1 Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement

Bernd Heinrich, Mathias Klier

1.1 Einleitung

Für eine ökonomische Betrachtung der Datenqualität (DQ) und insbesondere die Planung von DQ-Maßnahmen unter Kosten-Nutzen-Aspekten sind DQ-Metriken unverzichtbar (vgl. z.B. [Heinrich & Klier 2006; Naumann 2007; Pipino et al. 2002]). Deswegen wird im Folgenden die Fragestellung aufgegriffen, wie DQ zweckorientiert und adäquat quantifiziert werden kann. Dazu werden Metriken entwickelt und vorgestellt, die zum einen eine quantitative Analyse der zum Messzeitpunkt vorhandenen DQ ermöglichen sollen, um Handlungsbedarfe zu identifizieren. Zum anderen sollen Auswirkungen auf die DQ, wie z. B. zeitlicher Verfall oder die Durchführung von DQ-Maßnahmen, zielgerichtet – durch Vergleich des DQ-Niveaus zu zwei oder mehreren Messzeitpunkten – untersucht werden können.

Die Identifikation und Klassifikation von DQ-Dimensionen wird in einer Vielzahl von wissenschaftlichen und praxisorientierten Veröffentlichungen thematisiert (vgl. z. B. [Wang & Strong 1996, English 1999; Eppler 2003; Helfert 2002; Hinrichs 2002; Lee et al. 2002; Jarke & Vassiliou 1997; Redman 1996]). Nachfolgend werden die DQ-Dimensionen Vollständigkeit, Fehlerfreiheit, Konsistenz und Aktualität näher untersucht und mit entsprechenden Metriken versehen. Diese Dimensionen werden zum einen in wissenschaftlichen Veröffentlichungen besonders intensiv diskutiert. Zum anderen spielen die genannten Dimensionen aber auch in der Praxis eine wichtige Rolle: So fanden sich Vollständigkeit, Fehlerfreiheit, Konsistenz sowie Aktualität beispielsweise in einer Studie von Helfert, die unter 25 größeren Unternehmen in Deutschland, Österreich und der Schweiz durchgeführt wurde, allesamt unter den fünf meist genannten DQ-Dimensionen wieder (vgl. [Helfert 2002]).

Neben der Selektion von betrachteten DQ-Dimensionen wird zudem im Folgenden die „fachliche" DQ fokussiert, die hinsichtlich der Spezifikation des Datenmodells weitgehend automatisiert und objektivierbar gemessen werden soll (vgl. spezifikationsorientierte DQ bzw. Konformitätsqualität nach [Juran 1999] und [Seghezzi 1996]). Inwiefern den Anforderungen der Datenverwender bei der Spezifikation des Informationssystems Rechnung getragen wurde, ist dagegen den Bereichen Anforderungsmanagement und Bedarfsanalyse zuzurechnen und kann vor allem mittels Fragebögen und Interviews untersucht werden. Dieser Aspekt wird hier ebenso wie beispielsweise die Qualität der Datenrepräsentation, die eher auf die

„technische“ DQ im Sinne von Datenformat und Datenspeicherung abzielt, nicht weiter betrachtet.

Die Zusammenhänge zwischen DQ-Metriken und der Planung von DQ-Maßnahmen im Rahmen eines ökonomisch orientierten DQ-Managements lassen sich anhand des DQ-Regelkreises graphisch veranschaulichen (siehe Bild 1):

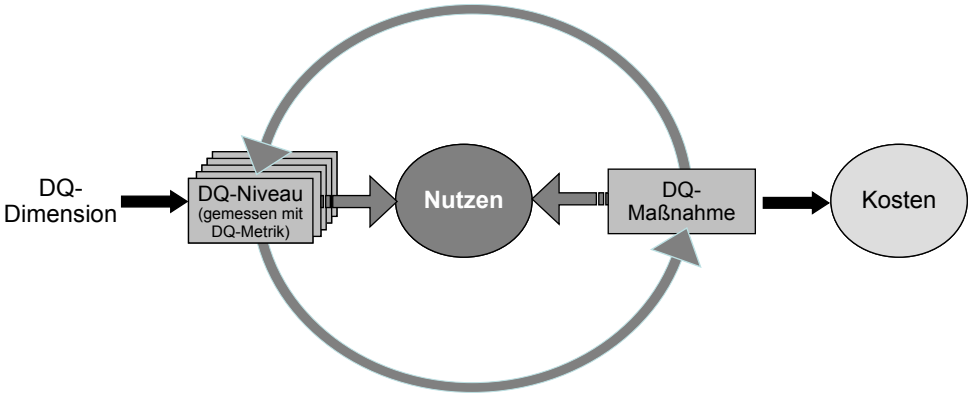

Abbildung 1: Datenqualitätsregelkreis

Den Regler, über den in den Regelkreis eingegriffen werden kann, stellen die DQ-Maßnahmen dar. Die Durchführung von DQ-Maßnahmen soll dabei eine Verbesserung der DQ – gemessen anhand der Metriken – zur Folge haben, wodurch ein entsprechender ökonomischer Nutzen resultiert. Ausgehend von einem bestimmten DQ-Niveau kann umgekehrt ebenfalls mithilfe der Metriken ex ante bzw. ex post die Steigerung der DQ durch entsprechende Maßnahmen abgeschätzt bzw. gemessen werden. Werden ökonomische Maßstäbe zugrunde gelegt, muss jedoch die Auswahl der DQ-Maßnahmen unter Berücksichtigung von Kosten-Nutzen-Gesichtspunkten erfolgen (vgl. z. B. [Campanella 1999; Feigenbaum 1991; Machowski & Dale 1998; Shank & Govindarajan 1994]). Beispielsweise würde man bei zwei zu bewertenden, sich ausschließenden Maßnahmen, aus denen (annähernd) der gleiche ökonomische Nutzen resultieren würde, nur die kostengünstigere in Betracht ziehen.

Im nächsten Abschnitt werden zunächst die allgemeinen Anforderungen an DQ-Metriken formuliert, bevor im darauf folgenden Abschnitt ein kurzer Überblick über ausgewählte Ansätze zur Quantifizierung von DQ gegeben wird. Im Anschluss daran werden Metriken für die vier oben genannten DQ-Dimensionen entwickelt sowie deren Eigenschaften diskutiert. Bevor abschließend die Ergebnisse zusammengefasst und weiterer Forschungsbedarf identifiziert werden, soll eine kurze Anwendung der Metrik für Aktualität im Customer Relationship Manage-

ment eines Mobilfunkanbieters einen Einblick geben, wie die Metriken im Rahmen eines ökonomisch orientierten DQ-Managements genutzt werden können.

