Das metadatenbasierte Datenqualitätssystem der Credit Suisse

Marcel Winter
Credit Suisse

Markus Helfert, Clemens Herrmann
Universität St. Gallen

Datenqualität ist ein entscheidender Erfolgsfaktor für eine erfolgreiche Umsetzung und Nutzung von Data-Warehouse-Systemen. Der Artikel beschreibt ein Datenqualitätssystem für ein umfassendes Datenqualitätsmanagement in Data-Warehouse-Systemen. Hierbei wird ein proaktiver Ansatz zugrundegelegt, der die Fehlervermeidung der nachträglichen Fehlerbereinigung vorzieht. Es wird eine konzeptionelle Architektur eines metadatenbasierten Datenqualitätssystems dargestellt, die anhand einer ersten konkreten Realisierungsstufe bei der Credit Suisse illustriert wird.

1 Einführung

Die Qualität von Daten und Informationen spielt in der heutigen Informationsgesellschaft eine immer wichtigere Rolle (vgl. Wolf 1999, S. 7f.). Neben einigen prominenten Beispielen aus der Presse für die teilweise verheerenden Auswirkungen mangelhafter Datenqualität[1], stellt die Qualität der Daten im Data Warehousing einen entscheidenden Erfolgsfaktor dar (vgl. English 1999, S. 4). Data-Warehouse-Projekte scheitern oft an einer unzureichenden Datenqualität (vgl. Helfert 2000, S. 65). Weitere Folgen ungenügender Datenqualität im Data Warehouse umfassen eine geringere Akzeptanz des Data Warehouse durch die Endbenutzer, schlechtere Entscheidungsprozesse auf der Basis qualitativ unzureichender Analyseergebnisse bzw. Berichte, Zusatzkosten z. B. durch Doppelerfassungen und

[1] Bekannte Beispiele sind der Absturz der ersten Ariane 5 Rakete am 4. Juni 1996 aufgrund einer falschen Datendefinition (vgl. Bange, Schinzer 2001) und der versehentliche Angriff der chinesischen Botschaft im Kosovokrieg am 8. Mai 1999 durch die NATO aufgrund falscher Adressdaten (vgl. Redman 2001, S. 39).

eine unzulängliche Unterstützung der Geschäftsprozesse (vgl. Helfert et al. 2001, S. 1f.).

Der Themenbereich Datenqualität im Data Warehousing wird bereits von einigen Autoren behandelt. 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: Vollständig, eindeutig, bedeutungsvoll und korrekt. 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 versäumt es jedoch, Überschneidungen und Beziehungen zwischen den einzelnen Merkmalen und den übergeordneten Kategorien darzustellen.

Im Rahmen einer empirische Untersuchung von Wang und Strong (vgl. Wang, Strong 1996) zur Bestimmung allgemeiner Datenqualitätsmerkmale werden vier Kategorien (Innere Datenqualität, kontextabhängige Datenqualität, Darstellungsqualität 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.

In dieser Arbeit soll ein alternativer Vorschlag zur Systematisierung und Konkretisierung des Begriffs Datenqualität gemacht werden, der für die praktische Fraugestellung bei der Credit Suisse geeignet erscheint. Darauf aufbauend wird die Konzeption eines metadatengestützen Datenqualitätssystem aufgezeigt und anschliessend anhand einer praktischen Umsetzung bei der Credit Suisse verdeutlicht. Der Artikel schliesst mit einer Zusammenfassung und einem Ausblick auf zukünftige Schritte.

