auch beim allgemeinen Qualitätsbegriff, wird der Begriff der Datenqualität in der Literatur in sehr unterschiedlichen Sichtweisen und mit verschiedenen Qualitätsmerkmalen beschrieben [vgl. Müller 2000, S. 15; Wang, Storey, Firth 1995]. Die Untersuchung der verschiedenen Ansätze zeigt übereinstimmend, dass diesen der anwenderorientierte Qualitätsbegriff (siehe Kapitel 2.1) zugrunde liegt. Der Datennutzer bestimmt demnach aus dem Anwendungskontext heraus das erforderliche Qualitätsniveau, sodass der Zweck bzw. das Ziel der Datennutzung erreicht wird.

Die verschiedenen Qualitätsmodelle basieren meist auf Datenqualitätskriterien, die beispielhaft in alphabetischer Ordnung in Tab. 2-1 genannt werden. Mit Hilfe dieser Kriterien soll der abstrakte Begriff der Datenqualität konkretisiert und operationalisiert werden. Meist werden diese Qualitätskriterien empirisch erfasst [vgl. Wang, Strong 1996] oder auf Basis der in der Literatur genannten Kriterien zusammengestellt [vgl. Naumann, Rolker 1999] und anschliessend in einem Datenqualitätsmodell strukturiert. Ein wesentlicher Aspekt eines Qualitätsmodells ist die Zerlegungssystematik von Qualitätseigenschaften [vgl. Wallmüller 1990, S. 46]. Der allgemeine Qualitätsbegriff, der durch Qualitätseigenschaften

charakterisiert ist, wird durch ableiten von Qualitätsmerkmalen weiter operationalisiert und durch festlegen von Indikatoren, den sogenannten Kenngrössen, bewertbar gemacht.

Aktualität	Allgemeingültigkeit	Alter	Änderungshäufigkeit	Aufbereitungsgrad
Bedeutung	Benutzbarkeit	Bestätigungsgrad	Bestimmtheit	Detailliertheit
Effizienz	Eindeutigkeit	Fehlerfreiheit	Flexibilität	Ganzheit
Geltungsdauer	Genauigkeit	Glaubwürdigkeit	Gültigkeit	Handhabbarkeit
Integrität	Informationsgrad	Klarheit	Kompaktheit	Kompression
Konsistenz	Konstanz	Korrektheit	Neutralität	Objektivität
Operationalität	Performance	Portabilität	Präzision	Problemadäquatheit
Prognosegehalt	Prüfbarkeit	Quantifizierbarkeit	Rechtzeitigkeit	Relevanz
Reliabilität	Richtigkeit	Robustheit	Seltenheit	Sicherheit
Signifikanz	Speicherbedarf	Standardisierungsgrad	Subjektadäquatheit	Testbarkeit
Umfang	Unabhängigkeit	Überprüfbarkeit	Übertragbarkeit	Validität
Verdichtungsgrad	Verfügbarkeit	Verfügungsmacht	Verknüpfbarkeit	Verlässlichkeit
Verschlüsselungsgrad	Verständlichkeit	Vertrauenswürdigkeit	Verwendungsbereitschaft	Vollständigkeit
Wahrheitsgehalt	Wahrscheinlichkeit	Wartungsfreundlichkeit	Wiederverwendbarkeit	Wirkungsdauer
Zeitadäquanz	Zeitbezug	Zeitoptimal	Zugänglichkeit	Zuverlässigkeit

Tab. 2-1: Häufig genannte Datenqualitätskriterien

Huang et al. strukturieren aufgrund empirischer Studien die Datenqualitätskriterien anhand der in Tab. 2-2 genannten Kategorien [vgl. Huang et al. 1999, S. 43]. Die Kategorie der inneren Datenqualität fasst Qualitätsmerkmale zusammen, die Daten aus ihrer Art heraus besitzen. In der Kategorie der kontextabhängigen Datenqualität werden Merkmale der Daten klassifiziert, die aus dem Anwendungszusammenhang resultieren. In der Kategorie der Darstellungsqualität werden Qualitätsmerkmale zusammengefasst, die aus der Darstellung der Daten entstehen. Zugriffsmöglichkeit und Sicherheit werden in der Kategorie Zugangsqualität zusammengefasst.

Kategorie	**Merkmale von Datenqualität**
Innere Datenqualität	Genauigkeit, Objektivität, Glaubwürdigkeit, Vertrauenswürdigkeit
Kontextabhängige Datenqualität	Relevanz, Vollständigkeit, Anwendungsbezug und zeitlicher Bezug, Informationsgehalt
Darstellungsqualität	Interpretierbarkeit, Widerspruchsfreiheit, knappe Darstellung, Verständlichkeit
Zugangsqualität	Zugriffsmöglichkeit, Sicherheit

Tab. 2-2: Datenqualitätskategorien nach Huang et al. [vgl. Huang et al. 1999, S. 43]

Als weiteres Beispiel ist das im Rahmen des europäischen Forschungsprogramms Data Warehouse Quality (DWQ) entwickelte hierarchische Datenqualitätsmodell in Abb. 2-1 dargestellt. Hierbei wird der Begriff der Datenqualität mittels der Kriterien Interpretierbarkeit, Nützlichkeit, Zugriffsmöglichkeit und Glaubwürdigkeit charakterisiert und anhand von weiteren Unterkriterien verfeinert.

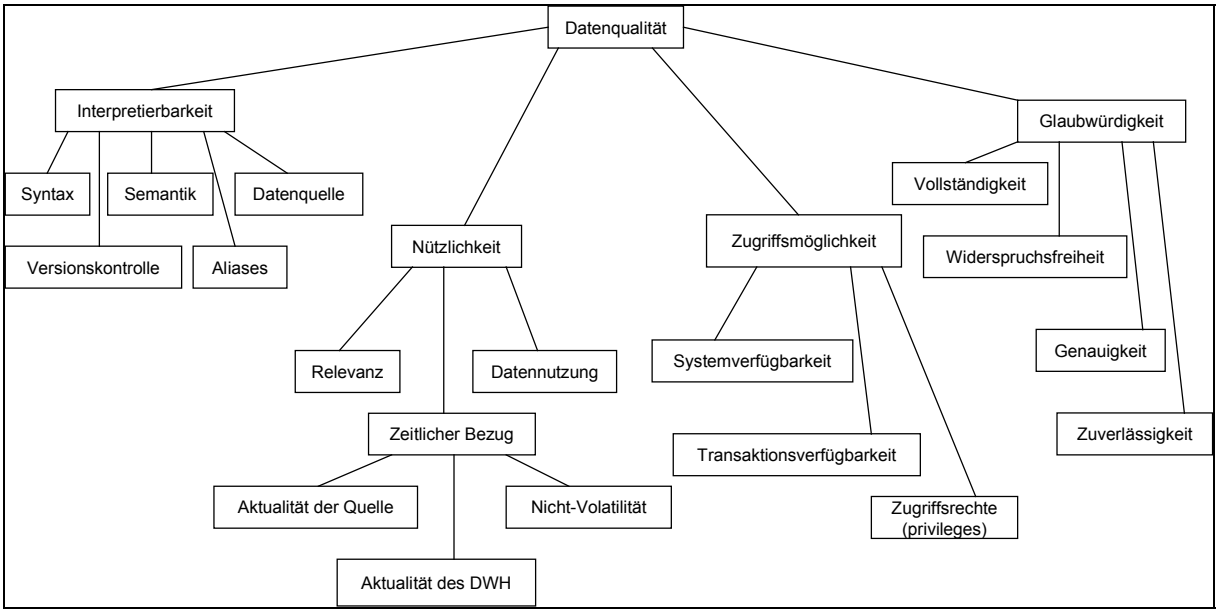

Abb. 2-1: Hierarchisches Datenqualitätsmodell nach Jarke und Vassiliou [Jarke, Vassiliou 1997]

2.3 Operatives Qualitätsmanagement

Nach DIN ISO 8402 umfasst Qualitätsmanagement alle Tätigkeiten der Gesamtführungsaufgabe, welche die Qualitätspolitik, die Qualitätsziele und die Verantwortungen für die Qualität festlegt [vgl. DIN 1995]. Die Elemente lassen sich grob, wie in Abb. 2-2 dargestellt, anhand des St. Galler Managementkonzepts [vgl. Bleicher 1992] einordnen.

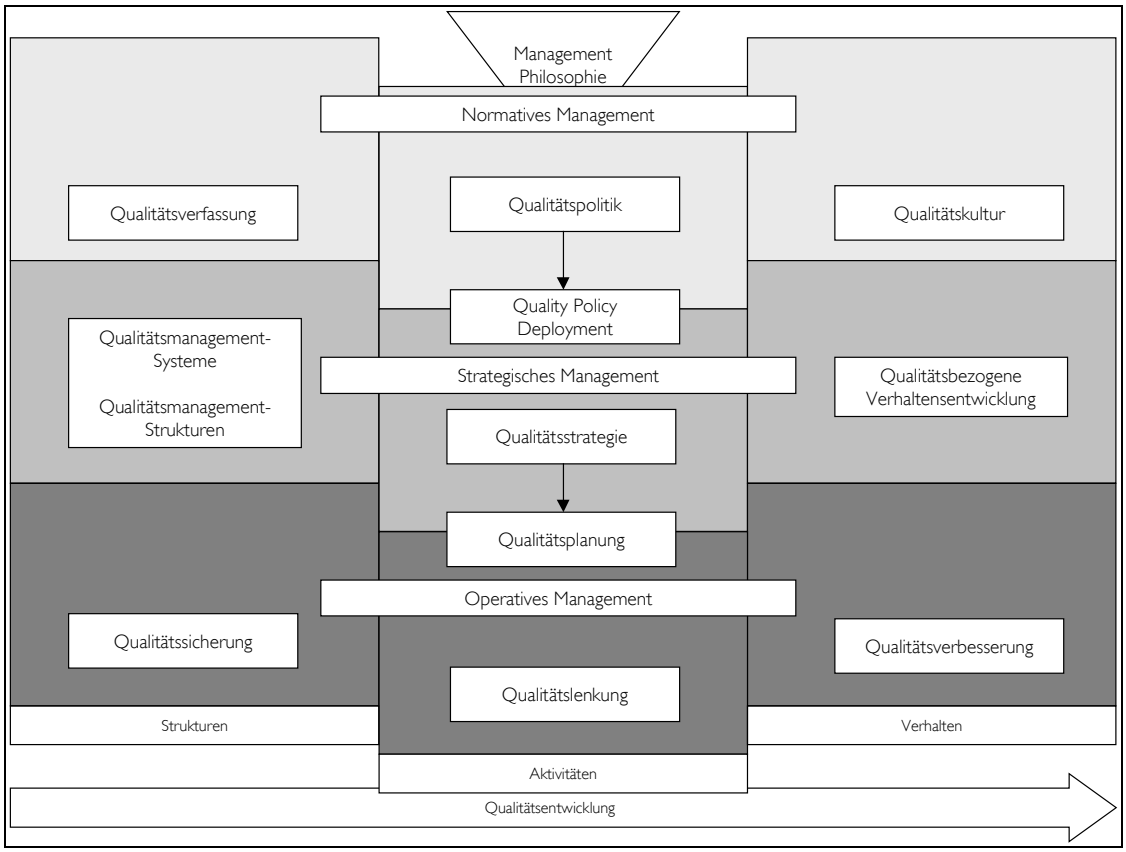

Abb. 2-2: Integrationsrahmen für ein ganzheitliches Datenqualitätsmanagement [vgl. Seghezzi 1996, S. 48]

Das Qualitätsmanagement wird in die drei Ebenen des normativen, strategischen und operativen Managements untergliedert. Die Visionen der Unternehmensführung sind auf der obersten Ebene angesiedelt. Diese werden durch Missionen auf der strategischen Stufe repräsentiert und deren Umsetzung erfolgt im operativen Qualitätsmanagement. Die mittlere Säule stellt die Aktivitäten dar, die sowohl durch die Strukturen unterstützt als auch durch das Verhalten der Führungskräfte und Mitarbeiter geprägt wird. Die dritte Dimension betrifft den zeitlichen Aspekt der Entwicklung der Qualitätsfähigkeit [vgl. Seghezzi 1996, S. 48ff.]. Die zur Erreichung von Qualität notwendigen Aktivitäten sind auf der operativen Ebene zu finden und werden daher im Folgenden eingehender betrachtet. Seghezzi ordnet die operativen Funktionsbereiche in den prozessorientierten Qualitätsansatz von Deming ein, wie Abb. 2-3 verdeutlicht.

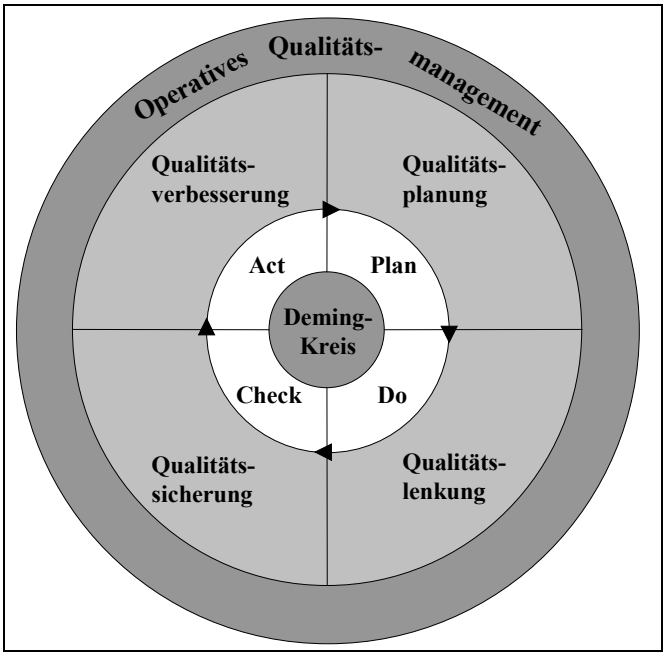

Abb. 2-3: Operatives Qualitätsmanagement nach dem Deming-Kreis [Seghezzi 1996, S. 53]

Das operative Qualitätsmanagement ist anhand der Prozesse auszurichten, um Probleme an den Prozessschnittstellen zu vermeiden. Die von Deming entwickelte Technik zur Prozessverbesserung umfasst die folgenden vier Schritte [vgl. English 1999, S. 42f.]:

- **Plan:** Diese Phase ist gleichzusetzen mit der **Qualitätsplanung**. Aufgabe ist es, Bedürfnisse und Erwartungen zu erfassen, diese in Vorgaben zu transformieren und Leistungen sowie Prozesse zu gestalten [vgl. Seghezzi 1996, S. 72]. Im Rahmen der Qualitätsplanung werden Qualitätsanforderungen an die Prozesse festgelegt. Es sind dafür Qualitätsmerkmale auszuwählen, zu klassifizieren und mit Gewichten zu versehen [vgl. Wallmüller 1990, S. 19].
- **Do:** Das Äquivalent hierzu ist die **Qualitätslenkung**, welche auf die Einhaltung von Spezifikationen und die Beherrschung der Prozesse abzielt [vgl. Seghezzi 1996, S. 76]. Hierfür sind zunächst geeignete Prozesse zu identifizieren und Massnahmen zum Erreichen der Prozesskonformität zu ergreifen. Produkt- und Prozessqualität müssen im Rahmen der Qualitätslenkung gemessen und in quantitativen Kennziffern ausgedrückt werden. Ein wichtiges Hilfsmittel für die Qualitätslenkung sind Qualitätsprüfungen [vgl. Wallmüller 1990, S. 19]. Letztlich sind Verantwortlichkeiten für die Qualitätslenkung festzulegen und die Messergebnisse als Rückkopplung in Regelkreisen zurückzuführen.
- **Check:** Dieser Schritt, auch als **Qualitätssicherung** bezeichnet, ist als strukturelle Unterstützung der Qualitätsplanung und Qualitätslenkung zu verstehen, der darauf abzielt, Risiken systematisch zu erkennen, aufzudecken und ihre Wirkung zu bekämpfen

[vgl. Seghezzi 1996, S. 108]. Voraussetzung der Qualitätssicherung sind Risikoanalysen, wie beispielsweise die der Fehlermöglichkeits- und -einflussanalyse (FMEA) [vgl. Seghezzi 1996, S. 99].