1.2 Anforderungen an Datenqualitätsmetriken

Um eine wissenschaftliche Fundierung zu gewährleisten und gleichzeitig eine praktische Anwendung zu ermöglichen, werden nachfolgend Anforderungen an DQ-Metriken definiert (in Teilen ähnliche Anforderungen finden sich auch bei [Even & Shankaranarayanan 2005] und [Hinrichs 2002]):

- [*Normierung*] Um die Interpretierbarkeit und Vergleichbarkeit der Metrikergebnisse zu gewährleisten, ist eine geeignete Normierung der Metrikergebnisse zu fordern.
- [*Kardinalität*] Um eine Analyse der zeitlichen Entwicklung der Metrikergebnisse (DQ-Niveau) und eine ökonomische Bewertung von Maßnahmen zu unterstützen, ist die Kardinalität der Metriken erforderlich (vgl. [Bamberg et al. 2007]).
- [*Sensibilisierbarkeit*] Um das DQ-Niveau zielgerichtet messen zu können, ist es notwendig, dass die Metriken für eine konkrete Anwendung sensibilisiert und für den jeweiligen Zweck, welcher der Messung zugrunde liegt, konfiguriert werden können.
- [*Aggregierbarkeit*] Um bei Zugrundelegung eines relationalen Datenbankschemas einen flexiblen Einsatz zu ermöglichen, soll die Metrik Ergebnisse auf Attributwert-, Tupel-, Relationen- sowie Datenbankebene liefern können. Dabei muss jedoch die Aggregierbarkeit der Metrikergebnisse auf einer Ebene zur nächst höheren Ebene gewährleistet sein.
- [*Fachliche Interpretierbarkeit*] In der praktischen Anwendung reicht in der Regel die bloße Normierung und Kardinalität der Metriken nicht aus. Vielmehr müssen die resultierenden Metrikergebnisse auch *fachlich* interpretierbar und reproduzierbar sein.

Auf Basis dieser Anforderungen lassen sich existierende Beiträge (beispielsweise [Ballou et al. 1998; English 1999; Helfert 2002; Hinrichs 2002; Lee et al. 2002; Naumann et al. 2004; Redman 1996; Wang et al. 1995]) analysieren. Im folgenden Abschnitt wird kurz auf ausgewählte Beiträge eingegangen.

1.3 Bisherige Beiträge zur Messung von Datenqualität

In der Literatur findet sich eine ganze Reihe von Ansätzen zur Quantifizierung von DQ, die sich neben den jeweils fokussierten DQ-Dimensionen vor allem in den zugrunde liegenden Messverfahren unterscheiden (vgl. z. B. Ansätze in [Wang et al. 1995]). So existieren nach [Helfert 2002] sowohl Verfahren, die auf der subjektiven Qualitätseinschätzung der Datenverwender beruhen, als auch Ansätze, die auf

einer direkten Analyse des Datenbestands oder einer prozessorientierten Betrachtungsweise basieren. In der Folge werden ausgewählte Ansätze vorgestellt.

Am Massachusetts Institute of Technology (MIT), das den Begriff des "Total Data Quality Managements" geprägt hat, wurde zur Messung der DQ die AIM Quality (AIMQ)-Methode entwickelt [vgl. z. B. Lee et al. 2002]. Diese besteht aus drei Komponenten. Die erste ist das Product-Service-Performance-Model, das eine vorgefertigte Menge von DQ-Dimensionen in vier Quadranten aufteilt. Unterschieden wird dabei zum einen auf Grundlage der Messbarkeit des Kriteriums. Dabei wird differenziert, ob die Konformität hinsichtlich einer formalen Spezifikation (z. B. Vollständigkeit) oder einer subjektiven Erwartung des Datennutzers (z. B. Interpretierbarkeit) bestimmt werden kann. Zum anderen wird nach der Qualität des Datenprodukts (z. B. Fehlerfreiheit) und des Services (z. B. Rechtzeitigkeit) unterschieden. Die Messung der DQ erfolgt dann, basierend auf obigem Modell, mittels eines zweiten Bestandteils in Form einer Befragung der Endanwender nach deren Qualitätseinschätzungen. Als dritte Komponente von AIMQ werden mit Benchmark-Gap- und Role-Gap-Analyse eine anwendungsunabhängige sowie eine anwendungsabhängige Qualitätsanalyse der Messergebnisse vorgeschlagen. Problematisch bei der AIMQ-Methode ist, dass die Messung der DQ in der Regel auf einer subjektiven Qualitätseinschätzung basiert und anhand von Befragungen vorgenommen wird. Dieses Vorgehen ermöglicht in der Regel keine automatisierte, objektivierbare und beliebig reproduzierbare Analyse der DQ bzw. der erhaltenen Ergebnisse der Messung. Zudem ist eine zielgerichtete und fokussierte Messung der DQ auf den konkreten Anwendungszweck und -kontext hin nicht explizit vorgesehen, auch wenn diese durch die Befragung der Datennutzer in der Role-Gap-Analyse natürlich implizit berücksichtigt wird. Da die Auswertung der Befragungsergebnisse über alle Datennutzer erfolgt, werden jedoch deren subjektive Qualitätseinschätzungen und somit die Anwendungskontexte vermischt. Dies kann zu Bewertungsproblemen führen, da die Nutzer in der Regel unterschiedliche Zielsetzungen verfolgen.

Ein anderes hier zu nennendes Verfahren ist der Ansatz von Hinrichs. Dieser entwickelt Metriken für ausgewählte DQ-Dimensionen, um die Qualität eines Datenbestandes bewerten zu können (vgl. [Hinrichs 2002]). Das zugrunde liegende Verfahren zur Quantifizierung der DQ ist dabei sehr aussichtsreich, da eine objektivierbare, zielgerichtete Bewertung angestrebt und eine weitgehend automatisierte Messung ermöglicht wird. Allerdings können beim Praxiseinsatz durchaus auch Probleme auftreten, da die vorgestellten Metriken nur schwer interpretierbar sind, was eine Begründung und Erklärung der Messergebnisse in der Diskussion beispielsweise mit der Fachseite erschweren dürfte. So basieren einige Metriken, wie z. B. diejenige für die DQ-Dimension Fehlerfreiheit, auf der Bildung von Quotienten der Form

$$\frac{1}{\text{Abstandsbewertung} + 1},$$

wobei die Abstandsbewertung aus dem Intervall [0; ∞] angibt, inwieweit ein Attributwert im Informationssystem von der entsprechenden Ausprägung der Realwelt-Entität abweicht. Dadurch wird zwar der Wertebereich der Metrik auf das Intervall [0; 1] beschränkt, jedoch geht durch die Quotientenbildung die Interpretierbarkeit der resultierenden Werte verloren (vgl. im Detail [Heinrich et al. 2007]). Zudem hängt die Größenordnung der Werte stark vom verwendeten Abstandsmaß und dessen Eigenschaften ab, was zusätzlich eine Vergleichbarkeit der Ergebnisse erschwert.

Der Ansatz von Helfert unterscheidet grundsätzlich – basierend auf den Ausführungen von [Seghezzi 1996] und [Juran 1999] – die beiden Qualitätsfaktoren Designqualität und Ausführungsqualität (vgl. [Helfert 2002]). Dabei bezeichnet die Designqualität den Grad der Übereinstimmung zwischen den Anforderungen der Datennutzer und der entsprechenden Umsetzung in der Spezifikation des Informationssystems. Die Ausführungsqualität, die Helfert schwerpunktmäßig betrachtet, drückt dagegen aus, in welchem Maße diese Spezifikation durch das Informationssystem (tatsächlich) eingehalten wird. Diese Unterscheidung ist im Hinblick auf die Messung der DQ sinnvoll, da somit die (subjektive) Einschätzung der Konformität zwischen dem spezifizierten Datenangebot und dem Datenbedarf des Nutzers von der (objektivierbaren) Analyse der Übereinstimmung von vorhandenem und spezifiziertem Datenangebot getrennt wird. Den zentralen Aspekt bei Helfert stellt die Integration des DQ-Managements in die Metadatenverwaltung dar, die ein weitgehend automatisiertes und werkzeugunterstütztes Management der DQ ermöglichen soll. Die Qualitätsanforderungen sind dabei durch eine Regelmenge repräsentiert. Die Analyse der DQ basiert dann hauptsächlich auf der (automatisierten) Überprüfung derartiger Regeln, d. h. anhand der Analyse werden Qualitätsaussagen im Sinne von

$$\text{Widerspruchsfreiheit} = 1 - \frac{\text{Anzahl verletzter Bedingungen}}{\text{Anzahl spezifizierter Bedingungen}}$$

abgeleitet. Derartige Qualitätsaussagen sollen in aggregierter Form als Größen zur Quantifizierung der DQ Verwendung finden und somit Auskunft über die Qualität des Datenbestands geben. Insgesamt stellt Helfert in seinen Ausführungen jedoch keine konkreten Metriken vor, sondern hat vielmehr den Anspruch, ein ganzheitliches, proaktives DQ-Management auf einer konzeptionellen Ebene zu beschreiben.

