

# **Datenqualitätsmanagement für Data-Warehouse-Systeme – Technische und organisatorische Realisierung am Beispiel der Credit Suisse**

**Marcel Winter**

Credit Suisse

**Clemens Herrmann**

Universität St. Gallen

**Markus Helfert**

Dublin City University

*Ein kritischer Erfolgsfaktor zur dauerhaften Etablierung von Data-Warehouse-Systemen in Unternehmen ist eine ausreichende Qualität der dadurch zur Verfügung gestellten Daten. Um ein hohes Datenqualitätsniveau langfristig zu sichern, reichen punktuelle Datenbereinigungsmassnahmen nicht aus. Stattdessen gilt es, ein umfassendes Datenqualitätsmanagement einzuführen, welches kontinuierlich die Qualität der Daten überwacht und bei Qualitätsabweichungen Massnahmen zur Beseitigung der Fehlerursachen einleitet. Am Beispiel der Credit Suisse wird ein solches Datenqualitätsmanagement sowohl aus technischem wie auch aus organisatorischem Blickwinkel betrachtet und detailliert erläutert.*

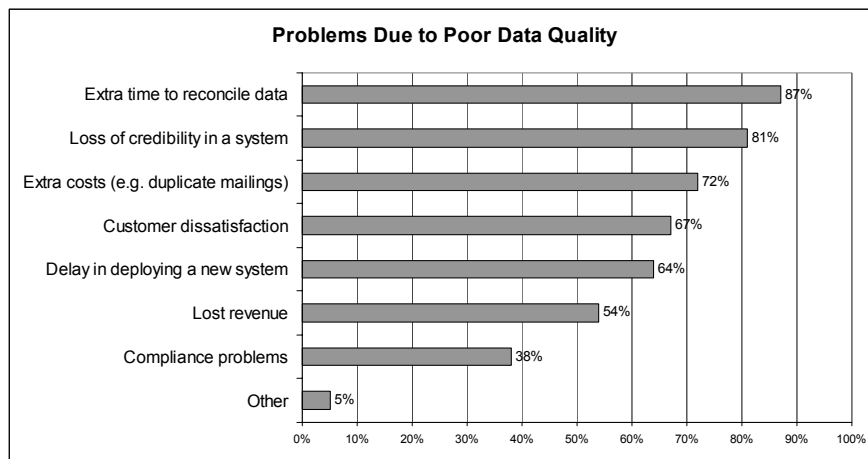
## **1 Motivation**

Die Qualität von Daten und Informationen spielt in der heutigen Informationsgesellschaft eine immer wichtigere Rolle (vgl. Wolf 1999, S. 7f.). Für die dauerhafte Etablierung eines Data-Warehouse-Systems im Unternehmen stellt die Qualität der Daten mittlerweile eine unabdingbare Notwendigkeit dar (vgl. English 1999, S. 4). Als Folgen unzureichender Datenqualität wurden in einer Studie des TDWI<sup>1</sup> der zusätz-

---

<sup>1</sup> Die Studie des TDWI (The Data Warehouse Institute) zum Themenbereich Datenqualität wurde im Jahr 2001 durchgeführt und basiert im Wesentlichen auf der Auswertung von 647 Fragebögen (vgl. Eckerson 2002).

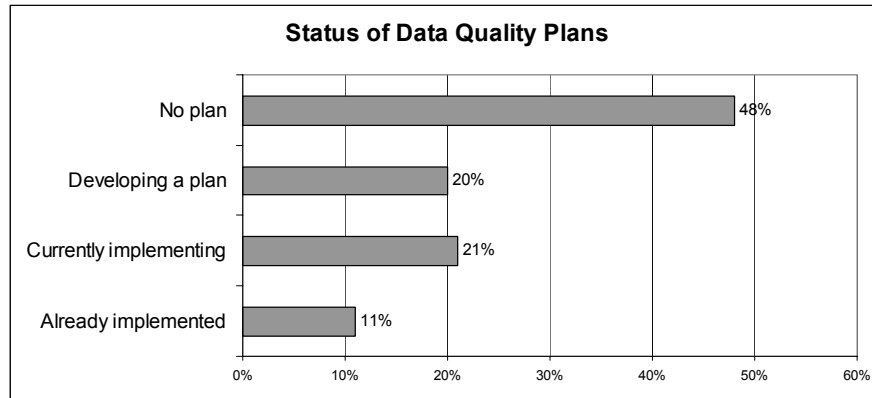
liche Zeitaufwand zur Integration bzw. Konsistenzerhaltung der Daten, der Verlust an Vertrauen in das Data-Warehouse-System und die zusätzlich entstehenden Folgekosten durch bspw. mehrfaches Versenden von Werbung am häufigsten genannt (vgl. Abb. 1). Weitere Konsequenzen, die aus einer mangelnden Datenqualität erwachsen können, sind Kundenunzufriedenheit, Verzögerungen bei der Einführung neuer Systeme bspw. durch Ungewissheit über die zugrunde liegende Datenqualität, Ertragsverluste z. B. durch fehlerhafte Rechnungen sowie Probleme bei der Erfüllung von gesetzlichen Auflagen z. B. bei der Bilanzerstellung.



**Abb. 1:** Probleme aufgrund unzureichender Datenqualität (Eckerson 2002, S. 10)

Trotz der negativen Auswirkungen schlechter Datenqualität auf das gesamte Unternehmen besaßen 48% der befragten Unternehmen noch keine Strategie, dieses Problem anzugehen und nur 11% hatten bereits umfassende Massnahmen zur Verbesserung der Datenqualität ergriffen (vgl. Abb. 2). Daher soll in diesem Artikel die erfolgreiche Umsetzung eines ganzheitlichen Datenqualitätsmanagements bei der Credit Suisse aufgezeigt werden und so als Beispiel für andere Unternehmen dienen, die eine systematische Verbesserung ihrer Datenqualität anstreben.

Nach einer Diskussion des Begriffs Datenqualität und dessen Konkretisierung anhand von Qualitätsmerkmalen wird in Kapitel 3 ein Konzept eines ganzheitlichen Datenqualitätsmanagements vorgestellt. Kapitel 4 beschreibt eingehend die Realisierung des Datenqualitätsmanagements bei der Credit Suisse sowohl aus technischer als auch aus organisatorischer Perspektive. Der Artikel schliesst mit einer Zusammenfassung und einem Ausblick auf zukünftige Schritte.