2 Datenqualität

Nach dem Systematisierungsansatz von Garvin lassen sich fünf Qualitätsvorstellungen unterscheiden (vgl. Garvin 1998, S. 40f.). Der *produktbezogene Ansatz* definiert Qualität über Produkteigenschaften, d. h. Qualität ist präzise messbar und eine inhärente Eigenschaft des Produktes selbst. Qualitätsdifferenzen sind demnach auf Unterschiede in den Eigenschaftsausprägungen der Produkte zurückzuführen. 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 individuellen Zweck der Benutzung durch den Kunden dient. Nach dem *prozessorientierten Ansatz* bedeutet Qualität die Einhaltung der Spezifikationen der Produktionsprozesse. Jede Abweichung von der Spezifikation bedeutet Verringerung der Qualität. Der *wertbezogene Ansatz* stellt einen Bezug zwischen Preis und Qualität im Sinne von Nutzen her. Ein Produkt ist dann von hoher Qualität, wenn der zu entrichtende Preis und die empfangene Leistung in einem akzeptablen Verhältnis stehen. Der *transzendente Ansatz* kennzeichnet Qualität als angeborene Vortrefflichkeit, Einzigartigkeit oder Superlative. Qualität wird zu einer absoluten und universell erkennbaren Eigenschaft. Die diesem eher abstrakt philosophischen Verständnis folgende Auffassung ist jedoch für die weitere Betrachtungen nicht relevant und soll daher nicht weiter verfolgt werden.

Diese verschiedenen Ansätze sind für unterschiedliche Ebenen des Produktionssystems geeignet und verfolgen unterschiedliche Zielsetzungen. Die Ansätze können auf den Ebenen der Anforderungsanalyse, der Produkt- und der Prozessentwicklung eingeordnet werden. Daher erscheint eine zusammenhängende Betrachtung der Ansätze sinnvoll. Der Auffassung von Garvin folgend können für den Datenqualitätsbegriff drei Sichten unterschieden werden:

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

Der anwenderbezogene Qualitätsansatz bezieht sich auf eine externe Sicht und stellt den Endbenutzer mit seinen Anforderungen in den Vordergrund. Von diesen Qualitätsforderungen wird eine Produktspezifikation und ein Produktionsplan abgeleitet. Die konzeptionelle Spezifikation bildet die Grundlage für die Gestaltung der Produktionsprozesse. Auf Grundlage dieser Qualitätsebenen lässt sich Qualität, wie in Abb. 1 dargestellt, grundsätzlich in zwei Faktoren untergliedern:

- Designqualität und
- Ausführungsqualität.

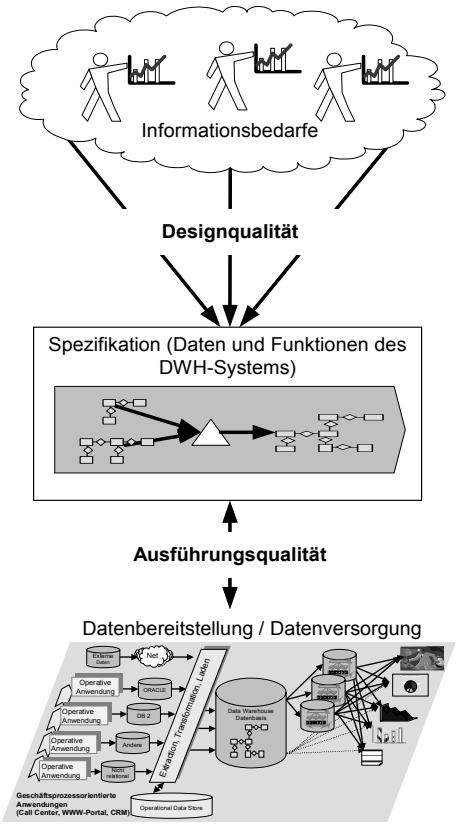

Abb. 1: Qualitätssichten

Zunächst werden die Qualitätsforderungen der Endbenutzer erfasst und durch eine Spezifikation konkretisiert. Es ist die Frage nach den geeigneten Produkteigenschaften zu beantworten. Es sind die Eigenschaften auszuwählen, welche die Bedürfnisse der Anwender am Besten erfüllen und so Kundenzufriedenheit erzeugen. In einer Datenbank werden durch Datenschemata Entitäten und Eigenschaften der zu erfassenden Datenobjekte festgelegt und können so als Spezifikation eingestuft werden.