Neben den wissenschaftlichen Ansätzen sollen auch die beiden bekannten Konzepte von English und Redman aus der Praxis genannt werden. English verfolgt dabei die Total Quality data Management-Methode (vgl. [English 1999]), die an die Konzepte des Total Quality Managements angelehnt ist. Dabei führt er Vorge-

hensmuster zur Messung der Datendefinitions- und Architekturqualität (das Informationssystem betreffend) sowie der Qualität der Datenwerte und der Datenrepräsentation an. Obwohl das Verfahren in einer Reihe von Praxisprojekten Verwendung gefunden hat, gibt es hier kein allgemeines, dokumentiertes Vorgehen zur Quantifizierung der DQ. Vielmehr wird der gesamte DQ-Regelkreis auf einer konzeptionellen Ebene betrachtet. Redman verfolgt im Gegensatz zu English einen stark prozessorientierten Ansatz und kombiniert Messverfahren für gezielt ausgewählte Abschnitte im Informationsfluss mit dem Konzept der statistischen Qualitätskontrolle (vgl. [Redman 1996]). Konkrete Metriken zur Quantifizierung der DQ werden dabei allerdings nicht entwickelt.

Da die bestehenden Ansätze die zuvor definierten Anforderungen jedoch nicht oder nicht vollständig erfüllen – auch infolge unterschiedlicher Zielsetzungen, da sie zum Teil beispielsweise die subjektive Qualitätseinschätzung der Datenverwender fokussieren –, wird im Folgenden ein eigener Ansatz vorgestellt, der einen Beitrag zur wissenschaftlichen Fundierung sowie zur praktischen Einsetzbarkeit von DQ-Metriken leisten will.

1.4 Metriken und Messverfahren für DQ

Die im Folgenden vorgestellten Metriken für die Dimensionen Vollständigkeit, Fehlerfreiheit, Konsistenz und Aktualität werden – orientiert an der Anforderung der Aggregierbarkeit – jeweils für die Attributwert-, Tupel-, Relationen- sowie Datenbankebene definiert. Dabei wird jede Metrik „bottom up" entwickelt – d. h. eine Metrik auf Ebene n+1 (z. B. Vollständigkeit auf Tupelebene) basiert auf der entsprechenden Metrik auf Ebene n (Vollständigkeit auf Attributwertebene).

1.4.1 Metrik für die DQ-Dimension Vollständigkeit

Unter Vollständigkeit wird hier die Eigenschaft verstanden, dass die Attribute im Informationssystem mit Werten belegt sind, die (semantisch) vom Wert *NULL* abweichen. *NULL* ist dabei kein erforderlicher oder definierter Attributwert, sondern lediglich ein Platzhalter für die Nichtbefüllung. Die Metriken auf Attributwert-, Tupel- sowie Relationenebene sollen dabei in Anlehnung an Hinrichs (vgl. [Hinrichs 2002]) definiert werden. Zusätzlich wird darauf eingegangen, welche Probleme bei der praktischen Anwendung der Metrik auftreten können und wie diesen zu begegnen ist. Auf Datenbankebene muss die Metrik zudem anders ausgestaltet werden, um eine objektivierbare Messung zu ermöglichen.

Auf Attributwertebene wird die Metrik für Vollständigkeit $Q_{Vollst.}(w)$ folgendermaßen definiert, wobei w einen Attributwert im Informationssystem symbolisiert:

$$Q_{Vollst.}(w) := \begin{cases} 0 & falls\ w = NULL\ oder\ w\ zu\ NULL\ (semantisch)\ äquivalent \\ 1 & sonst \end{cases}$$

Die Qualität eines Attributwertes wird also hinsichtlich Vollständigkeit mit dem Minimalwert von null bewertet, falls das entsprechende Attribut nicht befüllt ist oder einen zu *NULL* (semantisch) äquivalenten (Default-)Wert enthält (z. B. Dummy-Wert). Ansonsten ergibt sich der Wert der Metrik auf Attributwertebene zu eins.

Probleme bei der Metrik können dann auftreten, wenn ein Attributwert nicht aus Mangel an verfügbaren Daten mit *NULL* belegt ist, sondern, weil der entsprechende Wert in der Realwelt gar nicht existiert (z. B. Name des Ehepartners bei ledigen Personen). In diesem Fall wäre das entsprechende Attribut mit dem Wert *NULL* in der Tat richtig belegt und die Bewertung hinsichtlich Vollständigkeit müsste den Wert eins und nicht den Wert null liefern. Schwierigkeiten dieser Art können umgangen werden, indem Indikatoren dafür eingeführt werden, dass der entsprechende Wert in der Realwelt nicht existiert. So kann beispielsweise das Attribut *Name des Ehepartners* (automatisiert) mit „*nicht verheiratet*“ belegt werden, falls bei der Erfassung des Familienstandes *ledig* angegeben wird. Somit ist das entsprechende Attribut in der Datenbank befüllt und die obige Metrik liefert auf Attributwertebene den korrekten Wert eins. Vor der ersten Anwendung der Metrik müssen die Daten somit hinsichtlich der vorgestellten Problematik untersucht und ggf. auftretende Schwachstellen beseitigt werden. Hierbei wäre beispielsweise an eine Vervollständigung des Datenbestands mit Indikatoren zu denken, sofern dies technisch sowie fachlich möglich und sinnvoll ist und keine Seiteneffekte auf Ergebnisse anderer Metriken besitzt.

Im Folgenden wird, basierend auf den obigen Ausführungen zur Attributwertebene, die Metrik auf Tupelebene formuliert. Sei hierbei T ein Tupel mit den Attributwerten $T.A_1$, $T.A_2$, …, $T.A_{|A|}$ für die Attribute A_1, A_2, …, $A_{|A|}$ und $g_i \in [0; 1]$ die relative Wichtigkeit von A_i in Bezug auf Vollständigkeit. Dann ergibt sich unter Verwendung der Metrik auf Attributwertebene die Metrik auf Tupelebene als gewichtetes arithmetisches Mittel:

$$Q_{Vollst.}(T) := \frac{\sum_{i=1}^{|A|} Q_{Vollst.}(T.A_i) g_i}{\sum_{i=1}^{|A|} g_i}$$

Die Vollständigkeit eines Tupels wird folglich basierend auf der Vollständigkeit der enthaltenen Attributwerte berechnet. Dabei ist es möglich, diese je nach Zielsetzung mit Gewichtungen g_i zu versehen. Dies ist insofern sinnvoll, da je nach zugrunde liegendem Zweck in einem Anwendungskontext die Attribute von unterschiedlicher Bedeutung sind. So sind z. B. für die Durchführung von Mailingkampagnen Attribute wie *Name*, *Vorname*, *Adresse* oder *E-Mail* besonders relevant,

wohingegen bei telefonischen Kampagnen vor allem die *Telefonnummer* und nicht *Adresse* und *E-Mail* von Bedeutung sind.