**Abb. 2:** Datenqualitätsmanagement in der Praxis (Eckerson 2002, S. 6)

## 2 Datenqualität

### 2.1 Ansätze aus der Literatur

Der Themenbereich Datenqualität im Data Warehousing wird bereits von einigen Autoren behandelt. Im Folgenden seien einige ausgewählte Ansätze genannt.<sup>2</sup>

WAND und WANG (vgl. Wand, Wang 1996) fokussieren ihre Betrachtung auf die Entwicklung und den Betrieb eines Informationssystems. Datenqualitätsmängel treten bei Inkonsistenzen zwischen der Sicht auf das Informationssystem und der Sicht auf die reale Welt auf. Aus diesen Abweichungen können vier innere Datenqualitätsmerkmale abgeleitet werden: Vollständigkeit, Eindeutigkeit, Bedeutung und Korrektheit. WAND und WANG betrachten in ihrem Ansatz jedoch nicht die funktionalen Anforderungen der Endbenutzer an das Informationssystem.

ENGLISH (vgl. English 1999) unterscheidet zwischen Datendefinitions- und Architekturqualität, der Qualität der Datenwerte sowie der Qualität der Datenpräsentation. Diesen Kategorien ordnet er Merkmale zur detaillierteren Beschreibung zu. Er geht jedoch nicht detailliert auf Überschneidungen und Beziehungen zwischen den einzelnen Merkmalen und den übergeordneten Kategorien ein.

Im Rahmen einer empirischen Untersuchung von WANG und STRONG (vgl. Wang, Strong 1996) zur Bestimmung allgemeiner Datenqualitätsmerkmale werden vier Kategorien (Innere Datenqualität, kontextabhängige Datenqualität, Darstellungs-

<sup>2</sup> Ein ausführlicher Vergleich des State-of-the-art im Bereich der Datenqualität ist zu finden in (Helfert 2002, S. 68-79 und S. 121-130).

qualität und Zugangsqualität) mit jeweils unterschiedlichen Qualitätsmerkmalen ermittelt. Die empirische Untersuchung lief in zwei Stufen ab, wobei die Hauptanalyse auf 355 Fragebögen basiert.

JARKE et al. (vgl. Jarke et al. 1999; Jarke, Vassiliou 1997) gliedern die Datenqualitätsmerkmale anhand der drei Prozesse Entwicklung und Verwaltung, Softwareimplementierung sowie Datennutzung. Die sich hieraus ergebenden Merkmale werden weiter anhand von zugeordneten, auf die Datenwerte bezogenen Kriterien verfeinert.

HINRICHS (vgl. Hinrichs 2001; Hinrichs 2002, S. 29 ff.) charakterisiert Datenqualität ausgehend von einer empirischen Erhebung von STRONG et al. (vgl. Strong et al. 1997) anhand der Kategorien Glaubwürdigkeit, Nützlichkeit, Interpretierbarkeit und Schlüsselintegrität. Diesen werden insgesamt 13 unterschiedliche Datenqualitätsmerkmale zugeordnet. Es wird ausdrücklich auf das Problem der Vollständigkeit und Überschneidungsfreiheit einer solchen Klassifizierung hingewiesen.

Nachfolgend soll eine auf den Anforderungen des Datenqualitätsmanagements aufbauende Definition des Begriffs Datenqualität mit dazugehörigen Datenqualitätskriterien gegeben werden, die den Begriff näher beschreiben und operationalisieren sollen.

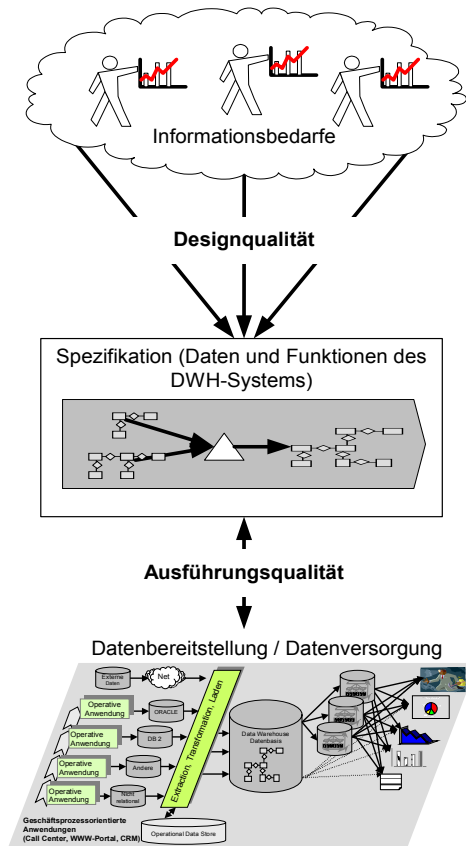
## 2.2 Das 3-Ebenen-Modell der Datenqualität

Der Qualitätsbegriff aus der industriellen Fertigung kann auf verschiedenen Ebenen betrachtet und in drei Sichten unterteilt werden (vgl. Helfert 2002, S. 66 ff.):

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

Der anwenderbezogene Qualitätsansatz bezieht sich auf eine externe Sicht und stellt den Endbenutzer mit seinen Anforderungen in den Vordergrund. Im Data Warehousing sind dies vor allem die Informationsbedarfe der Datenverwender. Von diesen Qualitätsforderungen werden eine Produktspezifikation und ein Produktionsplan abgeleitet. Die konzeptionelle Spezifikation eines Data-Warehouse-Systems mit dessen Daten und Funktionen kann auf dieser Ebene eingeordnet werden. Diese Spezifikation bildet die Grundlage für die Gestaltung der Produktionsprozesse. Hierunter werden im Data Warehousing die Datenbereitstellungs- und Datenversorgungsprozesse verstanden. Auf Grundlage dieser Qualitätsebenen lässt sich Qualität, wie in Abb. 3 dargestellt, grundsätzlich in zwei Faktoren untergliedern (vgl. Seghezzi 1996, S. 12 und S. 26):

- Designqualität
- Ausführungsqualität



**Abb. 3:** Qualitätssichten (vgl. Helfert 2002, S. 67)

Zunächst werden die Anforderungen der Endbenutzer erfasst und in Form einer Spezifikation konkretisiert. Es ist die Frage nach den geeigneten Produkteigenschaften zu beantworten. Es sind die Eigenschaften auszuwählen, welche die Bedürfnisse der Anwender am Besten erfüllen und so Kundenzufriedenheit erzeugen. In einer Datenbank werden durch Datenschemata Entitäten und Eigenschaften der zu erfassenden Datenobjekte festgelegt. Diese Datenschemata können so als Spezifikation eingestuft werden (vgl. Juran 1999, S. 1).