- **Act:** Die vierte Phase entspricht der kontinuierlichen Verbesserung (**Qualitätsverbesserung**) des operativen Qualitätsmanagements [vgl. Seghezzi 1996, S. 111]. 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.

Für die weitere Arbeit ist insbesondere die Qualitätsplanung und Qualitätslenkung von Interesse. Diese zwei Funktionen stellen die zentralen Aufgaben des zu entwickelnden Datenqualitätsmanagements dar [vgl. English 1999, S. 70ff.; Huang et al. 1999, S. 16]. Im Folgenden wird mit dem Begriff des Datenqualitätsmanagements demnach ausschliesslich der operative Teil bezeichnet, da dieser den für die Arbeit relevanten Bereich darstellt. Eine eingehendere Betrachtung der normativen und strategischen Ebene findet nicht statt.

2.4 Total Quality Management

Zunächst bietet sich eine kurze historische Betrachtung des Qualitätswesens in der Produktion an, da hierdurch der Wandel des Qualitätsgedankens deutlich wird. Ausgehend von einer Qualitätskontrolle im Sinne einer Produktendkontrolle rückte die Betrachtung des Produktionsprozesse in den Mittelpunkt. Im Rahmen einer ganzheitlichen Qualitätssicherung wird das Ziel verfolgt, fehlerfreie Prozesse zu etablieren und dadurch Qualität zu produzieren, statt nachträgliche Qualitätsverbesserungen am Produkt durchführen zu müssen.

Insbesondere amerikanische Qualitätsexperten wie Deming, Juran und Feigenbaum trugen massgeblich dazu bei, dass vor allem in Japan umfassendere Qualitätskonzepte weiterentwickelt wurden. Charakteristische Elemente dieser Entwicklung waren neben der konsequenten Anwendung statistischer Methoden, eine verstärkte Kundenorientierung, eine funktionsübergreifende innerbetriebliche Zusammenarbeit (Qualität als unternehmensweite Aufgabe), die Initiierung kontinuierlicher Qualitätsverbesserungsmassnahmen [vgl. Imai 1993, S. 21ff.], eine verstärkte Integration des Menschen als relevante Grösse (Organisationskonzept der „Quality Circles“), eine präventiv orientierte Qualitätssicherung sowie die Integration der Qualitätsziele in die Unternehmenspolitik.

Aus diesen Prinzipien entwickelte sich das Total Quality Management (TQM). Hierunter wird ein ganzheitlicher Denk- und Handlungsansatz zur Schaffung einer Qualitätskultur im Unternehmen verstanden. Drei wesentliche Grundprinzipien bilden den Kern des TQM [vgl. Schnauber et al. 1997, S. 9]:

- **Kundenorientierung:** Alle Aktivitäten des Qualitätsmanagements müssen sich am Kunden bzw. an dessen Zufriedenheit ausrichten, um eine dauerhafte Kundenbeziehungen zu etablieren. Weitere Anspruchsgruppen, die es zu berücksichtigen gilt, sind die Mitarbeiter, die Eigentümer bzw. Shareholder, die Lieferanten sowie das Unternehmen selbst. Die verschiedenen Ansprüche, Ziele und Zielkonflikte der Stakeholder müssen priorisiert und einvernehmlich gelöst werden.
- **Prozessorientierung:** Hierbei erfolgt eine Optimierung der Fertigungsprozesse durch eine kontinuierliche Verbesserung. Das Ziel ist eine hohe Qualität aller Prozessschritte und damit die Sicherstellung der Wirtschaftlichkeit. Jede Funktion innerhalb des Prozesses muss die qualitativen Anforderungen aller nachgelagerten Funktionen erfüllen. Erreicht werden kann dies durch den Abbau von Schnittstellen bzw. von Redundanzen, durch die Einführung prozessorientierter Organisationskonzepte sowie durch die Erweiterung der Arbeitsinhalte bzw. Verantwortungsbereiche der Mitarbeiter.
- **Mitarbeiterorientierung:** Die Einbeziehung und Beteiligung aller Mitarbeiter des Unternehmens wird als grundlegende Voraussetzung zur Umsetzung der Unternehmensziele angesehen. Dies bezieht sich einerseits auf die Qualität der Personalführung andererseits auf die Motivation der Mitarbeiter. Das Prinzip der internen Kunden-Lieferanten-Beziehung ist als wesentlicher Baustein anzusehen. Hierbei werden nachgelagerte Phasen im Prozess immer als Kunde der vorangegangen Phase (Lieferant) angesehen und dementsprechend behandelt [vgl. Töpfer, Mehdorn 1994, S. 23].

Das TQM kann in das operative Datenqualitätsmanagement eingeordnet werden und spezifiziert Aufgabenkomplexe für die einzelnen operativen Funktionen. So kann bspw. die kontinuierliche Prozessoptimierung der Qualitätsverbesserung zugeordnet werden und die Etablierung von prozessorientierten Organisationsstrukturen gehört dem Bereich der Qualitätssicherung an. Dies zeigt, dass das TQM Prinzipien bereitstellt, die den Strukturen des operativen Datenqualitätsmanagements zugeordnet werden können. Diese Ansätze können auf die Aspekte der Datenqualität übertragen werden und so ein umfassendes Datenqualitätsmanagement begründen [vgl. English 1999, S. 52-66].

3 Datenqualitätsmanagement in den Unternehmen

In den folgenden drei Unterkapiteln soll das Datenqualitätsmanagement in der Praxis näher beleuchtet werden. Der Fokus liegt dabei einerseits auf den konkreten Umsetzungen in den Unternehmen selbst. Hierbei sollen Beispiele für den Stand der Technik im Datenqualitätsmanagement präsentiert werden. Andererseits sollen sowohl Hindernisse, die dem Datenqualitätsgedanken entgegenstehen, als auch Massnahmen bzw. Voraussetzungen, die eine schnellere Verbreitung des Datenqualitätsmanagements in den Unternehmen unterstützen, auf Basis einer empirischen Untersuchung betrachtet werden. Die Bewertung erfolgt dabei durch Vertreter der Praxis und repräsentiert somit subsidiäre als auch prohibitive Einflussfaktoren in den Unternehmen.

3.1 Beispiele für Datenqualitätsmanagement in der Praxis

Im Rahmen des zweiten CC DW2 Workshops präsentierten die Vertreter der Partnerunternehmen ihre Ansätze im Bereich des Datenqualitätsmanagements. Hierbei kristallisierte sich heraus, dass Datenqualität ein relevanter Themenschwerpunkt für die Unternehmenspraxis darstellt und dass zahlreiche Anstrengungen unternommen werden, um Verfahren und Ansätze für das Datenqualitätsmanagement zu entwickeln. Beispielhaft sollen im Folgenden zwei innovative Methoden der Unternehmen SwissRe und die Credit Suisse beschrieben werden.[1]

Ausgehend von Datenqualitätsanforderungen[2] und der zugrundeliegenden Data-Warehouse-Architektur konnte die SwissRe verschiedene, die Datenqualitätsproblematik betreffende Problemgruppen identifizieren.

- Fehlerhafte Komponenten und Methoden (insb. die Datentransformation) im ETL-Prozess
- Heterogenität der operativen Systeme bzgl. der Datenqualität
- Neue Kennzahlen für neue Geschäftsbereiche
- Veränderte Organisationsstrukturen im Managementbereich
- Wachsendes Bewusstsein der Anwender für die Datenqualitätsproblematik

---

1 Als Quellen für die folgenden Abschnitte wurden die Präsentationen der Unternehmen SwissRe und Credit Suisse auf dem zweiten CC DW2 Workshop herangezogen.

2 Die SwissRe kategorisiert die Anforderungen nach dem Schema von Huang (vgl. Kapitel 1) in innere Datenqualität, kontextabhängige Datenqualität, Darstellungsqualität und Zugangsqualität.

Die Lösungsansätze können differenziert werden nach Datenkorrekturen, Methoden und Werkzeugen zur Verbesserung bzw. zur Sicherung der Datenqualität. Bei den Datenkorrekturen handelt es sich hauptsächlich um Data-Cleansing-Massnahmen, die bei auftretenden Fehlern angewendet werden (reaktiv). Im Bereich der Werkzeuge hat die SwissRe ein Tool entwickelt, mit dessen Hilfe der Datennutzer zu einem Datum Datenqualitätsanmerkungen in Form von Metadaten erstellen kann. Bei der Aggregation bzw. Transformation der Daten werden diese Anmerkungen „weitervererbt“ und stehen somit auf jeder Granularitätsstufe zur Verfügung. Hierdurch kann der Anwender die Datenqualität bereits vor der Verwendung der Daten abschätzen. Unter Methoden zur Qualitätssicherung fasst die SwissRe Testfälle, Regeln und Test Workshops zusammen. Der kritische Erfolgsfaktor dieses Konzepts liegt in der bisher noch nicht realisierten Quantifizierung der Qualität, d. h. eine Interpretierbarkeit der Ergebnisse über die vorliegende Datenqualität ist noch nicht möglich.

Einen auf Metadaten basierenden Ansatz verfolgt die Credit Suisse. Ausgangspunkt ist die Frage der Datennutzer nach der Verwendbarkeit der Daten. Es wird versucht, diese Fragestellung anhand der Dimensionen Aktualität, Repräsentativität und Fehlerauswirkung zu beantworten. Die hierbei verwendeten Metadaten sind zum einen die Scheduler-Ergebnisse über die Job-Abläufe und zum anderen Daten über den Ladevorgang. Das Scheduler-Protokoll gibt an, welche Ladeprozesse erfolgreich durchgeführt wurden, d. h. es ist bekannt, aus welchen operativen Systemen Daten in das Data Warehouse geladen wurden. Dadurch ist es möglich, Aussagen über die Aktualität der Daten abzuleiten. Wurde ein Ladevorgang durchgeführt, so sind die Daten des Data Warehouse auf einem aktuellen Stand. Hingegen ist die Aktualität der Daten beeinträchtigt, falls ein Ladevorgang nicht ausgeführt wurde. Neben diesen Informationen stehen noch Metadaten über den Anteil der in das Data Warehouse erfolgreich geladenen Daten bzgl. der Gesamtmenge aller in das Data Warehouse zu überführenden Daten zur Verfügung. Zusammengenommen kann die Datenqualität durch obige Metadaten beschrieben werden, jedoch kann eine genaue Einschätzung der vorliegenden Qualität bisher nur durch einen Experten vorgenommen werden, d. h. eine Verdichtung der Qualitätsdaten zu allgemein verständlichen Kennzahlen ist bislang nicht möglich.

Beide Beispiele zeigen exemplarisch die in der Praxis bereits eingesetzten Methoden und Verfahren im Bereich des Datenqualitätsmanagements. Es wird verdeutlicht, dass Verfahrensweisen zur Messung und Bewertung der Datenqualität existieren. Der vom CC

DW2 propagierte Ansatz kann als Ergänzung angesehen werden. Dieser fasst isolierte Messansätze und Techniken in einem integrierten Konzept zusammen.

3.2 Untersuchung betrieblicher Hindernisse

Im Rahmen des zweiten CC DW2 Workshop wurden mögliche Hindernisse für das Datenqualitätsmanagement im betrieblichen Umfeld diskutiert. In einem ersten Schritt wurden von den Vertretern der Partnerunternehmen und den Mitarbeitern des IWI-HSG mögliche Gründe und Ursachen identifiziert, die die Einführung eines Datenqualitäts-managements im Unternehmen be- bzw. verhindern. Dann erfolgte eine Priorisierung der Hindernisse nach deren Relevanz im Unternehmen. Im Folgenden werden die Ergebnisse dargestellt.

Sechs wesentliche betriebliche Hindernisse kristallisierten sich als Hauptursachen heraus. Die Stimmverteilung auf die einzelnen Problemkategorien zeigt Abb. 3-1:

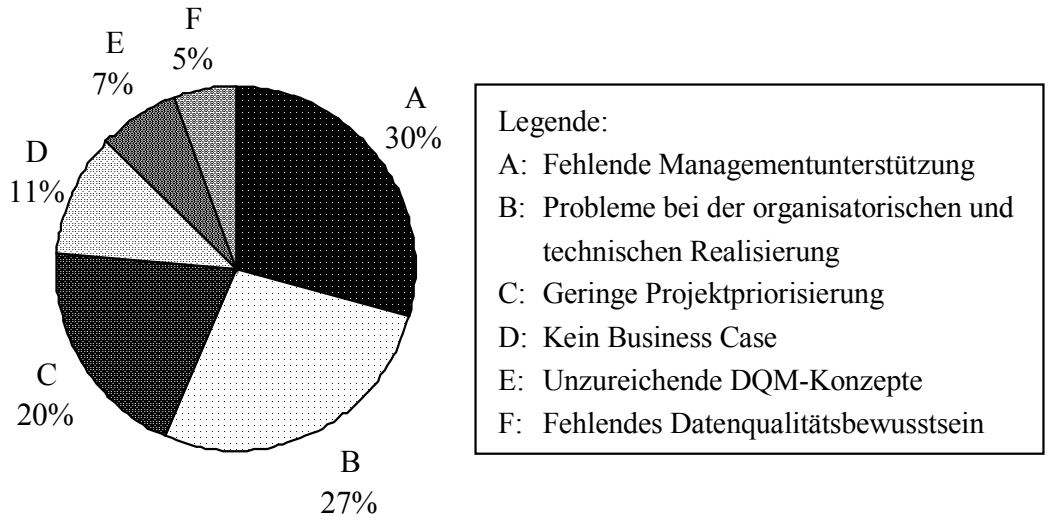

Abb. 3-1: Betriebliche Hindernisse

- **Fehlende Managementunterstützung:** Als wichtigster Aspekt, der der Implementierung eines Qualitätsmanagements entgegensteht, wird die fehlende Sensibilisierung bzw. das fehlende Verständnis für die Datenqualität und für deren Auswirkungen genannt. Diese Ursache bezieht sich insbesondere auf das verantwortliche Management, d. h. für die Einführung eines Datenqualitätsmanagements ist i. d. R. kein geeigneter Sponsor im Unternehmen vorhanden. Insgesamt entfallen 30% aller Stimmen auf diesen Punkt.