Im nächsten Schritt wird die Metrik auf Relationenebene definiert. Sei hierbei R eine nicht leere Relation oder ein mehrelementiger View. Dann ergibt sich die Vollständigkeit der Relation R auf Basis des arithmetischen Mittels der Vollständigkeitsbewertungen für die einzelnen Tupel T_j aus R ($j = 1, 2, \ldots, |T|$) wie folgt:

$$Q_{Vollst.}(R) := \frac{\sum_{j=1}^{|T|} Q_{Vollst.}(T_j)}{|T|}$$

Durch die Verwendung des arithmetischen Mittels werden dabei alle Qualitätsbewertungen auf Tupelebene gleich gewichtet und aufsummiert. Jedem enthaltenen Tupel kommt somit die gleiche Bedeutung zu. Dies ist deswegen sinnvoll, da in der Regel in einem Anwendungskontext die einzelnen, bereits selektierten Tupel (bspw. verschiedene Kundentupel in einer Marketingkampagne) nicht in unterschiedlicher Art und Weise behandelt oder genutzt werden. Sollte dies im Einzelfall notwendig sein, so sind ebenfalls Gewichtungsfaktoren nach obigem Muster denkbar.

Für die Definition der Metrik für Vollständigkeit auf Datenbankebene sei D eine Datenbank, die sich als disjunkte Zerlegung der Relationen R_k ($k = 1, 2, \ldots, |R|$) darstellen lässt – d. h., die gesamte Datenbank lässt sich in paarweise überschneidungsfreie Relationen R_k zerlegen, so dass jedes Attribut des Informationssystems in genau einer der Relationen enthalten ist (eine mathematische Formulierung dieses Sachverhaltes ist $D = R_1 \cup R_2 \cup \ldots \cup R_{|R|} \wedge R_i \cap R_j = \varnothing \; \forall i \neq j$). Weiter sei g_k die relative Wichtigkeit der Relation R_k in Bezug auf die Dimension Vollständigkeit. Dann wird die Vollständigkeit der Datenbank wiederum auf Basis der Vollständigkeit der Relationen R_k ($k = 1, 2, \ldots, |R|$) definiert:

$$Q_{Vollst.}(D) := \frac{\sum_{k=1}^{|R|} Q_{Vollst.}(R_k) g_k}{\sum_{k=1}^{|R|} g_k}$$

Über die Gewichtungsfaktoren $g_k \in [0; 1]$ ist es dabei im Vergleich zu Hinrichs, bei dem sich die Vollständigkeit der Datenbank als ungewichtetes arithmetisches Mittel ergibt, möglich, die relative Wichtigkeit der einzelnen Relationen gemäß der jeweiligen Zielsetzung zu berücksichtigen. Das Vorgehen von Hinrichs hat zur Folge, dass hinsichtlich der verfolgten Zielsetzung kaum relevante Relationen genauso stark in die Berechnung eingehen wie besonders wichtige Relationen. Zudem ist für den Fall, dass das ungewichtete arithmetische Mittel Verwendung fin-

det, die Quantifizierung der Vollständigkeit auf Datenbankebene von der betrachteten Zerlegung der Datenbank abhängig. So kommt beispielsweise der Relation R_k mit $k \neq 2$ bei der disjunkten Zerlegung $\{R_1, R_2, R_3, \ldots, R_{|R|}\}$ ein relatives Gewicht von $1/|R|$ zu, wohingegen dieselbe Relation bei Verwendung der disjunkten Zerlegung $\{R_1, R_2', R_2'', R_3 \ldots, R_{|R|}\}$ mit $R_2' \cup R_2'' = R_2$ und $R_2' \cap R_2'' = \varnothing$ nur mit dem Faktor $1/(|R|+1)$ eingeht.

Die Messung der Vollständigkeit mit Hilfe der Metrik kann in der Regel einfach mittels entsprechender SQL-Abfragen und bei Bedarf für den kompletten Datenbestand durchführt werden. Im nächsten Abschnitt wird eine Metrik für die DQ-Dimension Fehlerfreiheit vorgestellt.

1.4.2 Metrik für die DQ-Dimension Fehlerfreiheit

Unter Fehlerfreiheit wird hier die Eigenschaft verstanden, dass die Attributwerte im Informationssystem den zugehörigen Ausprägungen der modellierten Realwelt-Entität entsprechen – d. h., dass die im Informationssystem abgelegten Werte mit den tatsächlichen, realen Werten übereinstimmen. Nach Würthele existieren bei der Messung der Fehlerfreiheit grundsätzlich zwei Möglichkeiten:

Beim „Alles oder Nichts"-Ansatz wird bei der Überprüfung ausschließlich zwischen fehlerfrei (Attributwert stimmt vollständig mit der Ausprägung der modellierten Realwelt-Entität überein) und nicht fehlerfrei (es existiert mindestens eine Abweichung) differenziert (vgl. [Würthele 2003]). Im Gegensatz dazu wird beim Toleranz-Ansatz der Umfang der Übereinstimmung (beziehungsweise der Abweichung) zwischen Attributwert und Ausprägungen der modellierten Realwelt-Entität gemessen und ist damit als Analyseergebnis zulässig. So kann berücksichtigt werden, ob die entsprechenden Attributwerte nur geringfügig oder in größerem Umfang von den realen Ausprägungen abweichen.

Bei der im Folgenden entwickelten Metrik kann je nach verwendetem Abstandsmaß der „Alles oder Nichts"- oder der Toleranz-Ansatz Berücksichtigung finden. Die Vorteile der vorgestellten Metrik im Vergleich zu bisherigen Ansätzen liegen dabei vor allem in der Kardinalität und Interpretierbarkeit begründet. So können die resultierenden Werte grundsätzlich als prozentualer Wert für die Fehlerfreiheit des untersuchten Datenbestands verstanden werden. Im Weiteren wird die Metrik für Fehlerfreiheit dabei aus Platzgründen nur auf Attributwertebene vorgestellt. Sie kann jedoch analog zur Metrik für Vollständigkeit ebenfalls für die anderen Ebenen formuliert werden.

Sei w_I ein Attributwert im Informationssystem und w_R der entsprechende Attributwert in der Realwelt. Sei zudem $d(w_I, w_R)$ ein domänenspezifisches, auf das Intervall [0; 1] normiertes Abstandsmaß zur Bestimmung der Abweichung zwischen w_I und w_R. Mögliche Abstandsmaße sind beispielsweise folgende:

- Der domänenunabhängigen Abstandsfunktion

$$d_1(w_I, w_R) := \begin{cases} 0 & \text{falls } w_I = w_R \\ 1 & \text{sonst} \end{cases}$$

liegt der „Alles oder Nichts"-Ansatz zugrunde. Somit lassen sich zwei Fälle unterscheiden: Entweder der Attributwert im Informationssystem stimmt mit der Ausprägung der entsprechenden Realwelt-Entität überein (Abstand entspricht null) oder die Abweichung wird mit dem Maximalwert von eins festgelegt.