Sind die Anforderungen erfasst und in einer Spezifikation festgelegt, ändert sich die Zielsetzung des Qualitätsmanagements auf die Einhaltung der in der Spezifikation festgelegten Qualitätsforderungen. Nicht die Bedürfnisse der Anspruchsgruppen, sondern Konformität und fehlerfreie Erfüllung der in Spezifikationen niedergeschriebenen Anforderungen ist das Ziel (vgl. Juran 1999, S. 2). Die laufenden Prozesse sind dahingehend zu kontrollieren. *Designqualität* bezieht sich auf die Erfassung von Qualitätsforderungen aus Anwendersicht in einer Spezifikation, während *Ausführungsqualität* die Einhaltung der von den Anwendern festgelegten Spezifika-

tion umfasst. Eine unzureichende Gesamtqualität kann sowohl in einer mangelhaften Design- als auch in einer nicht ausreichenden Ausführungsqualität begründet sein.

Die Trennung in Design- und Ausführungsqualität lässt sich auf den Datenqualitätsbegriff übertragen. Es erfolgt dementsprechend eine Unterscheidung nach Datenschema und Datenwerten. Diese beiden übergeordneten Kriterien lassen sich weiter verfeinern in Unterkategorien und Merkmale. Eine ausführliche Beschreibung dieser Merkmale erfolgt in Tabelle 1 und 2. Der Begriff der Datenqualität wird dadurch konkretisiert und definiert.

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

**Tab. 1:** Qualitätsmerkmale bezogen auf das Datenschema (Helfert 2002, S. 83)

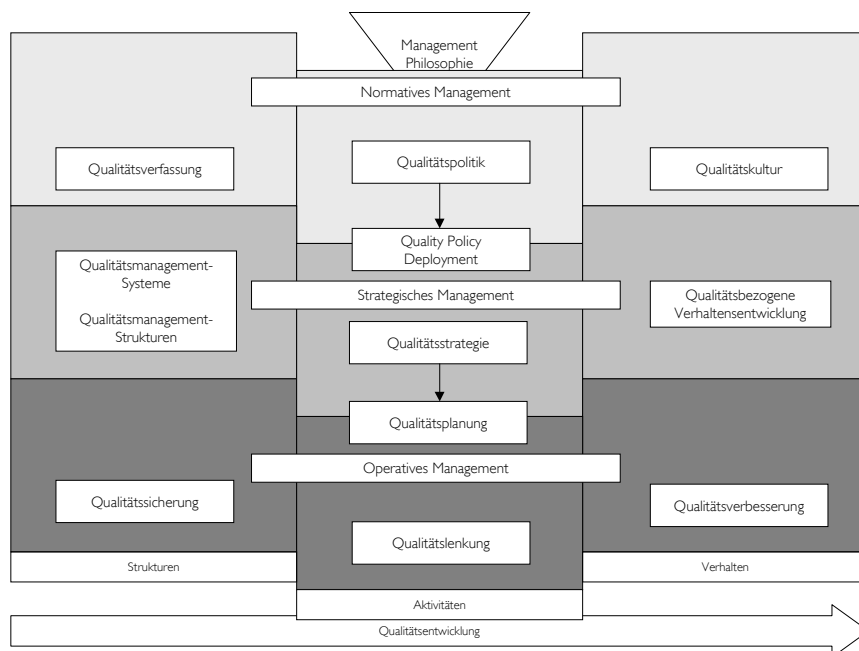
Kategorie	Merkmal	Beschreibung
Glaubwürdigkeit	Korrektheit	Die Daten stimmen inhaltlich mit der Datendefinition überein und sind empirisch korrekt.
	Datenherkunft	Die Datenherkunft und die vorgenommenen Datentransformationen sind bekannt.
	Vollständigkeit	Alle Daten sind gemäss Datenmodell erfasst.
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

**Tab. 2:** Qualitätsmerkmale bezogen auf die Datenwerte (Helfert 2002, S. 84)

### 3 Management der Datenqualität

#### 3.1 Operatives Qualitätsmanagement

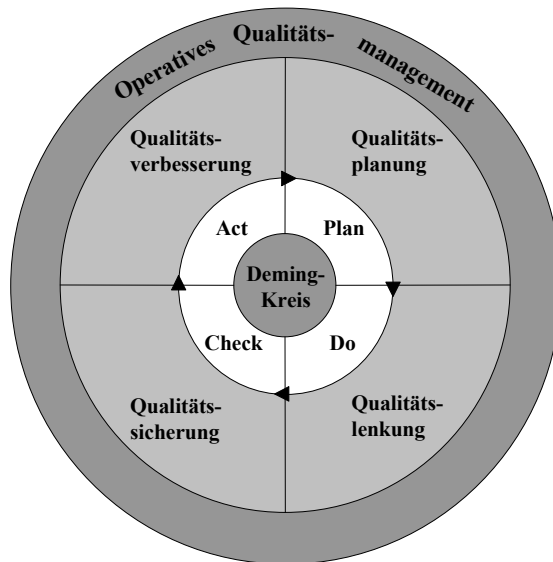
Nach Klärung des Begriffs der Datenqualität soll nun auf das Qualitätsmanagement näher eingegangen werden. Laut DIN ISO 8402 umfasst Qualitätsmanagement alle Tätigkeiten der Gesamtführungsaufgabe, welche die Qualitätspolitik, die Qualitätsziele und die Verantwortungen für die Qualität festlegt (vgl. o. V. 1995). Die Elemente lassen sich grob, wie in Abb. 4 dargestellt, anhand des St. Galler Managementkonzepts (vgl. Bleicher 1992) einordnen.