- **Probleme bei der organisatorischen und technischen Realisierung:** Ein zweites wesentliches Hindernis, das 27% der Stimmen auf sich vereint, ist der hohe technische und administrative Aufwand für den Aufbau und die Wartung eines gut funktionierenden Datenqualitätsmanagements. Eng verknüpft hiermit ist der organisatorische Aspekt der fehlenden bzw. unklaren Zuständigkeiten und Weisungsbefugnisse. Die Datennutzer gehören i. d. R. anderen Verantwortungsbereichen an als die Datenerzeuger bzw. -lieferanten, d. h. die einzelnen Applikationen des Data-Warehouse-Systems sind jeweils verschiedenen Abteilungen zugeordnet. So sind bspw. die operativen Systeme den jeweiligen Fachabteilungen zugeordnet, wohingegen die analytischen Informationssysteme z. B. dem Controlling-Bereich zuzurechnen sind. Ein umfassendes Datenqualitätsmanagement tangiert demnach zahlreiche Einflussbereiche, sodass Konflikte erwartet werden können.
- **Geringe Projektpriorisierung:** Das folgende Hindernis bezieht sich insbesondere auf den Stellenwert des Qualitätsmanagements in Data-Warehouse-Projekten. Hierbei ist zu beobachten, dass dem Datenqualitätsmanagement durch eine ungenügende Projektplanung lediglich eine niedrige Priorität eingeräumt wird. Diese nur geringe Berücksichtigung des Qualitätsgedankens wird weiter verstärkt durch den hohen Termin- und Kostendruck in Softwareentwicklungsprojekten. Mit 20% der Stimmen steht dieser Aspekt an dritter Stelle.
- **Kein Business Case:** Ein erhebliches Problem stellt die Rechtfertigung des Datenqualitätsmanagements gegenüber hierarchisch höheren Instanzen dar. Hierfür wird i. d. R. eine Wirtschaftlichkeitsanalyse gefordert. Da sich jedoch die Kosten, die aufgrund von schlechter Datenqualität anfallen, bzw. der Nutzen, der aus einem höheren Qualitätsniveau resultiert, nicht oder nur schwer quantifizieren lassen, ist eine derartige Wirtschaftlichkeitsrechnung nicht zu leisten.
- **Unzureichende DQM-Konzepte:** Ein weiterer Hinderungsrund stellt die fehlende bzw. noch nicht ausreichende Operationalisierung des Datenqualitätsmanagements dar. Pragmatische Ansätze zur Messung der Datenqualität sind ebenso wenig vorhanden wie ein systematisches Vorgehen zur Realisierung und Einführung des Datenqualitätsmanagements.
- **Fehlendes Datenqualitätsbewusstsein:** Der letztgenannte Aspekt betrifft die Datenqualität in den operativen Systeme. Diese ist erfahrungsgemäss nur schwer beeinflussbar und es hat i. d. R. eine Anspruchsniveauanpassung an die vorliegende Datenqualität stattgefunden. Daher ist auf dieser Ebene des Data-Warehouse-Systems

kein akuter Leidensdruck bzgl. der Datenqualität vorhanden. Auf der Ebene der analytischen Informationssysteme jedoch bestehen i. d. R. andere und qualitativ höhere Anforderungen an die Qualität der Daten. Diese unterschiedlichen Qualitätsanforderungen und die fehlende Motivation zur Schaffung einer höheren Datenqualität auf operativer Ebene stellen einen weiteren Hindernisgrund dar.

3.3 Untersuchung notwendiger Massnahmen

Im Zuge der Ursachenanalyse wurden auch notwendige Massnahmen zur Überwindung der Hindernisse bzw. zur Förderung des Datenqualitätsmanagements ermittelt. Das Erhebungsdesign entspricht dem des vorangegangenen Abschnitts.

Es konnten sechs mögliche Massnahmen identifiziert werden, deren Stimmgewichtung in Abb. 3-2 dargestellt ist. Teilweise sind dies gerade die Antonyme zu den bereits oben erläuterten Hindernissen. Im Folgenden sollen die einzelnen Aspekte kurz beleuchtet werden:

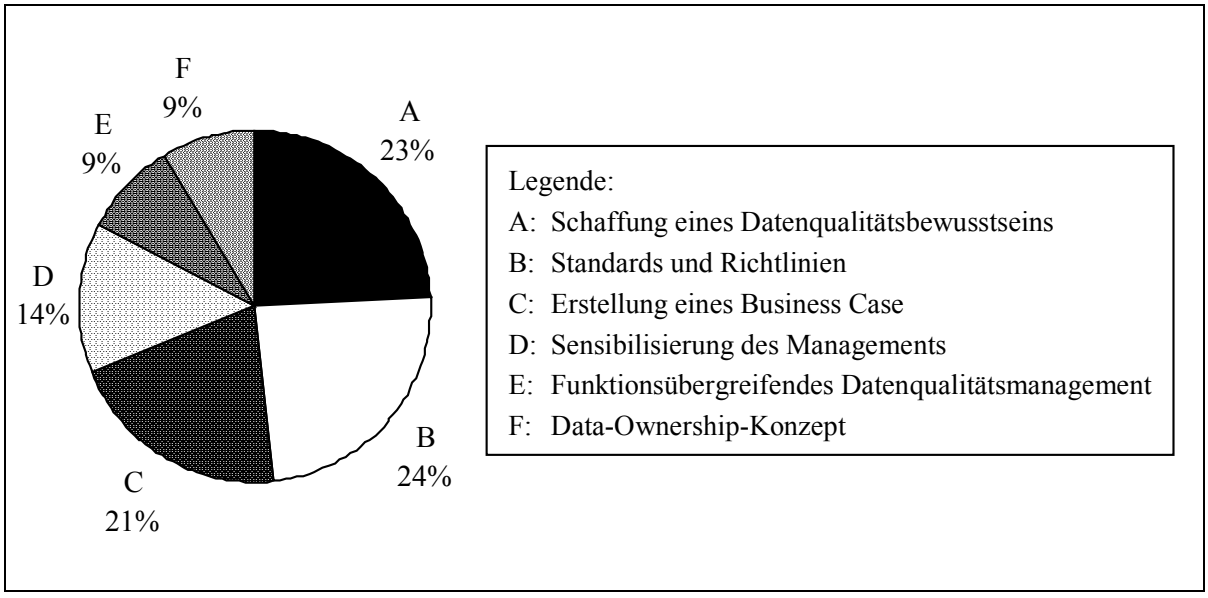

Abb. 3-2: Notwendige Massnahmen

- **Schaffung eines Datenqualitätsbewusstseins:** Eine notwendige Massnahme zur Unterstützung eines Datenqualitätsmanagements ist die Sensibilisierung aller Mitarbeiter, insbesondere der Datenerzeuger bzw. der Fachbereiche, für diesen Problembereich. Konkrete Massnahmen können bspw. die Einführung von Incentives für hohe Datenqualität sein oder die Aufnahme von Datenqualitätszielen in die Projektplanung. Auch durch eine Offenlegung von tatsächlichen Qualitätsproblemen und deren

Auswirkungen kann ein stärkeres Bewusstsein für die Problematik geschaffen werden. Insgesamt entfallen 24% der abgegebenen Stimmen auf diesen Massnahmenkomplex.

- **Standards und Richtlinien:** Das Datenqualitätsmanagement im Unternehmen kann durch festgelegte Standards und Richtlinien in den Bereichen der Qualitätsmessung und der organisatorischen Durchführung bzw. Einbettung gefördert werden. Hiermit kann eine systematische Herangehensweise an das Datenqualitätsmanagement erreicht werden und die Durchsetzbarkeit gegen bestehende Strukturen wird vereinfacht. Dieser Aspekt wird von 24% der Teilnehmer priorisiert.
- **Erstellung eines Business Case:** Durch eine Kosten-/Nutzenabschätzung des Datenqualitätsmanagements kann dessen Notwendigkeit besser belegt und damit eine Einführung erleichtert sowie unterstützt werden. Ziel muss es sein, eine transparente Wirtschaftlichkeitsbetrachtung zu erreichen bspw. durch die Berechnung geeigneter Kennzahlen oder durch die Formulierung eines Business Case. Die Wichtigkeit dieser Massnahme wird mit 22% bewertet.
- **Sensibilisierung des Managements:** Der vierte Massnahmenbereich betrifft die Unternehmensführung bzw. das obere Management. Diese müssen ähnlich wie beim Total Quality Management eine Datenqualitätskultur im Unternehmen schaffen und die Datenqualität in den Unternehmenszielen verankern. Insbesondere wird vom Management gefordert eine derartige Kultur „vorzuleben“ und somit das Datenqualitäts-management top-down einzuführen. Auch können Betriebs- und Rollenkonzepte eine derartiges Vorgehen unterstützen.
- **Funktionsübergreifendes Datenqualitätsmanagement:** Das Datenqualitätsmanagement im Unternehmen sollte als umfassendes Konzept verstanden werden. Ein lediglich auf einzelne Bereiche beschränktes Qualitätsmanagement ist zu vermeiden, stattdessen ist eine Einbindung aller betroffenen Abteilungen notwendig, um eine erfolgreiche Durchführung sicherzustellen.
- **Data-Ownership-Konzept:** Eine organisatorische Massnahme zur Regelung von Datenverantwortlichkeiten stellt das Data-Ownership-Konzept dar. Hierunter wird „die umfassende Verantwortung eines Fachbereichs für bestimmte Dateninhalte mit dem untrennbar damit verbundenen Metadatenmanagement“ [Meyer 2000, S. 72] verstanden. Hierdurch können unklare Zuständigkeiten vermieden und somit eindeutige Ansprechpartner für bestimmte Daten definiert werden.

4 Datenqualitätsmanagement für Data-Warehouse-Systeme

Nach der Betrachtung der betrieblichen Praxis im vorangegangenen Kapitel soll nachfolgend ein Datenqualitätskonzept für Data-Warehouse-Systeme vorgestellt werden. Ausgehend von den in Kapitel 1 beschriebenen, allgemeinen Ansätzen des Qualitätsmanagements wird die Konzeption eines Datenqualitätsmodells für Data-Warehouse-Systeme entwickelt und anhand eines auf dem zweiten CC DW2 Workshop erarbeiteten Beispiels konkretisiert. Im Anschluss daran wird für diesen Datenqualitätsansatz ein aus Phasen und Arbeitspaketen bestehendes Vorgehensmodell erläutert. Abschliessend erfolgt eine kritische Betrachtung des Ansatzes.

Aus dem operativen Datenqualitätsmanagement (siehe Kapitel 2.3) und den Prinzipien des TQM (siehe Kapitel 2.4) lässt sich ein Konzept für ein Datenqualitätsmanagement für Data-Warehouse-Systeme entwickeln [vgl. Helfert 2000b, S. 67f.]. Dieses stützt sich grundsätzlich auf drei zentrale Bereiche [vgl. Wolf 1999, S. 74]:

- Die Verpflichtung des Managements, Datenqualität als Philosophie und Unternehmenskultur vorzuleben. Auf Basis formulierter Unternehmensgrundsätze und -ziele ist eine Datenqualitätspolitik und eine Datenqualitätsstrategie abzuleiten. [vgl. Seghezzi 1996, S. 51]
- Ein Qualitätsmanagementsystem, welches den organisatorischen Rahmen darlegt, ist zu etablieren. Nach DIN ISO 8402 umfasst dieses die Aufbau- und Ablauforganisation, die Zuständigkeiten, Prozesse und Mittel für die Qualitätssicherung. Es stellt sicher, dass in allen Bereichen geeignete Prozesse, Richtlinien, Pläne sowie Test- und Prüfverfahren etabliert sind, die die geforderte Datenqualität gewährleisten. Hierzu ist eine ständige Überprüfung, Analyse und Verbesserung der gewählten Massnahmen und durchzuführenden Prozesse erforderlich.
- Zur Unterstützung der Mitarbeiter bei der Ausübung der Qualitätsprozesse sind in allen Phasen geeignete Methoden, Verfahren und Werkzeuge zur Verfügung zu stellen.

Zur Umsetzung eines DQM ist ein Messsystem notwendig, welches durch Qualitätskriterien und Qualitätsmessungen beschrieben werden kann [vgl. Jeusfeld, Jarke, Quix 1999, S. 10]. Die Anwendung eines Messsystems ermöglicht es,

- Qualitätsziele zu definieren,
- Aussagen über die vorhandene Datenqualität zu treffen sowie
- durchgeführte Verbesserungsmassnahmen zu bewerten.

Zur Unterstützung dieser Anforderungen wäre ein objektives Messsystem notwendig, das anhand von Qualitätskriterien Datenqualitätsziele operationalisiert und den Grad der Erreichung festgelegter Qualitätsziele objektiv misst. Bisher existiert jedoch noch kein universell anzuwendendes Messsystem.

In den folgenden Unterkapiteln soll daher ein Verfahren dargestellt werden, mit dessen Hilfe Qualitätskriterien definiert sowie Messgrössen spezifiziert werden können. Dies stellt damit einen Ansatz zur Operationalisierung des DQM-Gedankens dar. In Kapitel 4.1 erfolgt zunächst eine konzeptionelle Erläuterung des Ansatzes, die in Abschnitt 4.2 mittels eines Beispiels ergänzt, verfeinert und erläutert wird. Im Unterkapitel 4.3 wird das Vorgehen zur Umsetzung des Datenqualitätskonzepts mittels eines phasenorientierten Vorgehensmodells expliziert. Abschliessend wird in Kapitel 4.4 auf offene Fragen bei der Umsetzung im Unternehmen und den weiteren Forschungsbedarf eingegangen.

4.1 Konzeptionelle Darstellung des Datenqualitätsmodells

Der Kern des hier propagierten Ansatzes stellt ein Datenqualitätsmodell dar, welches aus drei Ebenen besteht und in Abb. 4-1 dargestellt ist. Das Modell umfasst die Aufgabenbereiche der Qualitätsplanung und Qualitätslenkung aus dem operativen Qualitätsmanagement.

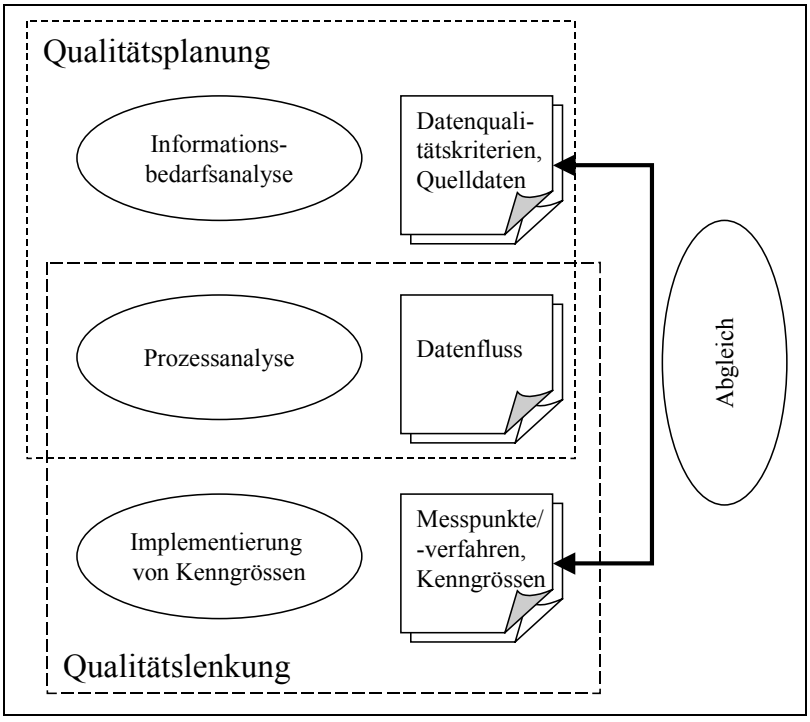

Abb. 4-1: Grundstruktur des Qualitätsmodells

Das Äquivalent zum Fertigungsprozess im herkömmlichen Qualitätsmanagement stellt der Datenproduktionsprozess im DQM dar. Letzterer wird auf der Ebene der *Prozessanalyse* betrachtet. Diese stellt den zentralen Baustein des Modells dar und bildet die Schnittstelle zwischen Qualitätsplanung und -lenkung. Es wird der gesamte Datenproduktionsprozess untersucht, der zur Erfüllung eines bestimmten Informationsbedarfs notwendig ist. Dies ist vergleichbar mit der ganzheitlichen Analyse des Fertigungsprozesses im TQM. Das Ergebnis der Prozessanalyse umfasst den gesamten Datenfluss im Data-Warehouse-System, d. h. von der Datenerfassung über das Data Warehouse selbst bis hin zu den analytischen Informationssystemen. Hierbei werden auf der Basis eines bestimmten Informationsbedarfs sämtliche Datenübertragungen bzw. -transformationen und die davon betroffenen Applikationen sowie Datenspeicher analysiert und in Form eines Datenflusses erfasst.