– Ein Abstandsmaß, das speziell bei numerischen Attributwerten eingesetzt werden kann, ist die Abstandsfunktion

$$d_2(w_I, w_R) := \left(\frac{|w_I - w_R|}{\max\{|w_I|, |w_R|\}} \right)^{\alpha},$$

die den Wert null ebenfalls nur bei vollständiger Übereinstimmung annimmt. Allerdings kann über den Parameter $\alpha \in \mathrm{IR}^+$ – je nach untersuchtem Attribut und verfolgter Zielsetzung der Messung – beeinflusst werden, wie stark die Metrik auf relative Abweichungen von w_I und w_R reagieren soll. So kann es beispielsweise im Fall einer Marketingkampagne bei der Untersuchung des Attributs *PLZ* notwendig sein, dass kleine Abweichungen relativ stark ins Gewicht fallen, da hierdurch eventuell das Kundenanschreiben nicht mehr zugestellt werden kann – hier ist $\alpha < 1$ zu wählen. Soll die Abstandsfunktion dagegen „toleranter" gegenüber kleinen Abweichungen sein, ist $\alpha > 1$ angebracht – wie z. B. beim Attribut *Hausnummer*, da die Zustellung hier trotzdem noch möglich ist. Bei Verwendung dieser Abstandsfunktion muss allerdings beachtet werden, dass die Normierung des Maßes auf das Intervall [0; 1] nur dann gegeben ist, wenn die Werte w_I und w_R gleiche Vorzeichen haben.

Andere Abstandsmaße $d(w_I, w_R)$, die es ermöglichen, die Ähnlichkeit von Zeichenketten zu bestimmen, können auf Basis von Editierabstand, Hamming-Distanz und N-Grammen gebildet werden, wobei hier zum Teil eine Normierung auf das Intervall [0; 1] notwendig ist:

– Der Editierabstand $d_{Edit.}(w_I, w_R)$ ist als kleinste Menge elementarer Operationen definiert, mit denen eine Zeichenkette in eine andere transformiert werden kann, wobei Einfügen und Löschen von einzelnen Zeichen ebenfalls als elementare Operationen zu betrachten sind. Wird zusätzlich das Ersetzen von Zeichen erlaubt, spricht man von der Levenshtein-Metrik $d_{Lev.}(w_I, w_R)$, die durch Hinzunehmen der Transposition (Vertauschung benachbarter Symbole) als weitere zulässige Operation zur so genannten Damerau-Levenshtein-Metrik $d_{Da.-Lev.}(w_I, w_R)$ ausgebaut werden kann, die speziell zur Tippfehlerkorrektur entworfen wurde. Bei Verwendung dieser Abstandsmaße muss der resultierende Wert noch auf das Intervall [0; 1] normiert

werden. Diese Normierung kann dadurch erfolgen, dass die Werte durch das Maximum der Längen der beiden Zeichenketten w_I und w_R dividiert werden.

- Die Hamming-Distanz $d_{Ham.}(w_I, w_R)$ summiert die Anzahl der Positionen, in denen sich die beiden Zeichenketten w_I und w_R unterscheiden. Definitionsgemäß existiert die Hamming-Distanz dabei nur für Zeichenketten gleicher Länge – bei Strings unterschiedlicher Länge kann jedoch der jeweils kürzere mit „Dummy-Zeichen" aufgefüllt werden, die als nicht übereinstimmend gelten. Für zwei gleich lange Strings w_I und w_R mit $|w_I| = |w_R| = m$ ergibt sich die auf das Intervall [0; 1] normierte Hamming-Distanz zu:

$$d_{Ham.}(w_I, w_R) := \frac{\left|\left\{i \in \{1,2,...m\} \mid w_I[i] \neq w_R[i]\right\}\right|}{m}$$

- N-Gramme betrachten das gemeinsame Auftreten von Substrings in den zu vergleichenden Zeichenketten. Ein N-Gramm ist dabei ein zusammenhängender Teil einer Zeichenkette und hat die Länge N. Für die Zeichenketten w_I und w_R werden dabei jeweils alle enthaltenen N-Gramme gebildet und in entsprechenden Mengen abgelegt. Danach wird die Anzahl der in beiden Mengen gleichermaßen enthaltenen N-Gramme ins Verhältnis zur Anzahl der insgesamt in $NG(w_I)$ und $NG(w_R)$ enthaltenen N-Gramme gesetzt. So ergibt sich die Abstandsfunktion, deren Wertebereich auf das Intervall [0; 1] beschränkt ist, zu:

$$d_{N-Gramm}(w_I, w_R) := 1 - 2 \cdot \frac{\left|NG(w_I) \cap NG(w_R)\right|}{\left|NG(w_I)\right| + \left|NG(w_R)\right|}$$

Basierend auf einem Abstandsmaß $d(w_I, w_R)$ kann die Metrik für Fehlerfreiheit auf Attributwertebene folgendermaßen definiert werden:

$$Q_{Fehl.}(w_I, w_R) := 1 - d(w_I, w_R)$$

Die Fehlerfreiheit eines Attributwertes wird somit mit dem Maximalwert von eins bewertet, falls der Attributwert im Informationssystem mit der modellierten Ausprägung der Realwelt-Entität (vollständig) übereinstimmt und das verwendete Abstandsmaß $d(w_I, w_R)$ den Wert null liefert. Bei einer Abweichung zwischen w_I und w_R fällt der Wert der Metrik je nach verwendetem Abstandsmaß geringer aus.

Allgemein ist bei Verwendung der Metrik für Fehlerfreiheit zu berücksichtigen, dass im Vorfeld möglicherweise (automatisierte) Data-Cleansing-Maßnahmen durchgeführt werden müssen. Dabei ist es insbesondere notwendig, dass eindeutig interpretierbare Abkürzungen über den gesamten Datenbestand hinweg „glatt gezogen" und vervollständigt werden, damit die zugrunde liegenden Abstandsmaße und somit die darauf basierende Metrik richtig ausgewertet werden. Ein Beispiel für eine solche Maßnahme ist das Ersetzen der Abkürzung „Str." durch „Straße". Nur durch Transformationen dieser Art kann sichergestellt werden, dass

die Abstandsmaße korrekte Attributwerte auch als solche identifizieren und angemessene Ergebnisse liefern.

Die Messung der Fehlerfreiheit kann dann direkt auf Basis obiger Metrik in Verbindung mit entsprechenden Abstandsmaßen zur Bestimmung der Fehlerfreiheit auf Attributwertebene erfolgen. Hierbei ist man in der Regel gezwungen, auf Stichproben zurückzugreifen und statistische Verfahren anzuwenden (vgl. z. B. [Helfert 2002]), da ein Abgleich zwischen den Attributwerten im Informationssystem und den tatsächlichen Ausprägungen der Realwelt-Entität erforderlich ist. Dieser Abgleich ist normalerweise nicht ohne weiteres technisch, automatisiert und mit akzeptablem Kostenaufwand für den gesamten Datenbestand durchführbar. Im Falle einer Stichprobe können jedoch bei ausreichend großem Umfang zumindest Schätzer für den Qualitätswert $Q_{Fehl.}(w_I, w_R)$ ermittelt und Rückschlüsse auf den gesamten Datenbestand gezogen werden.

Im nächsten Abschnitt wird eine Metrik für die DQ-Dimension Konsistenz erläutert.