**Abb. 4:** 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 einerseits durch die Strukturen unterstützt und andererseits durch das Verhalten der Führungskräfte und Mitarbeiter geprägt wird. Die dritte Dimension betrifft den zeitlichen Aspekt Qualitätsentwicklung (vgl. Seghezzi 1996, S. 48 ff.). 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 Funkti-

onsbereiche in den prozessorientierten Qualitätsansatz von DEMING ein, wie Abb. 5 verdeutlicht.

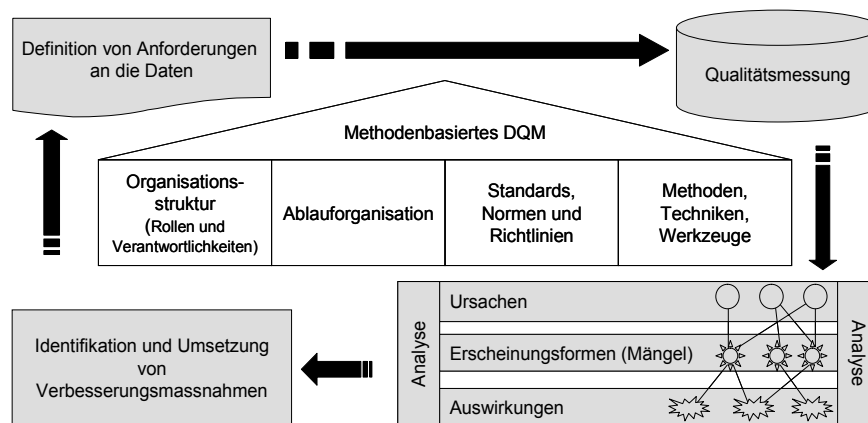


**Abb. 5:** Operatives Qualitätsmanagement nach dem Deming-Kreis (Seghezzi 1996, S. 53)

Das operative Qualitätsmanagement ist anhand der Geschäftsprozesse 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.



**Abb. 6:** Ganzheitliches Datenqualitätsmanagement (vgl. Helfert 2000, S. 68)

Ausgehend von obigem Verständnis des operativen Qualitätsmanagements soll im Folgenden ein Ansatz für ein ganzheitliches Datenqualitätsmanagement dargestellt werden.

### 3.2 Ganzheitliches Datenqualitätsmanagement

Das Datenqualitätsmanagement zielt auf eine kontinuierliche Verbesserung der Datenqualität ab und kann in vier Hauptprozesse untergliedert werden (vgl. Abb. 6). Im Anschluss an die Definition von Qualitätsanforderungen an die Daten, die sowohl technischen als auch fachlichen Charakter haben können, werden Qualitätsmessungen durchgeführt, deren Ergebnis Qualitätskennzahlen über den untersuchten Datenbestand sind. Darauf aufbauend werden Datenqualitätsmängel sowie deren Ursachen und Auswirkungen analysiert, so dass die Wirkungszusammenhänge bekannt sind. In der letzten Phase werden potenzielle Verbesserungsmaßnahmen auf Basis einer Problemanalyse identifiziert und umgesetzt. Diese vier Hauptaktivitäten stellen keinen einmalig zu durchlaufenden Prozess dar, sondern sind vielmehr

als iterativer Kreislauf zu verstehen, der eine kontinuierliche Datenqualitätsverbesserung sicherstellen soll (vgl. English 1999, S. 70 ff.; Helfert 2000, S. 67).

Zur erfolgreichen Umsetzung der Hauptprozesse des Datenqualitätsmanagements im Unternehmen sind die folgenden drei Aspekte zwingend zu berücksichtigen. (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.

Nach der Darstellung der begrifflichen Grundlagen und der fundamentalen Konzepte soll im Folgenden eine Konkretisierung anhand des Datenqualitätsmanagements bei der Credit Suisse erfolgen. Hierzu wird sowohl auf die technische Realisierung als auch auf die organisatorische Einbettung eingegangen.

## **4 Datenqualitätsmanagement der Credit Suisse**

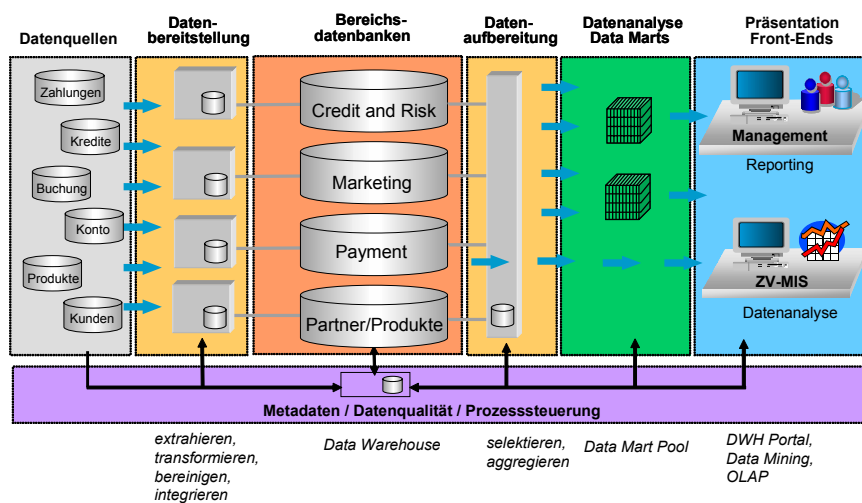
### **4.1 Data-Warehouse-Architektur der Credit Suisse**

Die Data-Warehouse-Architektur der Credit Suisse ist unterteilt in unterschiedliche Ebenen, die sich an der Data-Warehouse-Referenzarchitektur orientieren (vgl. Abb. 7):

- Datenquellen: Hierunter werden transaktionelle Systeme verstanden, die für das Data Warehouse relevante Daten enthalten. Die Daten werden für das Data Warehouse in Bereiche eingestellt, die als Feeder bezeichnet werden.
- Datenbereitstellung: Aus den Feedern werden die Daten extrahiert, auf der sog. Staging Area temporär zwischengespeichert und für das Data Warehouse aufbe-

reitet. Hierunter fallen Operationen wie Transformation, Integration und Bereinigung.