Die *Informationsbedarfsanalyse* liefert die für die Prozessanalyse notwendigen Informationen. Der Prozess der Informationsbedarfsanalyse kann grundsätzlich dreigeteilt werden (Abb. 4-2). So befasst sich der erste Teil mit der Ermittlung des eigentlichen Informationsbedarfs. Dieser kann unterschiedlich kommuniziert werden: mittels Aussagen der Datennutzer (Informationsbedarf), mittels bestehender Berichte (Informationsangebot), mittels vorhandener Prozessdokumentationen oder bestehender Datenmodelle der operativen Systeme. Der ermittelte Informationsbedarf ist zu konsolidieren und zu homogenisieren. Der zweite Teil betrifft die Quelldatenanalyse. Hierbei werden die für einen bestimmten Informationsbedarf benötigten Quelldaten bzw. Quelldatensysteme identifiziert. Diese liefern den Ausgangspunkt für eine Datenflussbetrachtung in der Prozessanalyse. Da ein Data Warehouse abhängig ist von den operativen Systemen, muss geprüft werden, ob der ermittelte Informationsbedarf überhaupt realisierbar ist. Verschiedene Probleme mit der Qualität der Quelldaten verhindern eine Bereitstellung der von den Benutzern gewünschten Informationen. Aus diesem Grunde ist es notwendig, dass die Benutzer für jede gewünschte Information entsprechende Qualitätskriterien angeben. Diese Kriterien entsprechen den subjektiven, anwendungsbezogenen Datenqualitätsanforderungen der Datennutzer und sind notwendig, um eine Bereitstellung der Daten in ausreichender Qualität zu ermöglichen.[3] Dies entspricht dem Prinzip der Kundenorientierung im TQM, da auch hier genau die vom Kunden (Datennutzer) nachgefragte Qualität produziert wird. Der dritte Teil der Informationsbedarfsanalyse befasst sich mit der Transformation der gesammelten Aussagen zum Informationsbedarf in ein Fachkonzept. Dieser Schritt ist jedoch für das

[3] Vgl. hierzu den „Anwenderbezogenen Ansatz“ zur Definition von Qualität im Kapitel 2.1.

Datenqualitätsmanagement nicht mehr von Bedeutung und wird deshalb nicht weiter betrachtet.

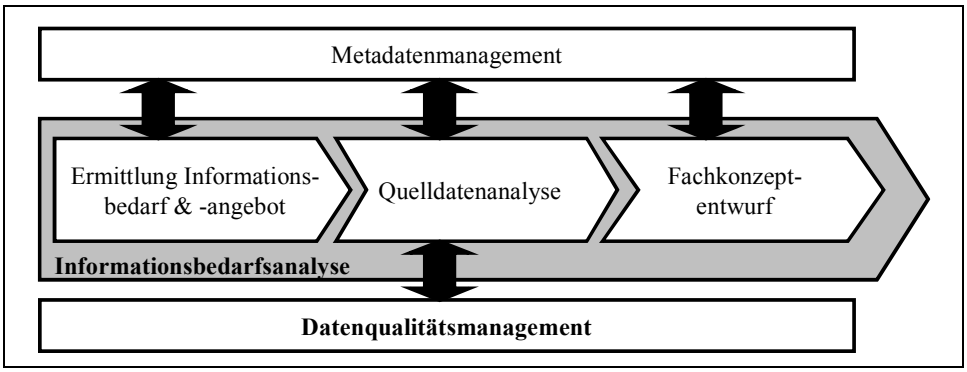

Abb. 4-2: Prozess der Informationsbedarfsanalyse

Das Problem der im Rahmen der Informationsbedarfsanalyse erfassten Qualitätskriterien ist meist deren fehlende direkte Messbarkeit. Die vom Datennutzer explizierten Qualitätsanforderungen korrespondieren i. d. R. mit den in Kapitel 2.2 dargestellten Kriterien der Datenqualität. Vollständigkeit, Korrektheit und Interpretierbarkeit sind Beispiele für derartige Anforderungen. Daher wird versucht, auf der Ebene der *Implementierung von Kenngrössen* Messpunkte und Messverfahren anhand des analysierten Datenflusses zu spezifizieren. Aufbauend auf der Prozessanalyse können durch die Prozessverantwortlichen bzw. durch Experten mögliche Fehlerquellen im Datenfluss identifiziert werden. Hierdurch lassen sich Messpunkte und dazugehörige Messtechniken ermitteln. Für die Messpunkte können nun sog. Kenngrössen (auch Kennzahlen oder Indikatoren genannt) festgelegt werden.

Diese messbaren Kenngrössen müssen mit den Qualitätskriterien des Datennutzers in Verbindung gebracht werden, sodass die nicht messbaren Qualitätskriterien mittels der Kennzahlen einer Quantifizierung bzw. Bewertbarkeit zugeführt werden. Im Rahmen des *Abgleichs* ist also die Beziehung zwischen den vom Anwender angegebenen Datenqualitäts-anforderungen und den Kenngrössen herzustellen. Dies kann auf unterschiedliche Weise realisiert werden:

- Beispielsweise können durch iterative Verfahren Zusammenhänge zwischen den Kenngrössen und den Datenqualitätskriterien über Sensitivitätsanalysen ermittelt werden. Hierbei werden punktuell auf einen Messpunkt im Datenfluss bezogene Änderungen an den Daten vorgenommen. Die Auswirkungen auf die Qualität der Daten des Anwenders werden bspw. anhand dessen Erfahrungswissens erfasst. Hierdurch kann nun ein Zusammenhang zwischen Kenngrössen und Qualitätskriterien hergestellt werden.

- Auch kann mit Hilfe von Expertenwissen oder durch eine Befragung der Datennutzer ein derartiger Abgleich stattfinden, indem ein hierarchisches Kennzahlensystem für die Datenqualitätskriterien entwickelt wird. Über sachlogische Zusammenhänge werden Hierarchien gebildet, an deren oberen Ende jeweils ein Qualitätskriterium steht, sodass erkennbar ist, welche Kenngrössen Einfluss auf das Qualitätskriterium haben.
- Das Prinzip der internen Kunden-Lieferanten-Beziehung aus dem TQM kann herangezogen werden, sodass jeder nachgelagerte Prozessschritt im Datenfluss als Kunde des vorangegangenen Schritts, der den Datenlieferanten darstellt, aufgefasst werden kann. Dieses Kunden-Lieferanten-Verhältnis kann auf den gesamten Datenfluss angewendet werden. Verbunden hiermit ist die Forderung, dass der Lieferant jeweils die von seinem Kunden geforderte Datenqualität erfüllen bzw. bereitstellen muss. Der Entscheider, der am Ende des Datenflusses steht, ist somit der letzte Kunde in der gesamten Kunden-Lieferanten-Kette. Ausgehend von dem vom Entscheider spezifizierten, notwendigen Qualitätsniveau wird der Datenfluss rückwärts traversiert bis zur Datenerfassung, die den ersten Datenlieferanten darstellt. Bei der Traversierung werden die jeweils notwendigen Qualitätsniveaus durch die Kunden spezifiziert und an die Datenlieferanten weitergegeben. Diese können dann wiederum ihre eigenen Qualitätsanforderungen festlegen und an den Lieferanten ihrerseits übermitteln. Hierdurch werden demnach, ausgehend vom geforderten Qualitätsniveau des Entscheiders, Datenqualitäts-anforderungen in den einzelnen Prozessschritten spezifiziert. Dies entspricht den gesuchten Kenngrössen für die einzelnen Prozessschritte. Die Prämisse für dieses Verfahren ist jedoch, dass in den jeweiligen Prozessschritten der Zusammenhang zwischen Input- und Outputqualität der Daten vorhanden sein muss. Für einfache Prozessschritte kann dieser Zusammenhang i. d. R. exakt beschrieben werden. Bei komplexeren Fällen müssen Schätzungen auf der Basis von Expertenwissen oder Testfällen herangezogen werden.

Durch den Abgleich kann erreicht werden, das die vom Datennutzer geforderte Datenqualität sichergestellt wird, indem Qualitätsniveaus für die Kenngrössen abgeleitet werden. Mit Hilfe des hier beschriebenen Modells können also vorhandene Qualitätsdefizite erkannt, Ansatzpunkte für den effizienten Einsatz von Verbesserungsmassnahmen identifiziert und die Auswirkungen solcher Massnahmen quantifiziert werden.

Auf dem zweiten CC DW2 Workshop wurde in Kooperation mit den Partnerunternehmen dieser Ansatz anhand eines Fallbeispiels diskutiert und verfeinert. Dies soll im nachfolgenden Kapitel dargestellt werden und dient der eingehenden Erläuterung des Verfahrens.

4.2 Exemplarische Darstellung des Datenqualitätsmodells

Der Aufbau dieses Kapitels orientiert sich an dem oben beschriebenen Datenqualitätsmodell (Abb. 4-1) und veranschaulicht dieses anhand eines mit Unternehmensvertretern erarbeiteten Beispiels. Im ersten Abschnitt wird die *Informationsbedarfsanalyse* beschrieben. Hierbei werden die notwendige Daten, wie Qualitätskriterien und Quelldaten, erhoben und erfasst. Kapitel 4.2.2 beinhaltet die *Prozessanalyse*, in der ausgehend von den Quelldaten der Datenfluss untersucht wird. Anschliessend erfolgt die Erläuterung der *Implementierung von Kenngrössen* und Messverfahren. Die Beschreibung des Datenqualitätsmodells endet mit der Beleuchtung des *Abgleichs* von Kenngrössen und Qualitätskriterien.

4.2.1 Informationsbedarfsanalyse

Der Informationsbedarf wird definiert als „die Art, Menge und Qualität der Informationen, die eine Person zur Erfüllung ihrer Aufgaben in einer bestimmten Zeit benötigt“ [Picot, Reichwald, Wigand 2001, S. 81]. Der erste Schritt der Informationsbedarfsanalyse besteht in der Erfassung des subjektiven Informationsbedarfs eines Entscheiders[4]. In dieser Arbeit wird angenommen, dass der Informationsbedarf in Form von analytischen Fragestellungen expliziert wird. Diese treten einerseits in semistrukturierten Entscheidungssituationen auf, haben eine hohe Priorität für den Entscheider und ihre Beantwortung determiniert i. d. R. massgeblich die Alternativenselektion. Andererseits dienen diese nicht nur der Entscheidungsfindung, sondern auch der Problemerkennung und der Unterstützung von Routinetätigkeiten durch die Bereitstellung relevanter Daten. Auf dem zweiten CC DW2 Workshop wurden Beispiele für analytische Fragestellungen, bezogen auf bestimmte Geschäftsvorfälle, erarbeitet. Für den Controllingbereich eines Versicherungsunternehmens stellt bspw. die Anzahl der Verträge pro Region eine solche analytische Fragestellung dar. Im Bankenbereich wäre bei der Kreditvergabe z. B. die Frage nach der Höhe des Kreditausfallrisikos zu stellen. Eine Auflistung möglicher Fragestellungen, die innerhalb ausgewählter Geschäftsvorfälle auftreten können, ist in Tab. 4-1 zu finden.

[4] Als Entscheider werden in dieser Arbeit die Anwender der analytischen Informationssysteme aufgefasst.

Geschäftsvorfall	Analytische Fragestellung
Geschäftsbereichscontrolling in der Versicherungsbranche	• Anzahl der Verträge pro Region/pro Sparte/pro Vertreter • Wachstum der Prämieneinnahmen in einem bestimmten Zeitraum • 20% der profitabelsten Kundenbeziehungen
Kreditvergabe in der Bankenbranche	• Kreditausfallrisiko anhand persönlicher Merkmale des Kunden • Maximales Kreditlimit des Kunden • Gesamtsumme aller laufenden Kredite
Mailing-Aktion im Marketing	• Spezifische Merkmale der Subjekte der Zielgruppe • Anzahl der neu gewonnenen Kunden durch das Mailing

Tab. 4-1: Auswahl möglicher analytischer Fragestellungen

Der zweite Schritt der Informationsbedarfsanalyse betrifft die Quelldatenanalyse. Hierbei werden sowohl die Daten bestimmt, die zur Beantwortung der analytischen Fragestellung notwendig sind, als auch die Quelldatensysteme identifiziert, in denen die benötigten Daten enthalten sind. Die Quelldatenanalyse soll im folgenden exemplarisch für die einfache analytische Fragestellung *„Anzahl der Verträge pro Region“* durchgeführt werden. Folgende Ausgangsdaten sind zur Beantwortung der Fragestellung notwendig:

- **Vertreterdaten:** Vertreter-Nr., Region-Nr., Vertreterstammdaten
- **Vertragsdaten:** Vertrags-Nr., Vertreter-Nr., Kunden-Nr., Produkt-Nr., Vertragsstammdaten
- **Kundendaten:** Kunden-Nr., Kundenstammdaten
- **Produktdaten:** Produkt-Nr., Produktstammdaten

Es wird willkürlich angenommen, dass diese vier Datenkategorien aus unterschiedlichen Quellsystemen stammen, sodass vier verschiedene Datentöpfe vorliegen, und zwar Vertreter-, Vertrags-, Kunden- und Produktdaten. Das zugrundeliegende Datenmodell zeigt Abb. 4-3.

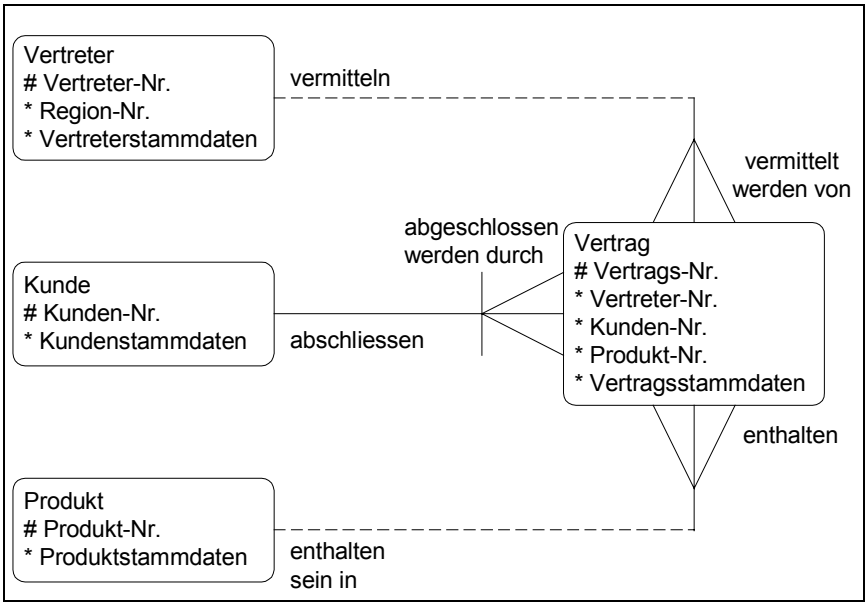

Abb. 4-3: Zugrundeliegendes Datenmodell[5]

Die Entität *Vertreter* enthält neben dem Schlüssel Vertreter-Nr. und den Vertreterstammdaten das Attribut Region-Nr., welches den Vertreter einer bestimmten Region zuordnet, in der er tätig ist. Dies ist für die analytische Fragestellung relevant, da die Anzahl der Verträge pro Region untersucht werden soll. Die Entität *Kunde* beinhaltet sowohl das Schlüsselattribut Kunden-Nr. als auch weitere Kundenstammdaten. Ähnlich hierzu ist die Entität *Produkt* aufgebaut. Diese besteht aus Produkt-Nr. und den Produktstammdaten. Beispiele für Produkte aus der Versicherungsbranche sind Lebensversicherungen, Krankenversicherungen und Haftpflichtversicherungen. Die Entität *Vertrag* besitzt als identifizierendes Attribut die Vertrags-Nr. Daneben ist enthalten, welcher Vertreter mit welchem Kunden einen Vertrag über welches Produkt abgeschlossen hat. Vertragsstammdaten sind bspw. das Datum des Vertragsabschlusses und die Vertragslaufzeit.