1.4.3 Metrik für die DQ-Dimension Konsistenz

Unter Konsistenz ist die Eigenschaft der Widerspruchsfreiheit des Datenbestandes zu verstehen. Die Überprüfung basiert dabei im Folgenden auf *logischen* Zusammenhängen, die für die betroffene Datenmenge gelten sollen und durch die Regelmenge $\mathfrak{R}$ repräsentiert werden. Regeln, die auf statistischen Zusammenhängen beruhen und somit nur bestimmten Signifikanzniveaus genügen (d. h. im betrachteten Datenbestand ist der statistische Zusammenhang nicht notwendigerweise exakt und vollständig erfüllt), werden im Weiteren nicht betrachtet. Die Datenmenge ist demnach konsistent, wenn sie $\mathfrak{R}$ entspricht vice versa. Die Vorteile der im Weiteren vorgestellten Metrik liegen insbesondere in der Interpretierbarkeit, die durch Vermeidung der Quotientenbildung und die Wahrung der Kardinalität gewährleistet ist. Die resultierenden Werte der Metrik (auf Relationen- und Datenbankebene) sind dabei als prozentualer Anteil der untersuchten Datenmenge zu verstehen, der hinsichtlich der Regelmenge $\mathfrak{R}$ konsistent beziehungsweise regelkonform ist. Im Gegensatz zu anderen Ansätzen wird dabei auf Attributwert- und Tupelebene keine Priorisierung und Gewichtung innerhalb der Regelmenge vorgenommen, sondern lediglich zwischen konsistent und nicht konsistent im Sinne einer 0-1-Entscheidung differenziert. Dies entspricht dem obigen Verständnis von Konsistenz auf Basis *logischer* Zusammenhänge und verbessert die Ergebnisinterpretation.

Im Weiteren wird die Metrik für Konsistenz nur auf Attributwert- und Tupelebene vorgestellt. Sie kann jedoch analog zur Metrik für Vollständigkeit ebenfalls auf Relationen- und Datenbankebene definiert werden.

Sei w ein Attributwert im Informationssystem und $\mathfrak{R}$ eine $|\mathfrak{R}|$-elementige Menge von Konsistenzregeln, die auf das entsprechende Attribut angewendet wird. Dabei

liefert jede Konsistenzregel $r_s \in \Re$ (s = 1, 2, …, $|\Re|$) den Wert null, falls der entsprechende Attributwert der Konsistenzregel genügt. Andernfalls ergibt die Auswertung der Regel den Wert eins:

$$r_s(w) := \begin{cases} 0 & \text{falls } w \text{ der Konsistenzregel } r_s \text{ genügt} \\ 1 & \text{sonst} \end{cases}$$

Daraus ergibt sich die Metrik zur Bewertung der Konsistenz eines einzelnen Attributwertes:

$$Q_{Kons.}(w, \Re) := \prod_{s=1}^{|\Re|} (1 - r_s(w))$$

Diese nimmt den Wert eins an, falls der Attributwert alle in der Regelmenge $\Re$ spezifizierten Konsistenzregeln erfüllt (d. h. $r_s(w) = 0 \ \forall r_s \in \Re$). Umgekehrt ist der resultierende Wert der Metrik auf Attributwertebene null, falls mindestens eine der spezifizierten Regeln verletzt ist (d. h. $\exists r_s \in \Re: r_s(w) = 1$). Als Konsistenzregel sind dabei unter anderem formalisierte Geschäftsregeln oder domänenspezifische Funktionen denkbar. Hierbei ist z. B. an Konsistenzregeln gedacht, die den Wertebereich eines Attributs überprüfen (z. B. $1067 \leq PLZ$, $PLZ \leq 99998$, $PLZ \in \{0, 1, \ldots, 9\}$[5] oder *Familienstand* ∈ {„ledig“, „verheiratet“, „geschieden“, „verwitwet“}).

Auf Tupelebene ergibt sich folgendes: Sei T ein Tupel und $\Re$ die Menge der vorhandenen Konsistenzregeln r_s (s = 1, 2, …, $|\Re|$), die auf das Tupel und die enthaltenen Attributwerte angewendet wird. Dann ergibt sich die Konsistenz des Tupels in Analogie zur Konsistenz auf Attributwertebene zu:

$$Q_{Kons.}(T, \Re) := \prod_{s=1}^{|\Re|} (1 - r_s(T))$$

Das Ergebnis der Metrik hängt dabei zum einen von Konsistenzregeln ab, die lediglich einen einzelnen Attributwert betreffen. Zum anderen können auch Regeln einfließen, die sich auf mehrere Attributwerte oder das ganze Tupel beziehen. Die Metrik auf Tupelebene wird dabei dahingehend „bottom up“ entwickelt, dass diese auch alle Konsistenzregeln und damit auch die Bewertung der Konsistenz auf Attributwertebene umfasst. Falls somit ein Attributwert eines Tupels nicht konsistent bezüglich der Regeln auf Attributwertebene ist, so wird das betrachtete Tupel auch auf Tupelebene als nicht konsistent bewertet. Sind im Gegensatz dazu die Konsistenzregeln für alle einzelnen Attributwerte eines Tupels erfüllt, so müssen zudem auch alle Konsistenzregeln auf Tupelebene erfüllt sein, damit die Konsistenz gewährleistet ist. Ist anderenfalls mindestens eine Regel, die mehrere Attributwerte des Tupels (gleichzeitig) betrifft, nicht erfüllt, so erfolgt (insgesamt) eine Bewertung als nicht konsistent.

Zusammenfassend wird ein Tupel somit nur dann als konsistent hinsichtlich der Regelmenge $\mathfrak{R}$ betrachtet, falls alle Regeln erfüllt werden ($r_s(T) = 0 \; \forall r_s \in \mathfrak{R}$). Ansonsten ergibt sich $Q_{Kons.}(T, \mathfrak{R})$ zu null, egal ob eine Regel oder mehrere verletzt werden ($\exists r_s \in \mathfrak{R}: r_s(T) = 1$). Als Konsistenzregeln auf Tupelebene sind dabei neben denen, die bereits auf Attributwertebene zulässig sind, zusätzlich attributübergreifende Regeln und Zusammenhänge wie z. B. (*Aktuelles Datum* – *Geburtsdatum* < 16 Jahre) $\Rightarrow$ (*Familienstand* = „ledig“) denkbar.

Die Messung der Konsistenz kann wiederum direkt mit Hilfe obiger Metrik in Verbindung mit entsprechenden SQL-Abfragen zur Prüfung der Konsistenzregeln erfolgen. Die Regeln auf Attributwert- und Tupelebene können dabei unter anderem unter Einbeziehung der Fachseiten auf Basis von Wertebereichen, Geschäftsregeln und logischen Zusammenhängen generiert werden.

Im folgenden Abschnitt wird die Metrik für die DQ-Dimension Aktualität entwickelt.

1.4.4 Metrik für die DQ-Dimension Aktualität

Unter Aktualität wird hier die Eigenschaft der Gegenwartsbezogenheit des Datenbestandes verstanden, d. h., inwiefern die im System erfassten Werte den aktuellen Gegebenheiten in der Realwelt entsprechen und nicht veraltet sind. Die Überprüfung basiert dabei – im Gegensatz zur Fehlerfreiheit – auf wahrscheinlichkeitstheoretischen Betrachtungen, um eine automatisierte Messung zu ermöglichen. Aktualität kann in diesem Zusammenhang als jene Wahrscheinlichkeit interpretiert werden, mit welcher die untersuchten Datenwerte noch aktuell sind. In dieser Interpretierbarkeit liegt auch der Vorteil der entwickelten Metrik im Vergleich zu existierenden Metriken, bei denen eine (wahrscheinlichkeitstheoretische) Interpretation der resultierenden Werte nicht möglich ist bzw. nicht vorgenommen wird. Die Metrik für Aktualität wird nur für die Attributwertebene vorgestellt, ist jedoch – analog zu oben – auch auf den anderen Ebenen definiert.