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. Die Produktionsprozesse sind dahingehend zu kontrollieren. *Designqualität* bezieht sich auf die Erfassung von Qualitätsforderungen aus Anwendersicht in einer Spezifikation, während *Ausführungs-*

qualität die Einhaltung der von den Anwendern festgelegten Spezifikation umfasst. Eine unzureichende Gesamtqualität kann sowohl in einer mangelhaften Design- als auch in einer nicht ausreichenden Ausführungsqualität begründet sein.

3 Metadatenbasiertes Datenqualitätssystem

3.1 Konzeptionelle Architektur

Das hier dargestellte Konzept des Datenqualitätssystems beschränkt sich nicht nur auf einen Teil des Data-Warehouse-Systems sondern betrachtet alle Ebenen inklusive der operativen Systeme. Die Messung der Datenqualität wird entlang des gesamten Datenflusses vorgenommen. Der Metadatenverwaltung kommt dabei eine besondere Bedeutung zu. Es werden vorwiegend Daten über die Transformationsprozesse und über die Datenschemata zur Messung herangezogen Daneben sind zur Beurteilung der Datenqualität auch manuelle und werkzeugunterstützte Analysen von Datenqualitätsexperten und das Urteil des Datenverwenders mit einzubeziehen. Das so entwickelte Konzept, das alle qualitätsrelevanten Daten entlang des Datenflusses ermittelt, ist in Abb. 2 dargestellt.

Kern des Ansatzes ist ein in die Metadatenverwaltung integriertes Datenqualitätssystem. Hier werden alle relevanten Datenqualitätsmetadaten verwaltet. Eine Regelmenge, in der bestimmte Regeln zur Prüfung der Datenqualität hinterlegt sind, ist wesentlicher Bestandteil des Systems. Neben den zu berücksichtigenden Messobjekten in den Regelbedingungen und den Zielwerten, werden hier auch deren Ausführungszeitpunkte spezifiziert. Die sich aus den Qualitätsprüfungen ergebenden Messergebnisse werden gespeichert und sind für Qualitätsaussagen verfügbar. Diese Qualitätsaussagen werden anhand der von den Datenverwendern spezifizierten Datenqualitätsvorgaben (Soll-Datenqualität) und geeigneten Qualitätskennzahlen erstellt bzw. generiert. Datenqualitätskennzahlen und -kennzahlensysteme erlauben die Aggregation der Messergebnisse zu verdichteten Qualitätsaussagen für die Endanwender. Damit eine direkte Interpretation dieser Aussagen durch die Datenverwender möglich wird, ist auf höchster Aggregationsstufe eine Aussage anhand von drei Zuständen wünschenswert:

- Die Daten sind verwendbar (z. B. Kennzeichnung grün),
- die Daten sind eingeschränkt verwendbar (z. B. Kennzeichnung gelb) und
- die Daten sind nicht zu verwenden (z. B. Kennzeichnung rot).

Prinzipiell können die Messwerte durch Erweiterung der Datenmodelle in den bereits vorhandenen Datenhaltungssystemen oder in einem separaten Datenhaltungssystem verwaltet werden. Aufgrund der Flexibilität wird hier die getrennte Daten-

haltung für Qualitätswerte bevorzugt. Eine weitere Komponente des Datenqualitätssystems sind Benachrichtigungsregeln. Hier werden Regeln und Ereignisse zur Benachrichtigung entsprechender Personen oder Personengruppen festgelegt. Qualitätsverantwortliche können dann bei Unterschreiten bestimmter Qualitätswerte auf elektronischem Wege (z. B. E-Mail), mobilem Telefon (z. B. SMS) oder sonstigen Kommunikationskanälen über das Ereignis in Kenntnis gesetzt werden. Sie können dann problemadäquate Massnahmen einleiten und so im Sinne einer Qualitätslenkung regelnd in den Prozess eingreifen.