Im Rahmen der Quelldatenanalyse müssen vom jeweils betroffenen Entscheider für seine analytischen Fragestellungen Qualitätskriterien definiert werden, mit deren Hilfe die aus dem Anwendungskontext heraus relevanten Qualitätseigenschaften der Daten beschrieben werden können. Diese Qualitätskriterien können gruppiert werden in quantitative, qualitative, technische und andere Merkmale. Derartige Kriterien wurden auf dem Workshop exemplarisch für die bereits oben ausgewählte, sehr einfache analytische Fragestellung *„Anzahl der Verträge pro Region“* erarbeitet. Die Fragestellung kann zwar weiter detailliert werden bzgl. der verschiedenen Vertragsarten, der Behandlung von Verträgen mit

Ehepartnern etc. Eine Konkretisierung ist jedoch für den weiteren Verlauf nicht notwendig und soll daher vernachlässigt werden. Für die Fragestellung lassen sich nun verschiedene Beispiele für Qualitätskriterien finden, die die notwendigen Qualitätsmerkmale der Daten beschreiben, welche dem Entscheider zur Beantwortung der analytischen Fragestellung geliefert werden. Ein quantitatives Qualitätskriterium ist die Vollständigkeit der in den Daten enthaltenen Verträge. Der Begriff Vollständigkeit kann sich dabei auf unterschiedliche Vertragsattribute, nämlich die Vertragsart, den Vertragstyp oder aber die Geschäftsbereichszugehörigkeit, beziehen. Die Interpretierbarkeit der Daten gehört zur Gruppe der qualitativen Qualitätsanforderungen, da hier eine exakte Quantifizierung kaum möglich ist. Die Interpretationsfähigkeit der Daten kann bspw. durch beigefügte Definitionen zu Kennzahlen oder Fachbegriffen erhöht werden. Unter technischen Qualitätskriterien wird z. B. die Antwortzeit oder aber die Aktualität der gelieferten Daten verstanden. Einen Überblick über mögliche Qualitätskriterien für die analytische Fragestellung „Anzahl der Verträge pro Region“ enthält Tab. 4-2. Diese ist keineswegs als vollständig anzusehen, sondern gibt nur exemplarische Beispiele für bestimmte Gruppen von Qualitätsanforderungen im Kontext des hier betrachteten Beispiels an.

Gruppenzuordnung	Qualitätskriterium
Quantitativ	• Vollständigkeit, bezogen auf die Vertragsarten, die Vertragstypen (aktiv, passiv, Antrag läuft), die Geschäftsbereiche etc. • Zeitlicher Bezug, d. h. richtige Zuordnung der Verträge zur Auswertungsperiode
Qualitativ	• Interpretierbarkeit, d. h. Definitionen von Fachbegriffen, Kennzahlen etc. müssen mitgeliefert werden • Glaubwürdigkeit, insb. Eindeutigkeit der Ergebnisse
Technisch	• Aktualität des DWH (z. B. monatlich oder täglich aktuell) • Antwortzeit bzw. Verfügbarkeit des DWH
...	• ...

Tab. 4-2: Qualitätskriterien für die analytische Fragestellung „Anzahl der Verträge pro Region“

Um die Qualitätskriterien zu vervollständigen, müssen vom Entscheider Qualitätsanspruchsniveaus quantifiziert werden, sodass die analytische Fragestellung im konkreten Anwendungsumfeld ausreichend gut beantwortet wird. Der Entscheider muss demnach die Datenqualität festlegen, die notwendig ist, um eine sinnvolle Entscheidung treffen zu können.

[5] Als Modellierungssprache wird die Krähenfussnotation verwendet [vgl. Haux et al. 1998, S. 95.].

Hieraus wird deutlich, dass diese Anspruchsniveaus zum einen subjektiv, d. h. vom Entscheider abhängig, und zum anderen anwendungsbezogen sind, d. h. vom konkreten Verwendungszweck determiniert werden. Derartige Anspruchsniveaus, bezogen auf die Qualitätskriterien aus Tab. 4-2, sind bspw.:

- Die Anzahl der Verträge, bezogen auf die Vertragstypen, darf lediglich um 2% vom realen Wert abweichen.
- Alle Regionen und Vertreter sind aufzuführen.
- 98% der Verträge müssen der richtigen Periode zugeordnet sein.
- Die im Bericht enthaltenen Kennzahlen müssen zu 90% definiert sein.
- Die Daten dürfen sich nachträglich nicht mehr ändern.
- Die Daten müssen monatlich aktuell sein.
- Die Antwortzeit muss weniger als 3 Minuten betragen.

Diese oben genannten Anspruchsniveaus sind natürlich nur fiktive Grössen, da eine sinnvolle Quantifizierung ausschliesslich in einem tatsächlichen Anwendungskontext möglich ist. Durch die Festlegung von Qualitätskriterien und Anspruchsniveaus werden die Anforderungen des Entscheiders an die Datenqualität expliziert. Die obige Auflistung zeigt, dass die einzelnen Datenqualitätskriterien den Merkmalen der Datenqualität aus Kapitel 2.2, wie bspw. Interpretierbarkeit, Glaubwürdigkeit etc., zugeordnet werden können. Problematisch ist jedoch, dass einige der Kriterien nicht direkt überprüfbar bzw. messbar sind. So ist es zwar möglich, die Antwortzeit eines DWH direkt zu messen, die Vollständigkeit der Vertragsdaten hingegen entzieht sich dieser Messbarkeit, da hierfür die notwendigen Vergleichsgrössen fehlen. Dieses Problem betrifft die Mehrzahl der Datenqualitätsmerkmale. Um eine Operationalisierung und damit eine Messbarkeit dieser Kriterien zu erreichen, müssen messbare Kenngrössen herangezogen werden. Zur Spezifizierung derartiger Indikatoren ist zuvor jedoch die Betrachtung des Datenflusses notwendig, was im folgenden Kapitel näher erläutert wird.

4.2.2 Prozessanalyse

Um messbare Kenngrössen zu identifizieren, ist die Analyse des Datenflusses (Datenproduktionsprozess) notwendig. Dieser reicht von der Datenentstehung bis zur Datenverwendung und umfasst sämtliche, die Daten betreffende Prozessschritte und Datenspeicher. Ein Datenfluss ist i. d. R. immer auf einen bestimmten Informationsbedarf

bezogen und orientiert sich an den Quelldaten, welche im Rahmen der Informationsbedarfsanalyse identifiziert wurden (siehe Kapitel 4.2.1).

Der in diesem Kapitel dargestellte Datenfluss bezieht sich auf die bereits oben ausgewählte analytische Fragestellung „Anzahl der Verträge pro Region“. Das Datenmodell (Abb. 4-3) und die Datenspeicher für Vertreter-, Vertrags-, Kunden- sowie Produktdaten (siehe Kapitel 4.2.1) stellen die Grundlage für eine eingehendere Betrachtung des gesamten Datenflusses dar. Dieser ist in Abb. 4-4 dargestellt und setzt bereits bei der Entstehung der Vertragsdaten an. Die Datentöpfe auf der linken Seite stellen Eingabedaten für die Funktionen bzw. Prozessschritte dar, die in Form von Rechtecken veranschaulicht werden, wohingegen rechts die Ausgabedaten zu finden sind. Im Folgenden soll der Datenfluss näher beschrieben werden.

Nach dem Verkaufsgespräch mit dem Kunden werden die Vertragsdaten entweder sofort elektronisch vom Vertreter im Laptop erfasst oder aber der Vertreter füllt ein Antragsformular aus, dessen Inhalt anschliessend in eine elektronische Form überführt wird. Das Antragsformular selbst besteht aus einzelnen Datenfeldern, die Informationen über den Kunden beinhalten. Hierbei wird zwischen Muss- und Kann-Feldern unterschieden. Einige der Felder besitzen auch Default-Werte und vorgegebene Wertebereiche, wie z. B. Datums- oder Postleitzahlenfelder. Zusätzlich zum eigentlichen Vertrag werden die relevanten Kundenstamm- und Kundenkontaktdaten erfasst, um das Kundenprofil zu ergänzen und damit das Beziehungsmanagement zu verbessern. Beispiele für derartige Kontaktdaten sind Wohnverhältnisse, Kundenreaktionen etc. Die Dateneingabe für den Versicherungsantrag erfolgt lokal am Arbeitsplatz des Vertreters. Daher ist eine periodische Datensynchronisation mit den Serversystemen der Versicherung notwendig, um neu abgeschlossene Verträge oder sonstige geänderte Daten in die zentralen Systeme zu übertragen. Hierbei werden insbesondere die operativen Systeme angesprochen, welche die Kunden- und Vertragsdaten verwalten. Die vom Vertreter erfassten und gesammelten Daten werden so in die operativen Systeme eingestellt und stehen zur weiteren Verarbeitung sowohl im operativen als auch im dispositiven Bereich zur Verfügung.

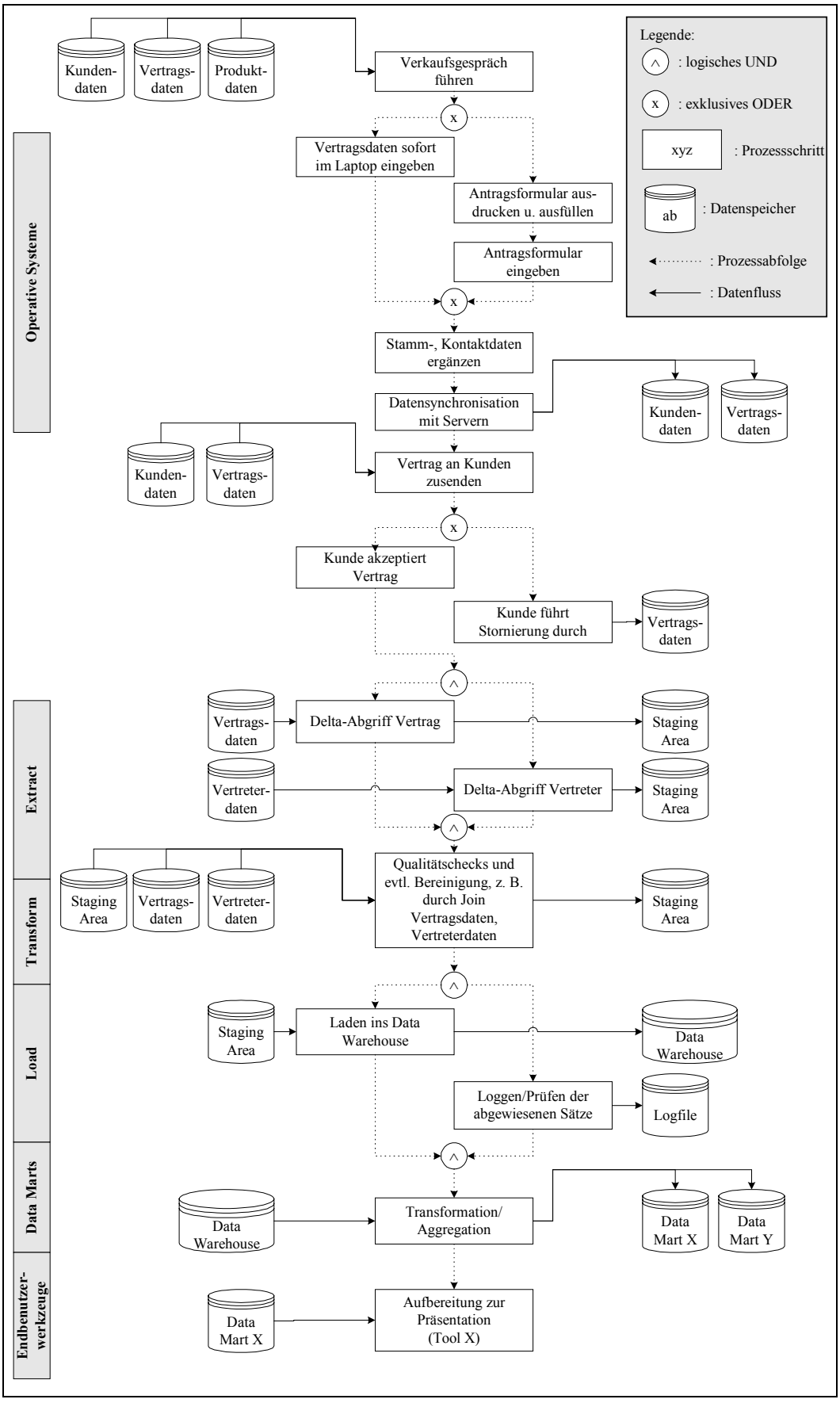

Abb. 4-4: Prozessschritte mit Ein- und Ausgabedaten

Nachdem der Versicherungsvertrag und die Vertragsbedingungen dem Kunden in schriftlicher Form übersandt wurden, hat dieser die Möglichkeit, den Vertrag zu akzeptieren oder nachträglich zu stornieren. Im ersteren Fall würde der Kunde i. d. R. überhaupt nicht reagieren, wohingegen bei der Vertragsannullierung der Kunde aktiv eine Stornierung durchführen muss. Die Vertragsauflösung hat eine Inaktivierung oder Löschung der Vertragsdaten aus dem operativen System zur Folge. Ein Inaktivierung kann durch Setzen bestimmter Felder erfolgen. Für aktive Vertragsdaten gilt, dass diese im operativen System erhalten bleiben und weiterverarbeitet werden können.

Im nächsten Schritt erfolgt ein Delta-Abgriff[6] der Vertrags- und Vertreterdaten aus den operativen Systemen. Diese enthalten alle seit der letzten Extraktion neu hinzugekommenen Daten über Verträge und Vertreter und stellen damit die zum Data Warehouse neu hinzuzufügenden Daten dar. Diese Delta-Loads werden zur weiteren Aufbereitung und Verarbeitung auf der Staging Area[7] zwischengespeichert. Durch einen Vergleich der extrahierten Daten mit den gesamten Vertrags- bzw. Vertreterdaten kann eine Qualitätsüberprüfung durchgeführt werden. Es kann bspw. festgestellt werden, ob Vertreter überhaupt vorhanden sind oder ob Dubletten in den Daten existieren. Auch sind weitere Data-Cleansing-Massnahmen zur Bereinigung bzw. Transformation der Daten denkbar. Die bereinigten und transformierten Daten werden wieder auf der Staging Area abgelegt, um anschliessend in das Data Warehouse geladen zu werden. Parallel hierzu wird ein Logfile für den Ladeprozess angelegt, in welchem die aufgetretenen Fehler protokolliert werden.