Sei A ein Attribut, w ein entsprechender Attributwert im Informationssystem und $Alter(w, A)$ das Alter des Attributwertes, das sich aus dem Zeitpunkt der Messung und dem Zeitpunkt der Datenerfassung errechnen lässt. Des Weiteren sei $Verfall(A)$ die (ggf. empirisch ermittelte) Verfallsrate von Werten des Attributs A. Diese gibt den Anteil an Datenwerten des entsprechenden Attributs an, der durchschnittlich innerhalb einer Zeiteinheit inaktuell wird. Dann stellt sich die Metrik für Aktualität auf Attributwertebene wie folgt dar:

$$Q_{Akt.}(w, A) := exp(-Verfall(A) \cdot Alter(w, A))$$

Unter der Annahme, dass die Gültigkeitsdauer der zugrunde liegenden Datenwerte exponentialverteilt mit dem Parameter $Verfall(A)$ ist, stellt der Wert $Q_{Akt.}(w, A)$ dabei die Wahrscheinlichkeit dar, mit welcher der vorliegende Attributwert w noch den aktuellen Gegebenheiten entspricht. Bei der Exponentialverteilung han-

delt es sich um eine typische Lebensdauerverteilung, die sich insbesondere im Rahmen der Qualitätssicherung bewährt hat.

Bei Attributen wie z. B. *Geburtsdatum* oder *Geburtsort*, die sich in der Realwelt nie ändern, gilt $Verfall(A) = 0$ und die Metrik für Aktualität ergibt sich somit grundsätzlich zu eins:

$$Q_{Akt.}(w, A) = exp(-Verfall(A) \cdot Alter(w, A)) = exp(-0 \cdot Alter(w, A)) = exp(0) = 1$$

Zudem wird die Aktualität von Attributwerten, die zum Betrachtungszeitpunkt neu erfasst werden – d. h. $Alter(w, A) = 0$ – ebenfalls mit eins bewertet:

$$Q_{Akt.}(w, A) = exp(-Verfall(A) \cdot Alter(w, A)) = exp(-Verfall(A) \cdot 0) = exp(0) = 1$$

Die erneute Erfassung eines Attributwertes wird somit als Aktualisierung eines bereits vorhandenen Attributwertes interpretiert.

Insgesamt ist festzuhalten, dass das Metrikergebnis und damit auch die DQ für ein bestimmtes, festes Alter umso geringer sind, je höher beim entsprechenden Attribut die Verfallsrate ist. Umgekehrt nimmt bei zunehmendem Alter die Wahrscheinlichkeit, dass der entsprechende Attributwert noch gültig ist, und somit das Metrikergebnis für die Aktualität auf Attributwertebene ab.

Für die praktische Anwendung der Metrik ist es notwendig, für jedes Attribut den Parameter $Verfall(A)$ der Wahrscheinlichkeitsverteilung festzulegen. Dieser ist als Verfallsrate zu verstehen und gibt an, welcher Datenanteil bezogen auf das jeweilige Attribut innerhalb einer Zeiteinheit inaktuell wird. Eine Verfallsrate von 0,2 drückt beispielsweise aus, dass im Laufe einer Periode von 100 Attributwerten des entsprechenden Attributs im Durchschnitt 20 Werte inaktuell werden. Dabei kann entweder auf Erfahrungswerte, statistische Werte (bspw. veröffentlichte Scheidungsraten des Statistischen Bundesamts als Grundlage zur Schätzung der Verfallsrate des Werts „*verheiratet*“ des Attributs „*Familienstand*“) zurückgegriffen oder mittels eigener Stichprobenuntersuchungen eine Schätzung vorgenommen werden. Betrachtet man z. B. eine Stichprobe vom Umfang M und misst für die entsprechenden Ausprägungen der Realweltobjekte die Änderungs- beziehungsweise Verfallszeitpunkte z_u ($u = 1, 2, \ldots, M$), dann ergibt sich ein im Sinne der Statistik erwartungstreuer Schätzer für den Verfallsparameter der zugehörigen Exponentialverteilung zu

$$\frac{M}{\sum_{u=1}^{M} z_u}.$$

Die Umsetzung der Messung hinsichtlich der DQ-Dimension Aktualität ergibt sich somit aus obiger Metrik in Verbindung mit den Schätzern für die Verfallsparameter und den Metadaten bezüglich des Zeitpunktes der Datenerfassung.

Der nächste Abschnitt skizziert die Anwendung der Metrik für Aktualität im Rahmen des Customer Relationship Managements eines Mobilfunkanbieters.

1.5 Praktische Anwendung der Metrik für Aktualität

Die praktische Anwendung der Metriken erfolgte im Rahmen des Kampagnenmanagement-Prozesses eines Mobilfunkanbieters. DQ-Probleme traten dabei u. a. bei der Kundenansprache auf. Diese führten bspw. bei Mailingkampagnen dazu, dass oftmals keine korrekte und individuelle Kundenansprache möglich war, was sich in geringeren Erfolgsquoten niederschlug.

Am Beispiel der Vermarktung einer Tarifoption gestaltet sich die Anwendung der Metrik für Aktualität auf Tupelebene wie folgt: Zunächst gilt es, die relevanten Attribute und deren relative Wichtigkeit im Rahmen der Kampagne zu bestimmen. Dies waren die Attribute „Name", „Vorname", „Kontakt" und „Produkt" mit den zugehörigen Gewichtungen von 0,9, 0,2, 0,8 und 1,0. Demzufolge war insbesondere der aktuelle Tarif des Kunden („Produkt") relevant, da eine Inanspruchnahme der Tarifoption nur für spezielle Tarife möglich war; der (korrekte) Vorname des Kunden hatte demgegenüber bspw. weniger Gewicht. Anschließend musste aus dem gegenwärtigen Zeitpunkt und dem Zeitpunkt der Datenerfassung bzw. der letzten Aktualisierung das Alter jedes einzelnen Attributwerts automatisiert berechnet werden. Im nächsten Schritt konnte dann, basierend auf empirisch bzw. mittels Stichprobentests ermittelten Verfallsparametern für die einzelnen Attribute, der Wert der Metrik auf Attributwertebene bestimmt werden. Für ein konkretes Beispiel siehe Tabelle 1:

Tabelle 1: Ermittlung der Aktualität anhand der entwickelten Metrik (Beispiel)

A_i	Name	Vorname	Kontakt	Produkt
g_i	0,9	0,2	0,8	1,0
$Alter(T.A_i,A_i)$ (in Jahren)	0,5	0,5	1,5	0,5
$Verfall(A_i)$ (in 1/Jahr)	0,02	0,00	0,20	0,40
$Q_{Akt.}(T.A_i,A_i)$	0,99	1,00	0,74	0,82

Hier ergibt sich der Wert der Metrik auf Tupelebene durch Aggregation der Ergebnisse auf Attributwertebene unter Berücksichtigung der relativen Wichtigkeiten g_i zu:

$$Q_{Akt.}(T, A_1,..., A_4) = \frac{0{,}99 \cdot 0{,}9 + 1 \cdot 0{,}2 + 0{,}74 \cdot 0{,}8 + 0{,}82 \cdot 1}{0{,}9 + 0{,}2 + 0{,}8 + 1} \approx 0{,}863\,.$$

Demzufolge liefert die Metrik für Aktualität für das Beispieltupel T einen Wert von 86,3% – d. h. das Tupel ist für den speziellen Anwendungsfall (Vermarktung einer Tarifoption) zu 86,3% aktuell. Derartige Werte können nun im Kampagnenmana-

gement genutzt werden. Bspw. wurden so aufgrund von Erfahrungswerten diejenigen Kunden, die einen Wert kleiner als 20% hatten, erst gar nicht angeschrieben. Auswertungen von früheren Kampagnen hatten beim Mobilfunkanbieter gezeigt, dass bei derartigen Kunden eine Erfolgsquote von nahezu 0 resultiert. Neben diesem kurzen Beispiel für die Anwendung der Metrik, bei dem die Kosten für die Kampagne gesenkt wurden, konnten eine Reihe weiterer DQ-Analysen durchgeführt werden, um Kosten zu sparen oder den Nutzen zu erhöhen.