- Bereichsdatenbanken: Diese Datenbanken sind nach bankfachlichen Aspekten getrennte, historisierte Datentöpfe.
- Data Marts: Die Abfragen der Endbenutzer werden im Wesentlichen auf den Data Marts ausgeführt. Hierunter werden Modelle der Bereichsdatenbanken verstanden, die für bestimmte Analysewerkzeuge optimiert sind.
- Präsentation-Front-Ends: Auf dieser Ebene werden den Endbenutzern die Abfrageergebnisse entweder grafisch oder textuell präsentiert.



**Abb. 7:** Data-Warehouse-Architektur der Credit Suisse

Zusätzlich zu oben beschriebenen Ebenen existiert eine Metadatenverwaltung, die zur Steuerung des gesamten Data-Warehouse-Systems und zur Verwaltung aller relevanten Metadaten eingesetzt wird. Auch die Daten des Datenqualitätsmoduls sowie das Datenqualitätsmodul selbst sind auf dieser Ebene einzuordnen.

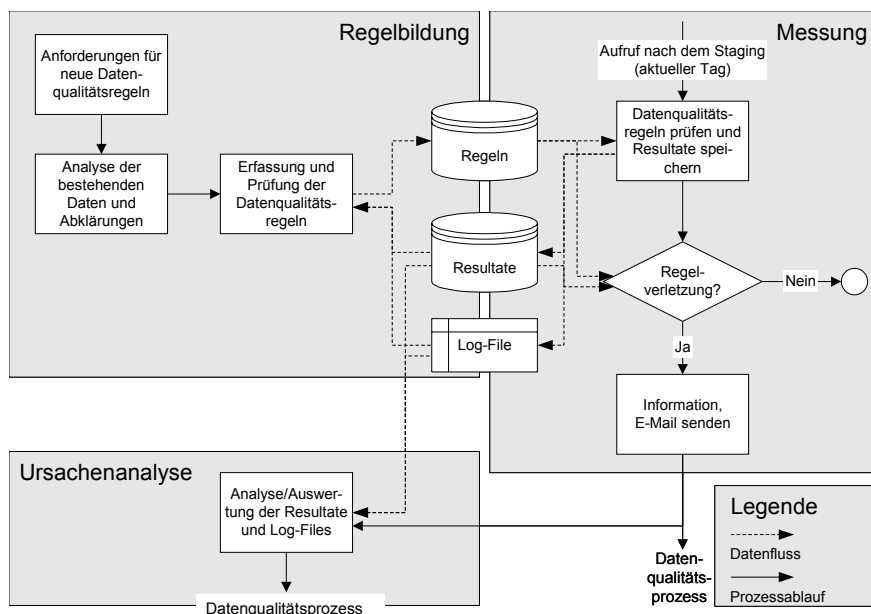
## 4.2 Technische Realisierung

Das Datenqualitätsmodul der Credit Suisse ist eingebettet in die Metadatenverwaltung. Es nutzt und erzeugt folgende Daten und Metadaten:

- Datenqualitätsregeln: Hierbei handelt es sich um Regeln, die auf den zu prüfenden Datenbestand angewendet werden können und feingranulare Qualitätsaussagen liefern. Es sind sowohl Regeln zur Überprüfung technischer als auch fachlicher Sachverhalte denkbar. Den einzelnen Regeln werden auch durchzuführende Aktionen bei Regelverletzungen zugeordnet.

- Datenqualitätsaussagen: Diese Metadaten resultieren aus der Anwendung der Qualitätsregeln auf die Daten. Anhand dieser Ergebnisse lassen sich Aussagen über die Qualität der Daten machen.
- Logfile als Fehlerprotokoll: Treten Regelverletzungen auf, so werden die fehlerhaften Datensätze in ein Logfile geschrieben, um die spätere Problemanalyse zu vereinfachen.

Den grundsätzlichen Aufbau des Datenqualitätsmoduls zeigt Abb. 8. Es werden die Bereiche Regelbildung, Messung und Ursachenanalyse unterschieden, auf die im Folgenden näher eingegangen werden soll.



**Abb. 8:** Datenqualitätsmodul der Credit Suisse

### Regelbildung

Die Spezifikation und Erfassung der Regelmengen zur Überprüfung der Datenqualität sind die wichtigsten und zugleich zeitaufwendigsten Aktivitäten des Datenqualitätsmanagements. Das Vorgehen zur Regelbildung gliedert sich in drei Schritte. Nach der Abklärung der Anforderungen an die neuen Datenqualitätsregeln werden die bestehenden Daten in Bezug auf die im ersten Schritt spezifizierten Anforderungen geprüft und analysiert. Anschliessend erfolgt die Erfassung der Datenqualitätsregeln in einer Regelmenge.

Regeln können in Form von SQL-Skripts spezifiziert und abgelegt werden. Pro Regel wird festgelegt, wie das Resultat bewertet wird und was nach der Prüfung ge-

schieht (z. B. Versendung einer E-Mail oder SMS). Die Regeln können jederzeit ergänzt oder verändert werden.

Die Basis zur Regelspezifikation können Integritätsregeln für Datenbanken bilden. Hierbei werden statische, transitionale und dynamische Bedingungen unterschieden. Erstere schränken einen einzelnen Datenbankzustand ein, wohingegen transitionale Bedingungen auf zwei Datenbankzustände bezogen sind. Es werden mögliche Zustandsübergänge eingeschränkt. Dynamische Integritätsbedingungen stellen eine Verallgemeinerung der transitionalen dar, indem Folgen von Zustandsübergängen eingeschränkt werden. (vgl. Heuer, Saake 2000, S. 496; Vossen 2000, S. 148f.). Eine weitere Möglichkeit zur Unterscheidung von Integritätsbedingungen stellt die Granularität der Bezugsobjekte dar. Bedingungen können sich auf Attribute, Tupel, Relationen oder Datenbanken beziehen (vgl. Heuer, Saake 2000, S. 507f.). Beispiele für Integritätsbedingungen sind:

- Ober- und Untergrenzen für Werte,
- Menge möglicher Werte,
- Pflichtfelder bzw. Ausschluss der Verwendung von Nullwerten,
- Schlüsselbedingungen,
- Fremdschlüsselbeziehungen und
- Aggregatbedingungen (z. B. Ober- und Untergrenze für die Summe der Guthaben).