Um die einzelnen Data Marts für die Endbenutzerwerkzeuge mit Daten zu speisen, ist wiederum jeweils ein ETL-Prozess durchzuführen, d. h. es werden für jeden Data Mart themenspezifisch Daten aus dem Data Warehouse extrahiert, sodass ein effizienter Zugriff der analytischen Informationssysteme sichergestellt werden kann. Im hier beschriebenen Beispiel muss der Data Mart X die Vertragsdaten für alle Regionen beinhalten, sodass der Entscheider, der das Endbenutzerwerkzeug X einsetzt, seine analytische Fragestellung „Anzahl der Verträge pro Region“ beantworten kann. Die Daten aus dem Data Mart werden vom Endbenutzerwerkzeug je nach Funktionalität grafisch aufbereitet und bspw. in Form

[6] Ein Delta-Abgriff (oder auch Delta-Load) enthält lediglich die fortlaufenden Änderungen und Ergänzungen der operativen Datenbestände im Gegensatz zu einem Full-Load, der die gesamten Daten der operativen Systeme enthält. Dieses Verfahren wird hauptsächlich beim Extraktionsprozess eingesetzt.

[7] Eine Staging Area bezeichnet einen Datentopf, der als temporärer Zwischenspeicher genutzt wird. Im Rahmen des ETL-Prozesses werden die Daten i. d. R. nicht sofort in das DWH geschrieben, sondern auf einer solchen Staging Area zwischengespeichert, um bspw. Data-Cleansing-Massnahmen oder andere Korrekturverfahren anzuwenden.

eines OLAP-Würfels dargestellt. Dieser hier beschriebene Datenfluss wurde zusammen mit den Partnerunternehmen innerhalb des Workshops entwickelt und stellt exemplarisch eine Möglichkeit dar, wie ein derartiges Datenschema aussehen könnte. Prämissen wie bspw. die zugrundeliegende Data-Warehouse-Architektur, die den Datenfluss wesentlich beeinflusst, werden hier nicht expliziert, um die Anschaulichkeit des Beispiels zu erhalten und die Komplexität möglichst gering zu halten.

Anhand des Datenflussprozesses können nun Messpunkte und Messverfahren spezifiziert werden. Darüberhinaus ist eine Operationalisierung des Datenqualitätsgedankens ist möglich. Dies wird im folgenden Kapitel dargestellt.

4.2.3 Implementierung von Kenngrössen

Anhand der spezifizierten Datenflüsse und Quellsysteme können nun potenzielle Fehlerquellen identifiziert werden, welche die Datenqualität beeinträchtigen. Derartige Fehlerquellen sind i. d. R. gleichzusetzen mit möglichen Messpunkten für die Datenqualität, da hier jeweils überprüft werden sollte, inwieweit tatsächlich Fehler aufgetreten sind. Eine Überprüfung kann jedoch nur stattfinden, falls geeignete Messverfahren für die jeweiligen Messpunkte definiert sind. Die Wahl des Verfahrens ist abhängig von der potenziellen Fehlerquelle.

Am Beispiel des Datenflusses aus Abb. 4-4 sollen nun in einem ersten Schritt mögliche Fehlerquellen identifiziert werden, welche die Datenqualität beeinträchtigen. Schon im ersten Prozessschritt bei der Erfassung der Vertragsdaten können Fehler durch Irrtümer oder Missdeutungen zwischen Vertreter und Kunde auftreten. Erfasst der Vertreter die Daten während des Verkaufsgesprächs elektronisch, so können bspw. Wörter, wie Namen oder Bezeichnungen, falsch geschrieben oder durch Unachtsamkeiten werden bestimmte Fakten, wie die Krankheitsgeschichte eines Kunden, nur unzureichend festgehalten werden. Es erfolgt also eine falsche oder unvollständige Erfassung der Daten bzgl. Vertrag oder Kunde. Eine weitere Fehlerquelle stellen die Vertragsformulare bzw. die elektronischen Eingabemasken selbst dar. Problematisch hierbei sind insbesondere Default-Werte, obligatorische Felder und die i. d. R. starre Struktur der Formulare. Vorgegebene Default-Werte werden bei der Datenerfassung oftmals unverändert übernommen oder Muss-Felder mit falschen Daten gefüllt, da hier eine Eingabe zwingend notwendig ist. Die häufig zu starre Struktur der Eingabemasken kann auch zu einer nicht intendierten Verwendung von Feldern führen. Hierbei könnten bspw. Felder für das Geburtsdatum zur Speicherung anderer Informationen genutzt werden, indem nicht plausible Geburtsdaten eingetragen werden.

Die Datensynchronisation mit den operativen Systemen birgt weiteres Fehlerpotenzial. So können Daten durch technische Fehler bei der Übertragung oder Speicherung korrumpiert werden. Auch durch eine falsche Interpretation oder veränderte Semantik der Daten können ungültige Werte geliefert werden. Produktnummern bspw. sind anhand bestimmter Kriterien aufgebaut, enthalten Informationen über das Produkt selbst und identifizieren dieses eindeutig. Wird die Semantik der Produktnummer im operativen System verändert, dies dem Vertreter jedoch nicht mitgeteilt, so kann die Einspeisung von „alten“ Produktnummern im schlimmsten Fall zur Zuordnung von Verträgen zu falschen Produkten führen. Auch kann eine Löschung von Vertrags- bzw. Kundendaten bei der Vertragsstornierung zur Minderung der Datenqualität führen. Unter Umständen treten Inkonsistenzen zwischen verschiedenen operativen Systemen auf, da die Löschung bspw. nicht in allen betroffenen Systemen durchgeführt wurde. Des Weiteren können durch eine fehlerhafte Referenzierung bei mehrfach vergebenen Kunden- bzw. Vertragsnummern Daten fälschlicherweise selektiert und gelöscht werden.

Im Rahmen des ETL-Prozesses können durch die Angabe von fehlerhaften Selektionskriterien die falschen Daten aus den operativen Systemen extrahiert werden. So kann bspw. durch die Angabe eines falschen Bezugszeitpunkts das falsche Delta gebildet werden, wodurch die Datenänderungen nur unvollständig an das Data Warehouse weitergegeben werden, d. h. Zwischenupdates gehen verloren. Denkbar ist auch, dass durch die unterschiedlichen Zeitzonen von operativen Systeme in verschiedenen geografischen Standorten die Delta-Bestimmung und Datenintegration fehlerhaft verläuft. Weitere Defekte können im Transformationsschritt auftreten, wie z. B. durch falsche Regeln oder Mappingfehler. Insbesondere bei Data-Cleansing-Prozessen sind zahlreiche Transformationen und Datenanpassungen notwendig, die eine Fülle von Fehlermöglichkeiten bieten. Die Überführung der Daten in das Data Warehouse (Extraktion) kann bspw. Datendefekte durch eine fehlerhafte Übertragung oder durch nicht beachtete Veränderungen im Datenschema hervorrufen. Dabei können z. B. Feldinhalte gegen neu festgelegte Wertebereiche verstossen. Der ETL-Prozess für die Data Marts weist die gleichen Fehlerquellen auf wie der des Data Warehouse.

Weitere Fehlerquellen sind sowohl beim Analysetool als auch beim Anwender selbst zu finden. Ist die Programmlogik des Analysetools fehlerhaft, so kann dies zahlreiche Folgen haben: So können aufgrund einer fehlerhaften Transformationskomponente Fehler in der Darstellung auftreten. Interpretiert das Analysewerkzeug bspw. die Grössenordnung von Zahlen falsch, so führt dies zu falschen Darstellungen, Berichten und Auswertungen. Jedoch

darf auch der menschliche Faktor als Fehlerquelle nicht unterschätzt werden. Benutzer, welche die dargestellten Daten nicht richtig interpretieren, ziehen falsche Schlussfolgerungen und treffen demzufolge falsche Entscheidungen. Die hier beschriebenen Fehlerquellen sind nochmals in Tab. 4-3 überblicksartig dargestellt. Diese Auflistung erhebt nicht den Anspruch der Vollständigkeit, sondern soll lediglich denkbare Fehlerquellen beispielhaft aufzeigen.

Prozessschritt	Fehlerquelle
Verkaufsgespräch führen	• Missverständnisse
Vertragsdaten sofort im Laptop eingeben	• Missverständnisse • Falscheingaben/Unzweckmässige Verwendung von Eingabefeldern
Antragsformular ausdrucken, ausfüllen und eingeben	• Missverständnisse • Falscheingaben/Unzweckmässige Verwendung von Eingabefeldern
Stamm-, Kontaktdaten ergänzen	• Falscheingaben/Unzweckmässige Verwendung von Eingabefeldern
Datensynchronisation mit Servern	• Technische Fehler bei der Soft- bzw. Hardware, z. B. Übertragungsfehler • Falsche Interpretation der Daten z. B. bei veränderter Semantik der Produktnummer
Kunde führt Stornierung durch	• Referenzierungsproblematik • Inkonsistenzen
Delta-Abgriff Vertrag/Vertreter	• Falsches Delta • Falscher Zeitpunkt, Zeitzonenproblematik • Verlorene Zwischenupdates
Qualitätschecks und evtl. Bereinigung	• Falsche Regeln • Mapping-/Transformationsfehler
Laden ins Data Warehouse	• Übertragungsfehler • Fehler durch erweitertes Datenschema
Transformation/Aggregation	• Falsche Regeln • Mapping-/Transformationsfehler
Aufbereitung zur Präsentation	• Transformationsfehler • Interpretationsfehler

Tab. 4-3: Potenzielle Fehlerquellen bezogen auf die Prozessschritte des Datenflusses

Aufbauend auf der Analyse der Fehlerquellen ist es möglich, Messpunkte im Datenfluss zu spezifizieren, an denen jeweils eine Kenngrösse tatsächlich gemessen werden kann, im Gegensatz zu den vom Entscheider festgelegten und i. d. R. nicht messbaren Qualitätskriterien in Kapitel 4.2.1. Für die einzelnen Messpunkte sind Messverfahren zu spezifizieren, die nach drei Kategorien klassifiziert werden können:

- **Vergleich mit realem Wert:** Dieses stellt das einfachste aller drei Verfahren dar. Hierbei wird der Datenwert mit dem tatsächlichen realen Wert, der aus einer anderen Quelle

herangezogen wird, verglichen. Dadurch kann zum einen bestimmt werden, welche Daten falsch bzw. richtig sind und zum anderen deren Qualität ermittelt werden. Voraussetzung hierbei ist jedoch, dass der reale Wert bekannt ist. Da das tatsächliche Datum i. d. R. nicht vorliegt, ist der Anwendungsbereich der Methode sehr eingeschränkt. Dadurch müssen Alternativlösungen herangezogen werden, um diese Einschränkung zu überwinden. Solche Messverfahren werden in den zwei folgenden Unterpunkten beschrieben.

- **Plausibilitätsprüfung:** Zu dieser Kategorie zählen Messverfahren, mit deren Hilfe lediglich überprüft werden kann, ob ein Datum nicht falsch ist. Es kann demnach nicht die Richtigkeit von Daten festgestellt werden. Zwei Unterkategorien können unterschieden werden: Zum einen kann die Plausibilität von Daten anhand von Erfahrungswerten überprüft werden. Hierbei vergleichen Fachexperten die vom System gelieferten Daten mit ihren jeweiligen erwarteten Werten, die auf längerfristigem Erfahrungswissen basieren. Durch diese Gegenüberstellung kann die vorliegende Datenqualität abgeschätzt werden. Ein einfaches Beispiel stellt der Load-Prozess für ein Data-Warehouse dar. Der Prozessverantwortliche kann anhand der Anzahl der in das Data Warehouse geladenen Datensätze unter Berücksichtigung seines Erfahrungswissens beurteilen, welche Datenqualität von diesem Prozessschritt „produziert“ wird. Die zweite Unterkategorie der Plausibilitätsprüfungen beinhaltet Methoden, die Sachverhalte durch einen Vergleich von Daten aus unterschiedlichen Systemen oder anhand von festgelegten Wertebereichen überprüfen. Es ist bspw. oftmals möglich, durch Berechnungen Werte zu generieren, die als Vergleichsgrössen herangezogen werden können. So ist es denkbar, die Daten über die Anzahl verkaufter Produkte aus dem Vertriebssystem den Lagerabgängen des Lagerverwaltungssystems gegenüberzustellen. Anhand der Mengenabweichungen kann die Datenqualität im operativen Vertriebssystem abgeschätzt werden. Obwohl durch die Plausibilitätsprüfungen wie bereits erwähnt nicht die tatsächliche Richtigkeit von Daten festgestellt werden kann, ist die Anwendung dieser Methoden automatisierbar und daher mit Kosteneinsparungen verbunden.
- **Statistische Analyse über Stichprobenergebnisse:** Bei dieser Verfahrensgruppe ist die reale Fehlerhäufigkeit unbekannt. Um dieses Informationsdefizit zu überwinden, werden Stichproben erhoben, die statistische Validität besitzen. Mit Hilfe des Stichprobenwertes kann dann das Ausmass der produzierten Datenqualität quantifiziert werden. Ein Beispiel hierzu ist die manuelle Dateneingabe bei Kassensystemen. Die hierbei auftretenden Falscheingaben können stichprobenartig gezählt werden, und ausserdem ist bei einer

ausreichend grossen Anzahl von Stichproben die Qualität bzw. Güte des Dateneingabeprozesses quantifizierbar. Problematisch hierbei sind jedoch die notwendigen manuellen Vorarbeiten, die zu erheblichem Zusatzaufwand und damit zu hohen Kosten führen.

Im Folgenden sollen beispielhaft für obige Fehlerquellen, die mit Messpunkten gleichzusetzen sind, konkrete Messverfahren angegeben werden.

- Die erstgenannte Fehlerquelle „Missverständnisse im Verkaufsgespräch“ kann mit Hilfe von Stichprobenerhebungen einer Quantifizierung zugeführt werden. Derartige Stichproben können erfasst werden, indem bei einer hinreichend grossen Zahl von Verkaufsgesprächen die Fehlerhäufigkeit gezählt wird. Dadurch lässt sich die Qualität der in dieser Funktion erzeugten Daten feststellen.
- Die zweite Fehlerquelle „Falscheingaben/Unzweckmässige Verwendung von Eingabefeldern“ kann mittels einer Plausibilitätsprüfung evaluiert werden. Hierbei können die Wertebereiche der Eingabefelder überprüft werden bzw. die eingegebenen Werte können den Durchschnittswerten gegenübergestellt werden. Denkbar ist auch die Betrachtung von Änderungshäufigkeiten bestimmter Felder, um dadurch auf die Datenqualität zu schliessen.
- Für die Fehlerquelle „Technische Fehler bei Soft- bzw. Hardware“ beim Datenupload ist eine genauere Spezifizierung des Defekts notwendig. Nimmt man als exemplarisches Beispiel hierfür einen Übertragungsfehler an, so kann dessen Häufigkeit mit Hilfe von Logfiles oder durch Fehlerprotokolle gemessen werden und damit die Datenqualität dieser Funktion ermittelt werden.

Die mit diesen Ansätzen gemessenen Kenngrössen beziehen sich immer nur auf die gerade betrachtete Funktion im Datenfluss und geben demnach keine Aussage über die Gesamtdatenqualität. Insbesondere fehlt i. d. R. der Zusammenhang zwischen den gemessenen Kennzahlen und den Qualitätskriterien bzw. Qualitätsanspruchsniveaus des Entscheiders. Hierzu ist ein Abgleich beider Grössen notwendig, der im folgenden Kapitel eingehender erläutert wird.