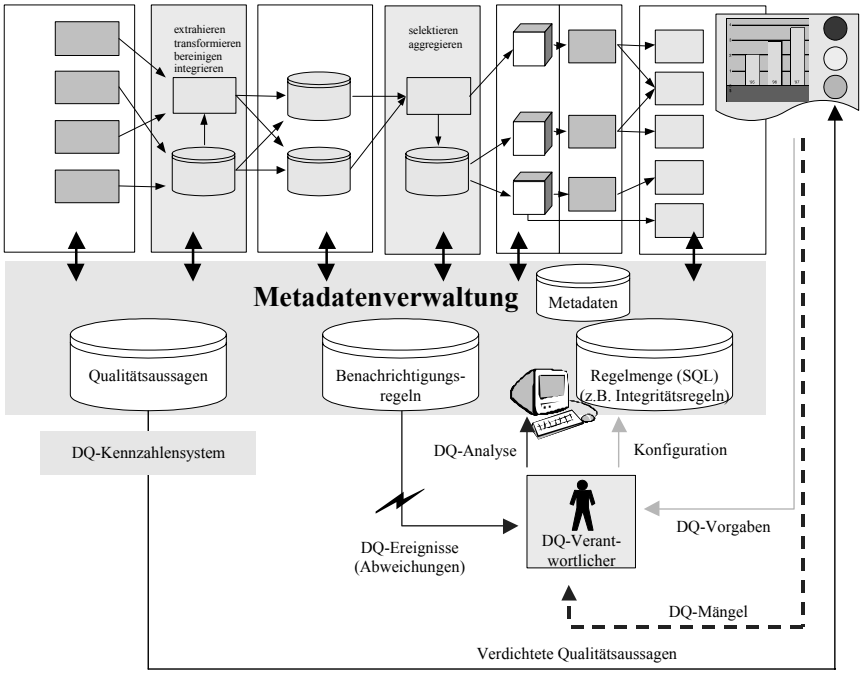

Abb. 2: Konzept eines metadatenbasierten Datenqualitätssystems

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 möglich 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 durchschnittliche Kreditvolumen verhält sich linear zur Anzahl der Kunden).

Derartige Bedingungen können aufgrund charakteristischer Eigenschaften der Daten gebildet werden. Hierzu werden sogenannte Qualitätsreferenzdaten untersucht, die einen idealtypischen Ausschnitt der Daten repräsentieren. Univariate und multivariate Methoden der deskriptiven Statistik stellen eine Möglichkeit zur Ableitung bestimmter Charakteristika der Daten dar. Des Weiteren können Verfahren des Data Mining genutzt werden, um typische Aussagen über die Daten zu generieren und daraus Regeln abzuleiten. Dieser Ansatz wird unter dem Schlagwort „Data Quality Mining“ zusammengefasst (vgl. weiterführend Soler, Yankelevich 2001; Grimmer, Hinrichs 2001).