Neben diesen Integritätsbedingungen sind noch weitere Regeln denkbar, wie z. B. (vgl. Elmasri, Navathe 1994, S. 149):

- Die Anzahl der Tupel einer Relation steht in Beziehung zur Anzahl der Tupel einer anderen Relation (z. B. die Anzahl der Konten ist grösser als die Anzahl der Kunden).
- Ein Wert ist zeitinvariant (z. B. das Geburtsdatum eines Kunden).
- Ein Attributwert zeigt im Zeitablauf ein ähnliches Verhalten wie ein zweiter Attributwert (z. B. das Kreditvolumen verhält sich linear zur Anzahl der Kunden).

Ein vereinfachtes Beispiel aus der Regelbasis zur Überprüfung der Anzahl der neu hinzugekommenen Zeilen nach einem Load stellt nachfolgender SQL-Ausdruck dar. Die Tabelle wird einmal im Monat aktualisiert und alle neu hinzugekommenen Tupel erhalten als Zeitstempel das Ladedatum, anhand dessen die Überprüfung stattfindet:

```
SELECT count (*)
FROM table_x a
WHERE a.date_per = to_date('31.01.2002', 'dd.mm.yyyy')
```

Aus der Erfahrung ist dem Fachexperten beispielsweise bekannt, dass zu `table_x` pro Monat ca. 1000 neue Tupel hinzukommen. Weicht das Ergebnis des SQL-Statements jedoch deutlich von diesem Wert ab, so muss eine Fehlerüberprüfung stattfinden.

Ein weiteres Beispiel stellt die folgende Regel dar, die alle Konten zählt, für die das „closed flag“ gesetzt ist, aber für die kein „closing date“ angegeben ist:

```
SELECT count(account_id)
FROM accounts
WHERE substr(appl_flags_1,8,1) = '1' AND
      account_closing_date is NULL
```

Das Ergebnis dieser Überprüfung muss Null ergeben, da es keine geschlossenen Konten ohne Enddatum geben darf. Grundsätzlich kann das Datenqualitätsmodul alle Regeln verarbeiten, die in Form von SQL-Statements spezifizierbar sind.

Die Oberfläche zur Verwaltung der Regeln zeigt Abb. 9. Einzelne Regeln können zu Regelmengen zusammengefasst werden, die jeweils abgeschlossene Sachverhalte überprüfen. Für jede Regel kann die gewünschte Wertemenge festgelegt werden, welche bei Fehlerfreiheit generiert wird. Hierbei kann bspw. unterschieden werden zwischen einem einzigen Wert und einem Intervall, in dessen Grenzen sich das Ergebnis der Qualitätsüberprüfung befinden sollte.

### *Messung*

Sobald Regeln existieren, kann eine Datenqualitätsmessung durchgeführt werden. Hierbei werden die Regeln auf den Datenbestand angewendet und die Messresultate gespeichert. Des Weiteren werden bei Regelverletzungen die falschen Datensätze in einem Logfile abgelegt und es erfolgt eine Benachrichtigung des Datenqualitätsverantwortlichen bzw. des Entwicklers.

Das Datenqualitätsmodul wird zurzeit täglich auf der Ebene der Datenbereitstellung (vgl. Abb. 9) zur Überprüfung der Datenqualität der Extrakte auf der Staging Area eingesetzt. Diese Qualitätskontrolle stellt den letzten Job dar, bevor die Daten endgültig in die Bereichsdatenbanken geladen werden. Beim Resultat pro Regelprüfung wird anhand von Kennzahlen mit dazugehörigen Massangaben und Bandbreiten zwischen drei unterschiedlichen Qualitätszuständen unterschieden. Entweder sind die Daten nutzbar (grün), eingeschränkt nutzbar (gelb) oder nicht zu verwenden (rot). Diese Bewertung kann sich z. B. an der Anzahl oder der prozentualen Menge fehlerhafter Records ausrichten. Liegt die Qualität im gelben oder roten Bereich, so werden alle fehlerhaften Records ins Log-File geschrieben (Regel, Datum, Schlüssel, wichtige Felder), damit sie analysiert und später richtig nachgeliefert werden können.