4.2.4 Abgleich

Der Abgleich von Kenngrössen und Qualitätskriterien kann auf unterschiedliche Weise realisiert werden, wie in Kapitel 4.1 beschrieben. Beispielhaft soll hier nur die Möglichkeit des Abgleichs durch die Bildung von Kennzahlensystemen betrachtet werden. Hierbei werden

die einzelnen Kennzahlen so verknüpft, dass eine Qualitätsaussage im Sinne des vom Entscheider spezifizierten Qualitätskriteriums möglich ist. Eine Verknüpfung kann grundsätzlich mathematisch oder sachlogisch erfolgen [vgl. Mutscheller 1996, S. 43]. In diesem Anwendungskontext ist lediglich die sachlogische Verknüpfung von Interesse, da i. d. R. die genauen mathematischen Zusammenhänge zwischen den Kenngrössen unbekannt sind. Es erfolgt eine Hierarchisierung bzw. Aggregation der Kennzahlen und damit eine Verdichtung der Einzelinformationen zum übergeordneten Qualitätskriterium. Im Folgenden soll dies an einem einfachen Beispiel verdeutlicht werden.

Ausgehend von dem Qualitätskriterium des Entscheiders, dass die gelieferten Daten monatlich aktuell sein müssen (siehe Kapitel 4.2.1), können Kennzahlen im Datenfluss ermittelt und zugeordnet werden. Für die Kennzahlen können Niveaus festgelegt werden, sodass die geforderte Aktualität sichergestellt wird. Kennzahlen zur Messung der Aktualität der Daten beziehen sich auf die Datenübertragungen bzw. Load-Funktionen im Datenfluss. Insbesondere ist von Interesse, wann diese Übertragungen stattgefunden haben. Auf den oben dargestellten Datenfluss bezogen, sind die relevanten Funktionen die Datensynchronisation der vom Vertreter erfassten Daten mit den operativen Systemen, der Ladevorgang des DWH und der Load-Prozess der Data Marts. Anhand von Protokolldateien kann ermittelt werden, wann diese Vorgänge durchgeführt wurden. Um nun eine Aktualität von einem Monat zu erreichen, darf die letztmalige Ausführung aller drei Funktionen nicht länger als einen Monat zurückliegen. Ausserdem müssen die Ausführungszeitpunkte aufeinander abgestimmt sein, sodass das Laden der Daten des Vertreters vor dem Update des DWH und dieses wiederum vor dem Load der Data Marts stattfindet. Diese Bedingungen stellen die sachlogischen Verknüpfungen dar. Sind die Bedingungen erfüllt, so ist sichergestellt, dass das Niveau des Qualitätskriteriums Aktualität erfüllt ist und die Daten in ausreichender Qualität beim Datennutzer vorliegen.

Ein weiteres, komplexeres Beispiel bezieht sich auf das Qualitätstkriterium „98% der Verträge müssen der richtigen Periode zugeordnet sein“ (siehe Kapitel 4.2.1). Dieses stellt ein vielschichtigeres Beispiel dar, da zahlreiche Funktionen im Datenfluss betroffen sind, die Messung der Datenqualität an den Messpunkten aufwendiger ist und die sachlogischen Verknüpfungen im Kennzahlensystem komplexer sind. Die für dieses Qualitätskriterium relevanten Funktionen sind die Eingabe der Vertragsdaten in den Laptop bzw. das Ausfüllen des Antragsformulars, die Datensynchronisation mit den Servern, der Transformationsschritt für das DWH, das Laden der Daten in das DWH und die Aggregation der Daten für die Data Marts. Da sich das Qualitätskriterium auf die Zuordnung von Verträgen zu Perioden bezieht,

ist insbesondere das Datumsfeld von Interesse, über welches diese Zuordnung stattfindet. Bei der Eingabe der Vertragsdaten im Laptop bzw. beim schriftlichen Ausfüllen der Verträge kann das falsche Datum erfasst werden. Die Fehlerhäufigkeit ist messbar über Stichprobenerhebungen, wie bereits in Kapitel 4.2.3 beschrieben wurde. Ein zweite Fehlerquelle stellt die Datensynchronisation mit den operativen Systemen dar. Hierbei können durch Übertragungsfehler die Datumsangaben korrumpiert werden. Die Fehlerwahrscheinlichkeit kann anhand der Fehlerprotokolle ermittelt werden. Weiteres Fehlerpotenzial beinhaltet der Transformationsschritt für das DWH. Die Integration der Daten aus unterschiedlichen operativen Systemen kann durch falsche Transformationsregeln zu Fehlern in den Datumsangaben führen. Insbesondere unterschiedliche Zeitzonen der operativen Systeme können derartige Fehler hervorrufen, da für jede Zeitzone die dazugehörige Zeitumrechnung in Form einer Regel vorhanden sein muss. Messbar ist die Fehlerhäufigkeit durch die Erhebung von Stichproben für den Transformationsschritt. Das Laden der Daten in das DWH stellt wieder eine mögliche Fehlerursache dar. Mittels der Protokolldatei kann der Fehlerumfang bei der Übertragung bestimmt werden. Auch die Aggregation bzw. Transformation der Daten für die Data Marts bzw. für das Analysewerkzeug kann fehlerhafte Transformationsregeln enthalten und so zu falschen Zuordnungen von Verträgen zu Perioden führen. Nach der Bestimmung der Fehlerwahrscheinlichkeiten der einzelnen Funktionen im Datenfluss sind diese derart in einem Kennzahlensystem zu verknüpfen, dass eine Aggregation zum übergeordneten Qualitätskriterium „Zuordnung der Verträge zur richtigen Periode“ möglich ist. Eine solche Verknüpfung kann zum einen von Fachexperten anhand von Erfahrungswissen durchgeführt werden. Zum anderen ist es möglich durch eine Befragung des Endbenutzers das Kennzahlensystem zu konkretisieren. Daher ist im hier vorliegenden Beispiel die genaue Ausgestaltung der komplexen Zusammenhänge nicht darstellbar.

Mit Hilfe des Abgleichs wird demnach die Verbindung zwischen messbaren Kenngrössen und anwenderbezogenen Qualitätskriterien hergestellt und so eine erste Operationalisierung des hier vorliegenden Datenqualitätsmodells erreicht. Zur weiteren Veranschaulichung und Konkretisierung wird im nachfolgenden Kapitel ein Vorgehensmodell eingeführt, welches das Datenqualitätsmodell phasenartig darstellt und Arbeitspakete für die einzelnen Schritte definiert.

4.3 Vorgehensmodell

Vorgehensmodelle werden häufig im Bereich des Software Engineering eingesetzt, können jedoch auch in anderen Domänen zur Anwendung kommen und haben die „Strukturierung des Entwicklungsprozesses und die Komplexitätsreduktion in Projekten durch eine idealtypische Gliederung in Phasen“ [Gabler Wirtschaftsinformatik Lexikon 1997, S. 756] zum Ziel. Das hier darzustellende Vorgehensmodell beschreibt die notwendigen Schritte zur Umsetzung des bereits ausführlich erläuterten Datenqualitätsmodells. Die Phasen lehnen sich dabei an die Ebenen des Datenqualitätsmodells an und beschreiben das bereits implizit zugrundeliegende Vorgehen. Dieses gliedert sich in fünf nacheinander zu durchlaufende Phasen, welche jeweils durch Arbeitspakete weiter detailliert werden (siehe Abb. 4-5). Die Reihenfolge der Arbeitspakete innerhalb der Phasen ist nicht festgelegt, da diese i. d. R. vom konkreten Projektverlauf abhängig ist. Im Folgenden werden die einzelnen Phasen eingehender betrachtet und die darin enthaltenen Arbeitspakete erläutert.

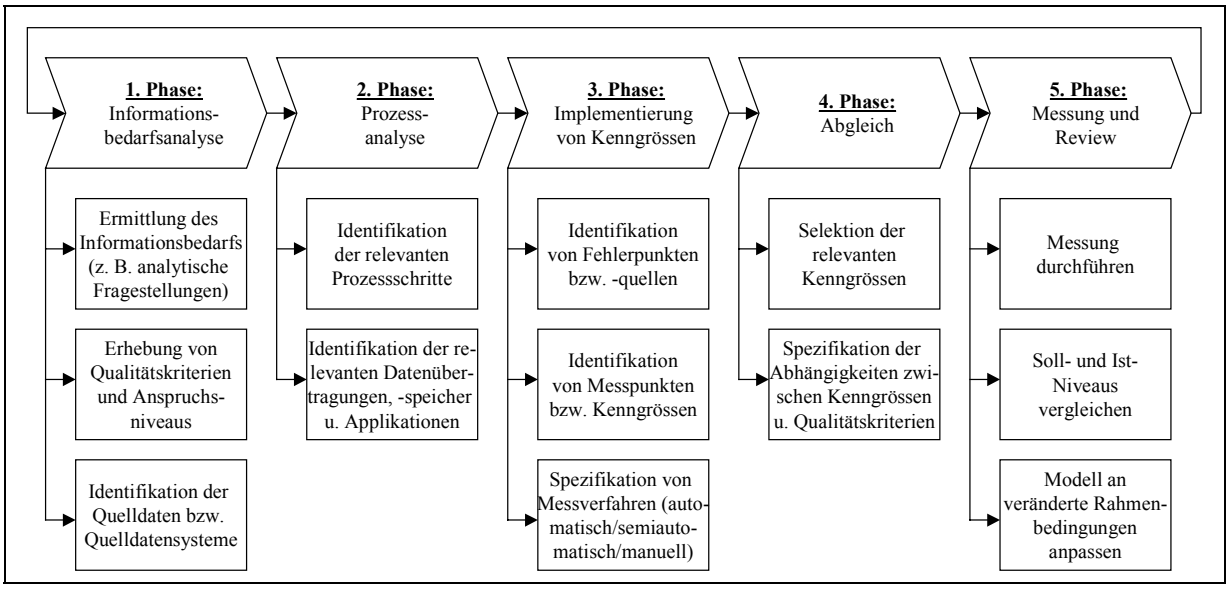

Abb. 4-5: Phasenartige Darstellung des Vorgehensmodells

1. Phase: Informationsbedarfsanalyse

Die Phase der Informationsbedarfsanalyse kann in drei Arbeitspakete zerlegt werden. Die erste Aufgabe besteht darin, den Informationsbedarf des Entscheiders zu ermitteln. Hierzu existieren zahlreiche Methoden, wie z. B. Interviews, Fragebögen, Beobachtungen, Aufgaben- oder Dokumentenanalysen [vgl. z. B. Beiersdorf 1995, S. 76ff.; Holten 1999, S. 120ff.]. Dabei wird zwischen deduktiven und induktiven Verfahren der Informationsbedarfsanalyse unterschieden. Während deduktive Verfahren versuchen, den

aufgabenbezogenen „richtigen“ Informationsbedarf zu ermitteln, fokussieren induktive Verfahren auf den personenbezogenen, subjektiven Informationsbedarf [vgl. Holten 1999, S. 120]. Im oben erläuterten Beispiel wurde der subjektive Informationsbedarf durch analytische Fragestellungen beschrieben.

Nachdem der Informationsbedarf ermittelt wurde, müssen darauf aufbauend dazugehörige Qualitätskriterien und Qualitätsanspruchsniveaus festgelegt werden. Dadurch werden die Qualitätsanforderungen des Entscheiders an den ermittelten Informationsbedarf expliziert. Diese Qualitätsanforderungen können im Rahmen der Informationsbedarfsermittlung, wie bereits oben beschrieben, erhoben werden. Hierdurch wird der Begriff der Qualität der Informationen konkretisiert und ein Ausgangspunkt zur Evaluierung der Datenqualität geschaffen. Es wird unterschieden zwischen direkt messbaren und nicht direkt messbaren Kriterien. Da erstere sofort überprüfbar sind, können für diese die Phasen zwei bis vier übersprungen werden. Ein Grossteil der Qualitätskriterien ist jedoch nicht unmittelbar messbar und muss daher das gesamte Vorgehensmodell durchlaufen.

Nach Erhebung des Informationsbedarfs und der qualitativen Anforderungen können die operativen Quelldaten bzw. die Quelldatensysteme, mittels derer der Informationsbedarf befriedigt wird, identifiziert werden. Hierbei wird der Informationsbedarf in seine Ausgangsdaten zerlegt. Unter den Ausgangsdaten werden die Daten verstanden, die zur Berechnung bzw. Generierung der nachgefragten Informationen notwendig sind. Würde bspw. der Umsatz einen Informationsbedarf darstellen, so wären die Verkaufsmengen und Verkaufspreise die dazugehörigen Ausgangsdaten. Anhand der Ausgangsdaten können dann die Quelldatensysteme ermittelt werden, d. h. es werden die Applikationen bzw. Datenspeicher identifiziert, die diese Quelldaten beinhalten.

2. Phase: Prozessanalyse

Im Rahmen der Prozessanalyse wird eine ganzheitliche Betrachtung des gesamten Data-Warehouse-Systems vorgenommen. Ausgehend von der Datenerfassung werden alle relevanten Prozessschritte, Datenübertragungen und Applikationen, die im Zusammenhang mit den in der ersten Phase identifizierten Quelldaten bzw. Quelldatensystemen für einen Informationsbedarf stehen, identifiziert. Zuerst werden alle Prozessschritte erfasst, in denen eine Bearbeitung, Transformation oder Nutzung der relevanten Daten stattfindet. Hierbei kann als Orientierungshilfe die bestehende Data-Warehouse-Architektur herangezogen werden. Diese Prozessanalyse muss bereits bei der Datenerfassung ansetzen und durchläuft bottom-up die Architektur bis zum analytischen Informationssystem bzw. Datennutzer. Für

jeden Prozessschritt werden dann die Ein- bzw. Ausgabedaten sowie die genutzten Applikationen ermittelt. Das Ergebnis dieser Phase stellt für einen bestimmten Informationsbedarf den Datenfluss dar, der genau aufzeigt, durch welche Prozessschritte die Daten transformiert bzw. übertragen werden und welche Datenspeicher hierbei tangiert werden.

3. Phase: Implementierung von Kenngrössen

Aufbauend auf dem Datenfluss werden in dieser Phase messbare Kenngrössen identifiziert. Zunächst jedoch erfolgt eine Analyse des Datenflusses nach Fehlerpunkten bzw. Fehlerquellen. Hierunter werden Fehlermöglichkeiten in den Prozessschritten verstanden, die die Qualität der in diesem Prozessteil verarbeiteten Daten negativ beeinträchtigen. Für derartige Fehler können sowohl technische als auch menschliche Aspekte sowie eine Kombination aus beidem verantwortlich sein. Die Fehlerquellen können von den Prozessverantwortlichen, die detailliertes Wissen über den Ablauf der einzelnen Prozessschritte besitzen, ermittelt werden.

Diese identifizierten Fehlerquellen stellen i. d. R. die Messpunkte dar, an denen eine Überprüfung der Datenqualität möglich und sinnvoll ist. Gemessen werden die sog. Kenngrössen, die jeweils für einen Fehlerpunkt im Prozessschritt spezifiziert werden. Für einen Fehler können meist mehrere Kenngrössen angegeben werden, die sich bzgl. des Messverfahrens[8] unterscheiden. Bei den Messungen werden automatische und manuelle Verfahren differenziert. Plausibilitätsprüfungen können meist automatisiert werden, wodurch ein relativ kostengünstiger Einsatz möglichist. Im Gegensatz dazu sind statistische Analysen über Stichprobenergebnisse sehr aufwendig, da teilweise manuelles Eingreifen und Vorarbeiten notwendig werden. Daher sollten Plausibilitätsüberprüfungen vorrangig eingesetzt werden und nur bei fehlenden Alternativen Stichprobenerhebungen zum Einsatz kommen.