3.2 Erste Realisierungsstufe

Das in Abschnitt 3.1 beschriebene Konzept eines Datenqualitätssystems wird im Rahmen eines Projektes bei der Credit Suisse umgesetzt. Dieses enthält alle oben beschriebenen Komponenten. In der ersten Realisierungsstufe ist lediglich die aggregierte Darstellung der vorliegenden Datenqualität für die Endbenutzer noch nicht enthalten, soll aber Gegenstand zukünftiger Ausbaustufen sein. Eine zentrale Komponente stellt die Regelbasis dar, die sowohl SQL-Statements als auch Korn-Shells enthalten kann. Ein vereinfachtes Beispiel aus der Regelbasis zur Überprüfung der Anzahl der neu hinzugekommenen Zeilen nach einem Load stellt folgender SQL-Ausdruck dar:

```
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.

Die Oberfläche zur Verwaltung der Regeln zeigt Abb. 3. Einzelne Regeln können zu Regelmengen zusammengefasst werden, die jeweils abgeschlossene Sachverhalte überprüfen. Für jede Regel kann das gewünschte Ergebnis festgelegt werden, welches bei Fehlerfreiheit generiert wird. Hierbei kann unterschieden werden zwischen einem einzigen Ergebnis und einem Intervall, in dessen Grenzen sich das Ergebnis befinden muss.

Das Datenqualitätsmodul wird zur Zeit täglich zur Überprüfung der Datenqualität der Extrakte auf der Staging Area eingesetzt. Diese Qualitätskontrolle stellt den letzten Job dar, bevor die Daten endgültig in das Data Warehouse geladen werden. Als problematisch hat sich in Zusammenarbeit mit den Datenverwendern die Identifikation der Felder herausgestellt, die für das Datenqualitätsmanagement eine hohe Priorität besitzen. Nicht alle Felder einer Tabelle müssen zwangsläufig zur Überprüfung der Datenqualität betrachtet werden. Beispielsweise sind aus technischer Sicht die Schlüsselattribute für die Datenqualität von besonderer Bedeutung. Auch aus fachlicher Sicht lassen sich derartige Präferenzen festlegen. So ist denkbar, dass z. B. die Qualität des Geburtsdatums eines Kunden wichtig ist, wohingegen das Feld Beruf nur eine untergeordnete Rolle spielt. Es hat sich jedoch gezeigt, dass derartige Prioritäten schwer zu identifizieren sind. Weiterhin werden die Regeln teilweise sehr komplex werden, da viele Ausnahmen zu berücksichtigen sind, wenn die Daten aus vielen unterschiedlichen Datenquellen stammen.

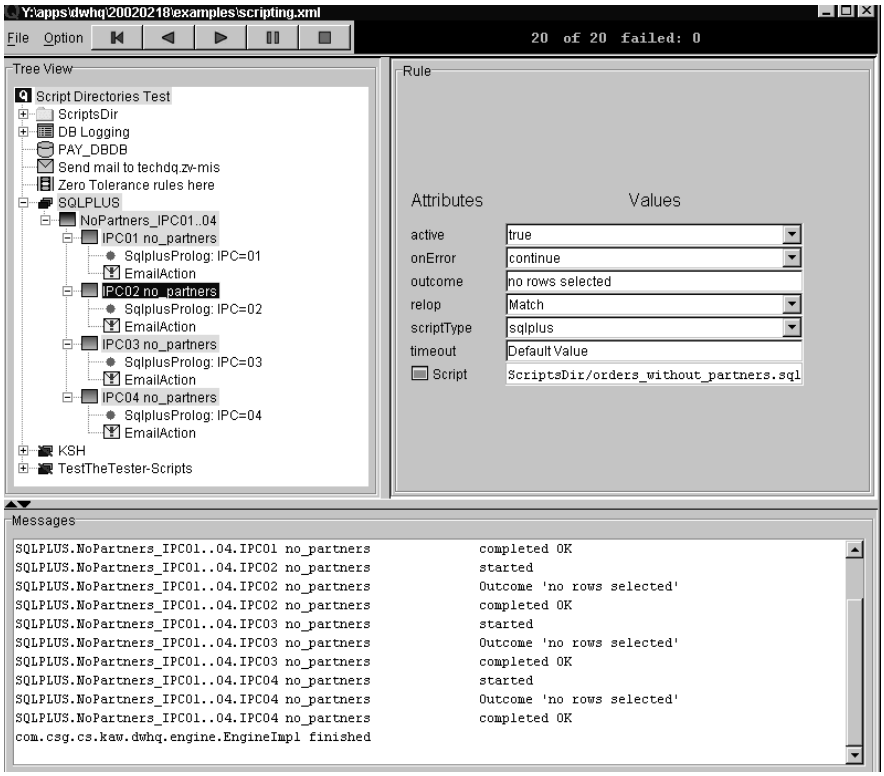

Abb. 3: Screenshot des Datenqualitätssystems (Regelverwaltung)

4 Zusammenfassung und Ausblick

Der Artikel beschreibt die konzeptionelle Architektur eines Datenqualitätssystems unter besonderer Berücksichtigung der Rahmenbedingungen der Credit Suisse. Es hat sich gezeigt, dass insbesondere das Metadatenmanagement eine entscheidende Rolle im Datenqualitätsmanagement einnimmt. Die wesentlichen Daten für ein Datenqualitätssystem stellen die Regelmenge, die Benachrichtigungsregeln und die Qualitätsaussagen dar. Diese sind allesamt im Metadatenmanagement anzusiedeln. Neben den einzelnen Komponenten beschreibt das Konzept den idealtypischen Ablauf eines proaktiven Datenqualitätsmanagements. Die zentrale Stellung hierbei nimmt der Datenqualitätsverantwortliche ein, der bei Regelverletzungen benachrichtigt wird, adäquate Verbesserungsmassnahmen koordiniert bzw. einleitet und die vorhandene Regelbasis pflegt und aktualisiert. Dies zeigt deutlich, dass ein funktionierendes Datenqualitätsmanagement sowohl auf technischer als auch