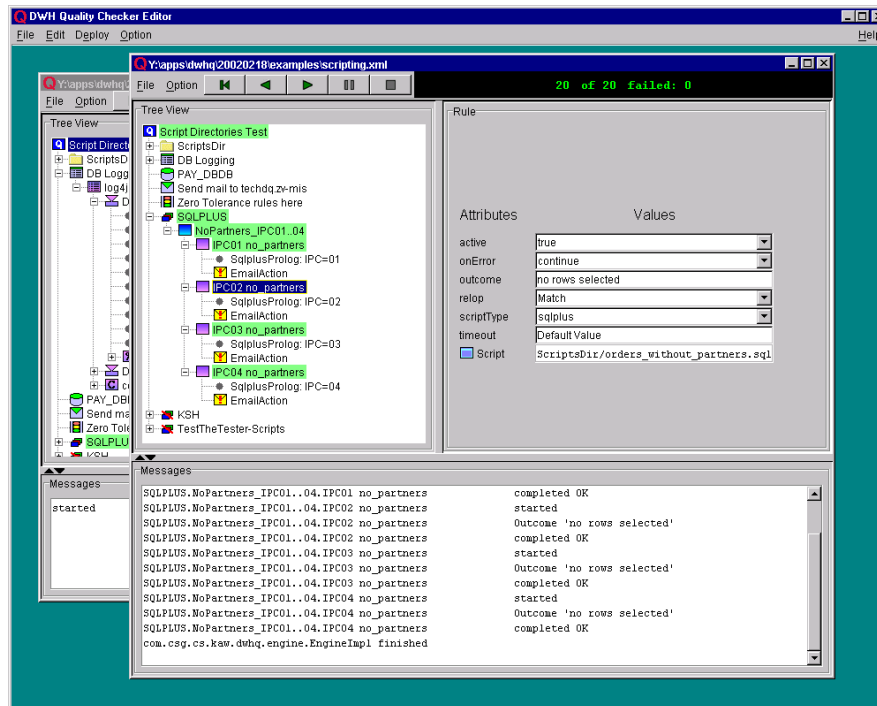


Abb. 9: Screenshot der Regelverwaltung

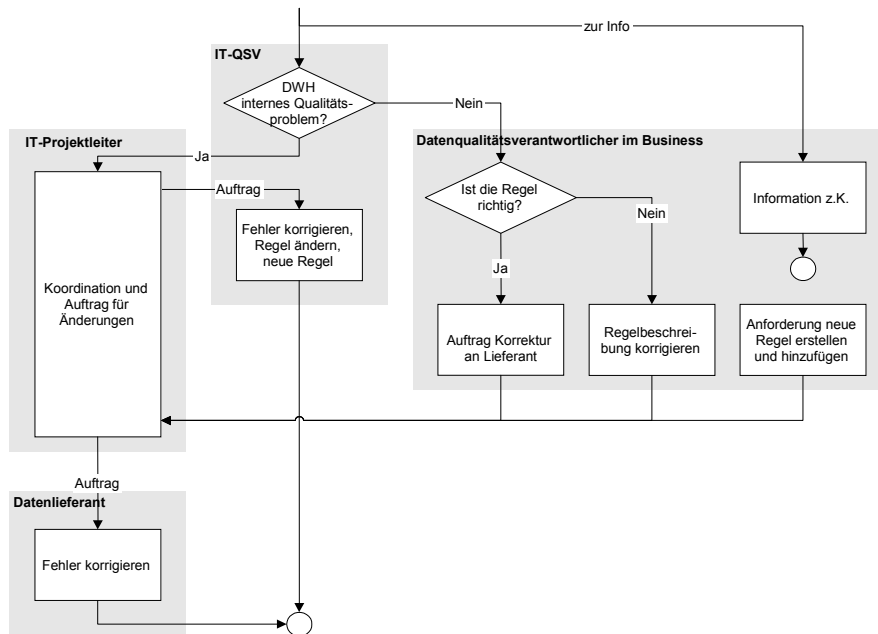
### Ursachenanalyse

In der abschließenden Ursachenanalyse werden die Messresultate und das Logfile zur eingehenden Analyse und Auswertung herangezogen, um eventuelle Massnahmen zur Qualitätsverbesserung zu identifizieren. Die Ursachenanalyse ist ein organisatorischer Prozess, der beim Auftreten von Fehlern angestoßen wird. Prozessverantwortlich hierfür sind zunächst die Datenqualitätsverantwortlichen der IT-Abteilung sowie der Fachabteilung. Eine genauere Erläuterung des Prozessablaufs erfolgt in Kapitel 4.3.

## 4.3 Organisatorische Einbettung

Neben der technischen Umsetzung wurde im Sinne eines ganzheitlichen Datenqualitätsmanagements auch ein organisatorischer Datenqualitätsprozess spezifiziert. Dieser wird entweder beim Auftreten von Datenqualitätsproblemen oder durch Anforderungen bzgl. neuer Regeln angestoßen und gibt sowohl die durchzuführenden Aktivitäten als auch die Verantwortlichkeiten vor (vgl. Abb. 10).

Die zentralen Rollen des Datenqualitätsprozesses besetzen zum einen der Datenqualitätsverantwortliche des IT-Bereichs und zum anderen der Datenqualitätsbeauftragte



**Abb. 10:** Datenqualitätsprozess der Credit Suisse

te des Fachbereichs. Wird ein Datenqualitätsproblem durch das Datenqualitätsmodul festgestellt, so werden beide Stellen informiert. Der DQ-Beauftragte im IT-Bereich steht jedoch in der Verantwortung, zunächst zu entscheiden, ob es sich um ein technisches Problem handelt, welches von der IT-Abteilung selbstständig korrigiert werden kann, oder ob die Fachabteilung hinzugezogen werden muss. Im ersteren Fall wird ein entsprechender Auftrag an den IT-Projektleiter gegeben, wohingegen im letzteren Fall der Fachbereich die betroffene Regel prüft und entweder die entsprechende fehlerhafte Regelbeschreibung korrigiert oder andernfalls einen Auftrag an den Lieferanten gibt, das dort vorliegende Problem zu beheben. Die zentrale Rolle zur Koordination aller Änderungsanforderungen hat der IT-Projektleiter inne. In Abhängigkeit von den Anforderungen ist dieser daher verantwortlich, Aufträge zur Fehlerbehebung an den Datenlieferanten oder die IT-Abteilung zu erteilen.

Nach der Beschreibung der technischen und organisatorischen Realisierung des Datenqualitätsmanagements bei der Credit Suisse sollen im Folgenden die Erfahrungen seit Einführung des Datenqualitätsmanagements geschildert werden.

#### 4.4 Erfahrungen

Die grössten Nutzenaspekte des Datenqualitätsmanagements liegen einerseits in der deutlichen Verbesserung der Qualität der Daten sowohl im Data Warehouse als auch in den Quellsystemen. Andererseits wird die Akzeptanz des Data-Warehouse-Systems bei den Endbenutzern gesteigert. Die Glaubwürdigkeit von Berichten und das

Vertrauen in vom Data Warehouse gelieferten Informationen wurden deutlich gestärkt.

Die Entwicklung des Datenqualitätsmoduls wurde in vier Monaten durchgeführt und erforderte einen Aufwand von 70 Personentagen. Das Budget verteilte sich wie folgt auf die unterschiedlichen Projektphasen:

- 50% Definition der Prüfungsregeln und Felderanalyse mit der Fachabteilung
- 30% Entwicklung des Datenqualitätsmoduls und Implementierung der Regeln
- 20% Testen und Einführung

Wie bereits aus der Budgetverteilung ersichtlich, waren insbesondere die Spezifikation betrieblicher Regeln durch die IT-Abteilung in Zusammenarbeit mit der Fachabteilung sowie die iterative Verfeinerung bzw. Prüfung der Regelmengen sehr zeit- und kostenintensiv. Kostentreiber waren sowohl die Identifikation der für das Datenqualitätsmanagement relevanten Felder als auch die Berücksichtigung von Ausnahmefällen, wodurch die Regeln teilweise sehr komplex wurden.