4. Phase: Abgleich

Die bisher dargestellten Phasen beziehen sich jeweils auf einen vom Entscheider spezifizierten Informationsbedarf. In der Phase Abgleich jedoch werden die für den Informationsbedarf festgelegten Qualitätskriterien aus der ersten Phase eingehender betrachtet. In einem ersten Schritt werden die für ein Qualitätskriterium relevanten

[8] Die verschiedenen Kategorien für Messverfahren wurden in Kapitel 4.2.3 vorgestellt. Es werden der Vergleich mit einem realen Wert, Plausibilitätsprüfungen und statistische Analysen über Stichprobenergebnisse unterschieden.

Kenngrössen aus der Menge aller in Phase drei identifizierten Kenngrössen selektiert. Diese Auswahl erfolgt anhand sachlogischer Zusammenhänge, d. h. tangiert eine Kenngrösse das Qualitätsniveau eines Qualitätskriteriums, so muss diese selektiert werden. Jedem Qualitätskriterium werden auf diese Weise eine oder mehrere Kenngrössen zugeordnet.

Im zweiten Schritt müssen die Abhängigkeiten zwischen dem Qualitätskriterium und den dazugehörigen Kenngrössen genauer spezifiziert werden. Für ein einfaches Qualitätskriterium können die Beziehungen durch mathematische Regeln angegeben werden, sodass die realen Zusammenhänge repräsentiert werden. Hierdurch können die messbaren Kenngrössen zum übergeordneten Qualitätskriterium anhand der Regeln aggregiert werden. Für komplexe Qualitätskriterien jedoch ist die genaue Quantifizierung der Abhängigkeiten i. d. R. nicht möglich. Statt mathematischer Operationen müssen hier Zusammenhänge bspw. in Form von Wahrscheinlichkeiten definiert werden, die auf Erfahrungswissen, Schätzungen und Sensitivitätsanalysen beruhen. Hierdurch erfolgt eine Hierarchisierung der Kenngrössen, an deren oberem Ende wiederum das Qualitätskriterium steht. Durch den Abgleich wird also die Brücke zwischen den subjektiven, vom Entscheider spezifizierten, aber nicht messbaren Qualitätskriterien auf der einen Seite und den messbaren, jedoch auf einen einzelnen Prozessschritt beschränkten Kenngrössen auf der anderen Seite geschlagen.

5. Phase: Messung und Review

In der letzten Phase des Vorgehensmodells wird die eigentliche Messung der Datenqualität durchgeführt. In Abhängigkeit vom Verfahren erfolgt die Messung der einzelnen Kenngrössen automatisch, semiautomatisch oder manuell. Anhand der in Phase vier spezifizierten Abhängigkeiten können die Messergebnisse zu den übergeordneten Qualitätskriterien verdichtet werden. Das Ergebnis stellt die Ist-Datenqualität dar. Diese kann den Qualitätsanspruchsniveaus des Entscheiders (Soll-Datenqualität) gegenübergestellt werden. Qualitätsdefizite der Daten werden durch Abweichungsanalysen transparent. Dies stellt auch den Ausgangspunkt für den Einsatz von Verbesserungsmassnahmen zur Steigerung der vorliegenden Datenqualität dar.

Die durch obiges Vorgehen identifizierten Qualitätskriterien, Kenngrössen und Abhängigkeiten sind i. d. R. im Zeitablauf durch sich verändernde Rahmenbedingungen nicht stabil. Daher muss ein periodischer Review stattfinden, der das Datenqualitätsmodell an die veränderten Umweltbedingungen anpasst, sodass eine zuverlässige Datenqualitätsmessung gewährleistet wird.

Das oben beschriebene Datenqualitätsmodell und das dazugehörige Vorgehen bilden den Grundstein für ein umfassendes und operationalisierbares Datenqualitätsmanagement für Data-Warehouse-Systeme. Dennoch existieren eine Reihe von offenen Punkten, die im nachfolgenden Kapitel betrachtet werden sollen.

4.4 Kritische Betrachtung

Das in diesem Kapitel beschriebene Verfahren stellt einen ersten Ansatz dar zur Operationalisierung von Qualitätskriterien, die vom Entscheider festgelegt werden. Dem Ansatz liegt ein anwendungsbezogenes, subjektives Qualitätsverständnis zugrunde. Ausgehend von subjektiven Anspruchniveaus und einer Datenflussbetrachtung werden einerseits Fehlerquellen, die die Datenqualität innerhalb des Datenflusses beeinträchtigen, identifiziert. Andererseits werden Messpunkte und Messverfahren spezifiziert, mit deren Hilfe eine Quantifizierung von Kennzahlen möglich ist. Über einen Abgleich von Kennzahlen und Qualitätskriterien kann die geforderte Datenqualität des Datennutzers sichergestellt werden. Ein wichtiges Merkmal dieses Ansatzes ist die ganzheitliche Betrachtung des Data-Warehouse-Systems, beginnend mit der Datenerfassung über das eigentliche Data Warehouse bis hin zum Entscheider. Jedoch ist nicht nur die rein technische Umsetzung bedeutend, sondern darüber hinaus eine Sensibilisierung aller Betroffenen für das Datenqualitätsproblem notwendig, sodass bspw. auch die an der Datenerhebung massgeblich beteiligten Mitarbeiter ihrer Verantwortung bzgl. der Datenqualität bewusst werden. Insbesondere das Management ist hierbei gefordert, das Datenqualitätsverständnis z. B. durch Anreizsysteme zu fördern.

Das hier präsentierte Vorgehen beschreibt einen konzeptionellen Rahmen für das Datenqualitätsmanagement. Vor einer tatsächlichen Umsetzung gilt es, einige noch offene Fragen zu klären. Die Analyse der Quelldaten und der Datenflüsse stellt einen dieser Problembereiche dar. Hier wird die Frage aufgeworfen, in welchem Umfang eine derartige Analyse durchzuführen ist. Dies kann von individuellen Untersuchungen für jede analytische Fragestellung bis hin zu einer einzigen, generellen Betrachtung für das gesamte Unternehmen reichen. Hierbei sind auch Überlegungen bzgl. einer Kosten-/Nutzenbewertung des Verfahrens zu berücksichtigen. Ein dritter Problemaspekt betrifft die organisatorische Umsetzung. Da das Verfahren den gesamten Datenfluss umfasst, werden zahlreiche Verantwortungsbereiche tangiert. Um die Mitarbeit aller Abteilungen sicherzustellen, sind hierfür zum einen organisatorische Überlegungen in Form von Betriebs- und Rollenkonzepten zu entwickeln. Zum anderen ist die Unterstützung durch das Management unabdingbar.

5 Zusammenfassung und Ausblick

Zielsetzung dieses vorliegenden Arbeitsberichts ist es, ein Konzept zur Sicherstellung der Datenqualität für Data-Warehouse-Systeme zu entwickeln. Ausgangspunkt hierfür stellt eine aktuelle Umfrage von Unternehmensvertretern dar, aus der hervorgeht, dass u. a. aufgrund von fehlenden operationalisierbaren Konzepten ein umfassendes Datenqualitätsmanagement in den Unternehmen bisher noch nicht realisiert wurde.

Dieser Arbeitsbericht soll einen Beitrag zur Überwindung dieses Defizits leisten, indem nach der Klärung der definitorischen Grundlagen (siehe Kapitel 1) ein Datenqualitätsmodell für Data-Warehouse-Systeme entwickelt wird. Ein wesentliches Gestaltungsmerkmal des hier propagierten Datenqualitätskonzepts ist dessen umfassende Berücksichtigung des gesamten Data-Warehouse-Systems und damit die holistische Herangehensweise an die Datenqualitätsproblematik. Dieses Grundkonzept verdeutlicht, dass partielle Betrachtungen nicht ausreichend sind, sondern der ganzheitliche Ansatz eine essentielle Voraussetzung für ein Datenqualitätsmanagement für Data-Warehouse-Systeme darstellt. Ein weiterer Schwerpunkt des Ansatzes liegt insbesondere auf dem Aspekt der Operationalisierung und Anwendbarkeit. Dementsprechend wird nach einer kurzen theoretischen Fundierung des Datenqualitätsmodells (siehe Kapitel 4.1) dessen Umsetzung anhand eines ausführlichen, mit Praxisvertretern erarbeiteten Beispiels detailliert aufgezeigt und konkretisiert (siehe Kapitel 4.2). Im Anschluss daran wird die implizit zugrunde liegende Reihenfolge der Massnahmen durch ein phasenorientiertes Vorgehensmodell expliziert (siehe Kapitel 4.3). Hierbei werden die durchzuführenden Arbeitspakete innerhalb der Phasen eingehend beschrieben und somit ein Leitfaden zur Implementierung des Datenqualitätsmodells im Besonderen für die praxisorientierte Leserschaft zur Verfügung gestellt.

Im Bereich des Datenqualitätsmanagement für Data-Warehouse-Systeme besteht weiterer Forschungsbedarf. Das hier vorgestellte Datenqualitätsmodell ist durch konkrete Umsetzungen in der Praxis zu validieren und zu ergänzen. Insbesondere ist das Modell um organisatorische Aspekte zu erweitern, ausserdem müssen standardisierte Methoden und Vorgehensweisen für die einzelnen Phasen entwickelt werden. Des Weiteren ist die Formulierung eines Business Case für das Datenqualitätsmanagement zur weiteren Verbreitung des Datenqualitätsgedankens in der Praxis unabdingbar. Das Kompetenzzentrum Data Warehousing 2 sieht vorrangig den Bereich des Metadatenmanagements und, damit verbunden, die Einordnung und Modellierung von Datenqualitätsaspekten im Rahmen eines zentralen Metadatenmanagementkonzepts als Forschungsschwerpunkt an.

Literatur

[Beiersdorf 1995]

Beiersdorf, Holger: Informationsbedarf und Informationsbedarfsermittlung im Problemlösungsprozess "Strategische Unternehmungsplanung", Hampp, München, 1995.

[Bleicher 1992]

Bleicher, Knut: Das Konzept integriertes Management, Campus, Frankfurt a. M. et al., 1992.

[DIN 1995]

o. V.: Qualitätsmanagement und Statistik : Verfahren 3 : Qualitätsmanagementsysteme : Normen, DIN Deutsches Institut für Normung (Hrsg.), Beuth, Berlin, 1995.

[English 1999]

English, Larry P.: Improving Data Warehouse and Business Information Quality, Wiley, New York et al., 1999.

[Gabler Wirtschaftsinformatik Lexikon 1997]

o. V.: Gabler Wirtschaftsinformatik Lexikon, Gabler, Wiesbaden, 1997.

[Garvin 1998]

Garvin, David A.: What does ‘Product Quality’ really mean?, in: Sloan Management Review, Fall, 1998, S. 25-43.

[Häussler 1998]

Häussler, Christa: Datenqualität, in: Wolfgang, Martin (Hrsg.): Data Warehousing, ITP GmbH, Bonn, 1998, S. 75-89.

[Haux et al. 1998]

Haux, R., Lagemann, A., Knaup, P., Schmücker, P., Winter, A.: Management von Informationssystemen: Analyse, Bewertung, Auswahl, Bereitstellung und Einführung von Informationssystemkomponenten am Beispiel von Krankenhausinformationssystemen, Teubner, Stuttgart, 1998.

[Helfert 2000a]

Helfert, Markus: 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 2000b]

Helfert, Markus: Massnahmen und Konzepte zur Sicherung der Datenqualität, in: Jung, Reinhard (Hrsg), Winter, Robert (Hrsg.): Data-Warehousing-Strategie: Erfahrungen, Methoden, Visionen, Springer, Berlin et al., 2000, S. 61-77.

[Holten 1999]

Holten, Roland: Entwicklung von Führungsinformationssystemen: Ein methodenorientierter Ansatz, Dt. Univ.-Verl., Wiesbaden, 1999.

[Huang, Lee, Wang 1999]

Huang, Kuan-Tsae, Lee, Yang W., Wang, Richard Y.: Quality Information and Knowledge, Prentice Hall, Upper Saddle River, NJ, 1999.

[Imai 1993]

Imai, Masaaki: Kaizen : Der Schlüssel zum Erfolg der Japaner im Wettbewerb, Langen-Müller/Herbig, München, 1993.

[Jarke et al. 2000]

Jarke, Matthias, Lenzerini, Maurizio, Vassiliou, Yannis, Vassiliadis, Panos: Fundamentals of data warehouses, Springer, Berlin et al., 2000.

[Jarke, Vassiliou 1997]

Jarke, Matthias, Vassiliou, Yannis: Foundations of Data Warehouse Quality – A Review of the DWQ Project, in: Strong, Diane M. (Hrsg.), Kahn, Beverly K. (Hrsg.): Proceedings of the 2nd International Conference on Information Quality, Cambridge, MA, 1997, S. 299-313.

[Jeusfeld, Jarke, Quix 1999]

Jeusfeld, Manfred A., Jarke, Matthias, Quix, Christoph: Qualitätsanalyse im Data Warehousing, in: EMISA Forum, Nr. 1, Vol. 9, 1999, S. 21-30.

[Meyer 2000]

Meyer, Markus: Organisatorische Gestaltung des unternehmensweiten Data Warehousing, Difo-Druck OHG, Bamberg, 2000.

[Müller 2000]

Müller, Jochen: Transformation operativer Daten zur Nutzung im Data Warehouse, Diss. Universität Bochum, 2000.

[Mutscheller 1996]

Mutscheller, Andreas M.: Vorgehensmodell zur Entwicklung von Kennzahlen und Indikatoren für das Qualitätsmanagement, Diss. Universität St. Gallen, 1996.

[Naumann, Rolker 1999]

Naumann, Felix, Rolker, Claudia: Do metadata models meet IQ requirements?, in: Proceedings of the international Conference on Information Quality (IQ), Cambridge, MA, 1999, S. 99-114.

[Picot, Reichwald, Wigand 2001]

Picot, Arnold, Reichwald, Ralf, Wigand, Rolf T.: Die grenzenlose Unternehmung: Information, Organisation, Management, 4. Auflage, Gabler, Wiesbaden, 2001.

[Schnauber et al. 1997]

Schnauber, Herbert, Grabowski, Sabine, Schlaeger, Sabine, Zülch, Joachim: Total Quality Learning: Ein Leitfaden für lernende Unternehmen, Springer, Berlin et al., 1997.

[Seghezzi 1996]

Seghezzi, Hans D.: Integriertes Qualitätsmanagement – das St. Galler Konzept, Hanser, München, Wien, 1996.

[Töpfer, Mehdorn 1994]

Töpfer, Armin, Mehdorn, Hartmut: Total Quality Management: Anforderungen und Umsetzungen im Unternehmen, 3. Auflage, Luchterhand, Neuwied, Kriftel, Berlin, 1994.

[Wallmüller 1990]

Wallmüller, Ernest: Software-Qualitätssicherung in der Praxis, Hanser, München et al., 1990.

[Wang, Storey, Firth 1995]

Wang, Richard Y., Storey, Veda C., Firth, Christopher P.: A framework for analysis of data quality research, in: IEEE Transactions on Knowledge and Data Engineering, Nr. 4, Vol. 7, 1995, S. 623-640.

[Wang, Strong 1996]

Wang, Richard Y., Strong, Diane M.: Beyond Accuracy: What Data Quality Means to Data Consumers, in: Journal of Management Information Systems, Nr. 4, Vol. 12, 1996, S. 5-33.

[Wolf 1999]

Wolf, Peter: Konzept eines TQM-basierten Regelkreismodells für ein "Information Quality Management" (IQM), Verlag Praxiswissen, Dortmund, 1999.