organisatorischer Ebene etabliert werden muss. Weiterhin wird die erste Realisierungsstufe des konzeptionell beschriebenen Datenqualitätssystems bei der Credit Suisse beschrieben. Hierbei wird insbesondere auf die Regelbasis und deren Verwaltung eingegangen.

In Zukunft ist ein weiterer Ausbau der Regelmenge sowie die Bereitstellung der Ergebnisse der Datenqualitätsmessung in aggregierter Form für die Datenverwender geplant. Auch soll eine intensivere organisatorische Einbettung des Datenqualitätssystems erfolgen, beispielsweise durch die Etablierung von standardisierten Prozessabläufen.

5 Literatur

Bange, C.; Schinzer, H.: Am Anfang steht die Datenqualität. Computerwoche, 2001, Nr. 44.

Elmasri, R.; Navathe, S. B.: Fundamentals of Database Systems. 2. Aufl. Reading u. a. 1994.

English, L. P.: Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. New York u. a. 1999.

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

Grimmer, U.; Hinrichs, H.: A Methodological Approach to Data Quality Management Supported by Data Mining. In: Pierce, E. M.; Kaatz-Haas, R. (Hrsg.): Proceedings of the Sixth International Conference on Information Quality. Cambridge 2001, S. 217-232.

Helfert, M.: Massnahmen und Konzepte zur Sicherung der Datenqualität. In: Jung, R., Winter, R. (Hrsg.): Data Warehousing Strategie: Erfahrungen, Methoden, Visionen. Berlin u. a. 2000, S. 61-77.

Helfert, M.; Herrmann, C.; Strauch, B.: Datenqualitätsmanagement. Arbeitsbericht des Instituts für Wirtschaftsinformatik der Universität St. Gallen, BE HSG/CC DW2/02, 2001.

Heuer, A., Saake, G.: Datenbanken. Konzepte und Sprachen. 2. Aufl. Bonn 2000.

Jarke, M.; Jeusfeld, M.; Quix, C.; Vassiliadis, P.: Architecture and Quality in Data Warehouses: An Extended Repository Approach. In: Information Systems 24 (1999) 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. Cambridge 1997, S. 299-313.

Redman, T. C.: Data Quality: the field guide. Boston u. a. 2001.

Soler, S. V., Yankelevich, D.: Quality Mining: A Data Mining Method for Data Quality Evaluation. In: Pierce, E. M.; Kaatz-Haas, R. (Hrsg.): Proceedings of the Sixth International Conference on Information Quality. Cambridge 2001, S. 162-172.

Vossen, G.: Datenmodelle, Datenbanksprachen und Datenbankmanagementsysteme. 4. Aufl. München u. a. 2000.

Wand, Y., Wang, R. Y.: Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM. 39 (1996) 11, S. 86-95.

Wang, R. Y., Strong, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers. In: Journal of Management Information Systems. 12 (1996) 4, S. 5-33.

Wolf, P.: Konzept eines TQM-basierten Regelkreismodells für ein „Information Quality Management“ (IQM). Dortmund 1999.