## 5 Zusammenfassung und Ausblick

Der Artikel konkretisiert den Begriff der Datenqualität, indem verschiedene Betrachtungsebenen unterschieden werden und daraus die zwei Qualitätsfaktoren Designqualität und Ausführungsqualität abgeleitet werden. Des Weiteren werden Qualitätskategorien und -merkmale für diese Faktoren beschrieben, die den Qualitätsbegriff charakterisieren, detaillieren und damit auch operationalisieren. Weiterhin wird der iterative Prozess eines umfassenden Datenqualitätsmanagements aufgezeigt, das nicht die einmalige Datenbereinigung, sondern die kontinuierliche Verbesserung der Datenqualität in allen Informationssystemen des Unternehmens zum Ziel hat.

Die theoretischen Ausführungen werden anhand des Fallbeispiels der Credit Suisse konkretisiert. Hierbei wird das ganzheitliche Datenqualitätsmanagement der Credit Suisse sowohl aus technischer Sicht als auch anhand des organisatorischen Datenqualitätsprozesses eingehend beschrieben. Es zeigt sich, dass vor allem die Spezifikation der Datenqualitätsregeln als auch die Regelverwaltung und -verfeinerung den grössten Teil des Datenqualitätsprojekts der Credit Suisse ausmachen. Das Projekt hat zu einer erheblichen Akzeptanzsteigerung bei den Nutzern des Data Warehouse sowie zu einer nachhaltigen Verbesserung der Datenqualität insbesondere auch in den operativen Systemen geführt.

Für zukünftige Ausbaustufen des Datenqualitätsmoduls ist geplant, sowohl die Messresultate als auch zusätzliche Informationen nicht nur der IT- sondern auch der Fachabteilung zugänglich zu machen, um so eine noch bessere Entscheidungs-

grundlage zu bieten, wann Daten hinsichtlich ihrer Qualität nutzbar, bedingt nutzbar oder gar nicht verwendbar sind. Auch ist vorgesehen, das Datenqualitätsmodul auf allen Ebenen des Data-Warehouse-Systems einzusetzen und nicht nur auf der Ebene der Datenbereitstellung für die Bereichsdatenbanken. Weitere Einsatzgebiete sind Migrationsprojekte und Qualitätsvergleiche beim Wechsel von Datenquellen.

## Literatur

- Bleicher, K.: Das Konzept integriertes Management; Campus, Frankfurt a. M. u. a. 1992.
- Eckerson, W. W.: Data Quality and the Bottom Line; TDWI Report, The Data Warehouse Institute 2002.
- Elmasri, R., Navathe, S. B.: Fundamentals of Database Systems; 2. Aufl., Addison-Wesley, Reading u. a. 1994.
- English, L. P.: Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits; Wiley, New York u. a. 1999.
- Helfert, M.: Massnahmen und Konzepte zur Sicherung der Datenqualität; in Jung, R., Winter, R. (Hrsg.): Data Warehousing Strategie: Erfahrungen, Methoden, Visionen; Springer, Berlin u. a. 2000, S. 61-77.
- Helfert, M.: Planung und Messung der Datenqualität in Data-Warehouse-Systemen; Dissertation Universität St.Gallen, Bamberg 2002.
- Heuer, A., Saake, G.: Datenbanken. Konzepte und Sprachen; 2. Aufl., mitp-Verlag, Bonn 2000.
- Hinrichs, H.: Datenqualitätsmanagement in Data Warehouse-Umgebungen; in Heuer, A., Leymann, F., Priebe, D. (Hrsg.): Datenbanksysteme in Büro, Technik und Wissenschaft, 9. GI-Fachtagung BTW 2001, Springer, Berlin u. a. 2001, S. 187-206.
- Hinrichs, H.: Datenqualitätsmanagement in Data Warehouse-Systemen; Dissertation Universität Oldenburg, 2002.
- Jarke, M., Jeusfeld, M., Quix, C., Vassiliadis, P.: Architecture and Quality in Data Warehouses: An Extended Repository Approach; Information Systems, 24. Jg. (1999), Nr. 3, S. 229-253.
- Jarke, M., Vassiliou, Y.: Foundations of Data Warehouse Quality – A Review of the DWQ Project; in Strong, D. M., Kahn, B. K. (Hrsg.): Proceedings of the 1997 Conference of Information Quality, MIT, Cambridge, MA 1997, S. 299-313.
- Juran, J. M.: How to think about Quality; in Juran, J. M., Godfrey, A. B. (Hrsg.): Juran's Quality Handbook, 5. Aufl., McGraw Hill, New York u. a. 1999, S. 1-18.
- o. V.: Qualitätsmanagement und Statistik : Verfahren 3 : Qualitätsmanagementsysteme : Normen; DIN Deutsches Institut für Normung (Hrsg.), Beuth, Berlin 1995.

- 
- Seghezzi, H. D.: Integriertes Qualitätsmanagement – das St. Galler Konzept; Hanser, München, Wien 1996.
- Strong, D. M., Lee, Y. W., Wang, R. Y.: Data Quality in Context; Communications of the ACM, 40. Jg. (1997), Nr. 5, S. 103-110.
- Vossen, G.: Datenmodelle, Datenbanksprachen und Datenbankmanagementsysteme; 4. Aufl., Oldenburg, München u. a. 2000.
- Wallmüller, E.: Software-Qualitätssicherung in der Praxis; Hanser, München u. a. 1990.
- Wand, Y., Wang, R. Y.: Anchoring Data Quality Dimensions in Ontological Foundations; Communications of the ACM, 39. Jg. (1996), Nr. 11, S. 86-95.
- Wang, R. Y., Strong, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers; Journal of Management Information Systems, 12. Jg. (1996), Nr. 4, S. 5-33.
- Wolf, P.: Konzept eines TQM-basierten Regelkreismodells für ein „Information Quality Management“ (IQM); Verl. Praxiswissen, Dortmund 1999.