Insgesamt konnte beim Mobilfunkanbieter duch die Anwendung der Metriken ein direkter Zusammenhang zwischen den Ergebnissen der DQ-Messung und den Erfolgsquoten von Kampagnen hergestellt werden. Dies hatte zur Folge, dass der Prozess der Kundenselektion für die Kampagnen deutlich verbessert werden konnte. Zudem konnten der Einsatz von DQ-Maßnahmen auf Basis der Metriken gezielter erfolgen und der damit einhergehende ökonomische Nutzen besser abgeschätzt werden.

Der folgende Abschnitt fasst die Ergebnisse zusammen und würdigt diese kritisch.

1.6 Zusammenfassung und Ausblick

Im Beitrag wurde die Fragestellung aufgegriffen, wie DQ adäquat quantifiziert werden kann. Ziel war dabei, Metriken für die DQ-Dimensionen Vollständigkeit, Fehlerfreiheit, Konsistenz und Aktualität vorzustellen, die eine objektivierbare, zielgerichtete und weitgehend automatisierbare Messung auf den Ebenen Attributwert, Tupel, Relation und Datenbank ermöglichen. Dabei wurde im Gegensatz zu bestehenden Ansätzen der Fokus insbesondere auf die Anforderung der Kardinalität der Metriken gelegt, um eine Untersuchung von DQ-Maßnahmen unter Kosten-Nutzen-Gesichtspunkten zu unterstützen. Die Metriken ermöglichen somit eine Quantifizierung der DQ und bilden die Basis für eine ganze Reihe ökonomischer Analysen. So können zukünftige Auswirkungen auf die DQ, wie z. B. zeitlicher Verfall oder die Durchführung von DQ-Maßnahmen, untersucht und damit ex ante Planungswerte mit ex post Messwerten verglichen werden. Dies sowie die Eignung der Metriken konnte in Zusammenarbeit mit Unternehmen bereits für ausgewählte Fälle auch unter praktischen Gesichtspunkten verdeutlicht werden (vgl. [Heinrich & Klier 2006; Heinrich et al. 2007]).

Zukünftig ist darüber hinaus an modellbasierten Ansätzen zur ökonomischen Planung von DQ-Maßnahmen zu arbeiten, für deren Operationalisierung Metriken und Messverfahren für DQ unbedingt erforderlich sind. Daneben sind die vorgestellten Metriken zu erweitern und zu verbessern: Beispielhaft ist hier die Metrik für Konsistenz zu nennen, für die neben logischen Zusammenhängen ebenfalls auch eine fundierte Formulierung für statistisch ermittelte Zusammenhänge erforderlich ist. Darüber hinaus stellen die Weiterentwicklung der Metrik für Aktualität für den Fall, dass die Annahme einer exponentialverteilten Gültigkeitsdauer nicht gerechtfertigt ist, sowie Ansätze zur Aggregation der Bewertungen für die einzel-

nen DQ-Dimensionen zu einem Gesamtqualitätswert weiteren Forschungsbedarf dar.

Literaturverzeichnis

[Ballou et al. 1998] Ballou, D. P.; Wang, R. Y.; Pazer, H.; Tayi, G. K.: Modeling information manufacturing systems to determine information product quality. In: Management Science 44 (1998) 4, S. 462–484.

[Bamberg et al. 2007] Bamberg, G., Baur, F., Krapp, M.: Statistik. Oldenburg 2007.

[Campanella 1999] Campanella, J.: Principles of quality cost. Milwaukee 1999.

[English 1999] English, L.: Improving Data Warehouse and Business Information Quality. New York 1999.

[Eppler 2003] Eppler, M. J.: Managing Information Quality. Berlin 2003.

[Even & Shankaranarayanan 2005] Even, A.; Shankaranarayanan, G.: Value-Driven Data Quality Assessment. In: Proceedings of the 10th International Conference on Information Quality. Cambridge 2005, S. 221-236.

[Feigenbaum 1991] Feigenbaum, A. V.: Total Quality Control. New York 1991.

[Heinrich et al. 2007] Heinrich, B.; Kaiser, M.; Klier, M.: Metrics for measuring data quality - foundations for an economic oriented management of data quality. In: Proceedings of the 2nd International Conference on Software and Data Technologies. Barcelona 2007.

[Heinrich & Klier 2006] Heinrich, B.; Klier, M.: Ein Optimierungsansatz für ein fortlaufendes Datenqualitätsmanagement und seine praktische Anwendung bei Kundenkampagnen. In: Zeitschrift für Betriebswirtschaft 76 (2006) 6, S. 559-587.

[Helfert 2002] Helfert, M.: Planung und Messung der Datenqualität in Data-Warehouse-Systemen. Dissertation. Bamberg 2002.

[Hinrichs 2002] Hinrichs, H.: Datenqualitätsmanagement in Data Warehouse-Systemen. Dissertation. Oldenburg 2002.

[Jarke & Vassiliou 1997] Jarke, M.; Vassiliou, Y.: Foundations of Data Warehouse Quality – A Review of the DWQ Project. In: Proceedings of the 2nd International Conference on Information Quality. Cambridge 1997, S. 299–313.

[Juran 1999] Juran, J. M.: How to think about Quality. In: Juran, J. M.; Godfrey, A. B. (Hrsg.): Juran's Quality Handbook. New York 1999, Kap. 2, S. 1-18.

[Lee et al. 2002] Lee, Y. W.; Strong, D. M.; Kahn, B. K.; Wang, R. Y.: AIMQ: a methodology for information quality assessment. In: Information & Management 40 (2002) 2, S. 133–146.

[Machowski & Dale 1998] Machowski, F.; Dale, B. G.: Quality costing: An examination of knowledge, attitudes, and perceptions. In: Quality Management Journal 5 (1998) 3, S. 84-95.

[Naumann et al. 2004] Naumann, F.; Freytag, J.-C.; Leser, U.: Completeness of integrated information sources. In: Information Systems 29 (2004) 7, S. 583-615.

[Naumann 2007] Naumann, F.: Aktuelles Schlagwort: Datenqualität. In: Informatik Spektrum 30 (2007) 1, S. 27-31.

[Pipino et al. 2002] Pipino, L.; Lee, Y.; Wang, R.: Data quality assessment. In: Communications of the ACM 45 (2002) 4, S. 211-218.

[Redman 1996] Redman, T. C.: Data Quality for the Information Age. Norwood 1996.

[Seghezzi 1996] Seghezzi, H. D.: Integriertes Qualitätsmanagement – das St. Galler Konzept. München 1996.

[Shank & Govindarajan 1994] Shank, J. M.; Govindarajan, V.: Measuring the cost of quality: A strategic cost management perspective. In: Journal of Cost Management 8 (1994) 2, S. 5-17.

[Wang et al. 1995] Wang, R. Y.; Storey, V. C.; Firth, C. P.: A Framework for analysis of data quality research. In: IEEE Transaction on Knowledge and Data Engineering 7 (1995) 4, S. 623-640.

[Wang & Strong 1996] 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.

[Würthele 2003] Würthele, V. G.: Datenqualitätsmetrik für Informationsprozesse. Norderstedt 2003.