

Datenqualitätsmanagement in der Praxis

*Bericht über ein Projekt zum
Datenqualitätsmanagement in der Logistikbranche*

Forschungsbericht GFFT-2008-006

Schlüsselworte (Suchkriterien): Datenqualität, Data Quality, Data Profiling,
Datenqualitätsmanagement

Version 1.0 vom 31.07.2008

Geheimhaltungsgrad: Confidential

Autor: Jens Bleiholder

Herausgeber: GFFT e.V.

GFFT e.V.
Taunusstraße 23
61138 Niederdorfelden

INHALT

| | | |
|----------|---------------------------|----------|
| 1 | Einleitung | 3 |
| 2 | Ansatz | 3 |
| 3 | Ergebnisse | 4 |
| 4 | Empfehlungen | 6 |
| 5 | Literatur | 7 |
| 6 | Kontakt | 7 |

1 Einleitung

Das Thema Datenqualität gewinnt in der öffentlichen Wahrnehmung und insbesondere auch innerhalb von Unternehmen an Bedeutung. Eine strukturierte Vorgehensweise zur Definition, Messung und Verbesserung der Datenqualität ist allerdings in Unternehmen noch sehr wenig verbreitet. So berichtet die Computerzeitung z.B. in ihrer Ausgabe vom 14.1.2008 über die Ergebnisse einer Umfrage unter Firmen, dass das Thema Datenqualität von vielen Firmen als wichtiges Zukunftsthema erachtet wird. Das Bewusstsein für qualitativ hochwertige Daten ist vorhanden, eine strukturierte und nachhaltige Vorgehensweise zur Messung und Verbesserung der Datenqualität aber weitgehend nicht.

Das vorliegende Dokument stellt die Ergebnisse einer Vorstudie zum Datenqualitätsmanagement zusammen mit einem renommierten Logistikunternehmen vor. Es werden einmal die wichtigsten Rollen in einem Datenqualitätsmanagement vorgestellt und ein Vorgehensmodell präsentiert, wie in einem großen Unternehmen Datenqualität kontinuierlich beachtet und verbessert werden kann. Zum zweiten wurde ein konkreter Datenbestand auf Qualitätsprobleme hin untersucht. Das vorliegende Dokument beschreibt das hierbei gewählte Vorgehen.

Die folgenden Aspekte motivieren die Studie:

Nachweisbarkeit, Rechtliche Verpflichtungen: Wenn Auflagen nicht erfüllt werden, weil z.B. Lieferfristen nicht eingehalten werden, müssen möglicherweise Konventionalstrafen gezahlt werden. Neben dem direkten finanziellen Verlust der dadurch entsteht ist auch der Imageverlust und darauf folgend der Verlust von Folgeaufträgen zu berücksichtigen. Sind die Daten zu einer Sendung korrekt im System erfasst, kann im Streitfall nachgewiesen werden, dass Sendungen pünktlich ankamen. Verstößt ein Unternehmen aufgrund falscher Zahlen gegen Berichtspflichten, z.B. in der Bilanz, droht eine Klage, mit unter Umständen negativen finanziellen Folgen und Imageschaden z.B. durch negative Schlagzeilen.

Kundenansprache: Hierbei handelt es sich um das klassische Beispiel Master Data Management und die Dublettenproblematik. Unternehmen sind unter Umständen unter vielen verschiedenen Bezeichnungen in der eigenen Kundendatenbank abgelegt. Betrachtet man nur einen dieser Einträge ist der Gesamtumsatz den man mit dieser Firma tätigt unbekannt. Dies kann dazu führen, dass Kunden nicht angemessen angesprochen, bzw. kein angemessener Rabatt gewährt wird. Auf der anderen Seite können auch keine Rabatte bei anderen Firmen ausgenutzt werden (z.B. bei Fluggesellschaften), weil nicht genau bekannt ist, wie viel Umsatz mit welcher Fluglinie getätigt wird. Eine Konzentration auf 20 Fluglinien, mit jeweils mehr Umsatz, erlaubt das Verhandeln besserer Konditionen. Dies ist nur mit vollständigen und korrekten Kennzahlen möglich.

Die Standardliteratur bietet eine Reihe unterschiedlich motivierter Herangehensweisen unterschiedlicher Granularität und Komplexität zum Datenqualitätsmanagement an (siehe z.B. [Redman 2000], [Olson 2002], [Naumann 2002]). Die Herangehensweisen unterscheiden sich oft nur im Detail; das in diesem Projekt erstellte Vorgehensmodell vereinigt geeignet Elemente der verschiedenen Herangehensweisen. Gemeinsamkeiten der Herangehensweisen sind z.B. die Definition spezieller Rollen und eines festen Prozesses, oft mit Rückkopplung, sowie eine Definition von Datenqualitätskriterien, die wiederum die Definition von Zielen und Projekten unterstützen.

2 Ansatz

In Zusammenarbeit eines Forschungsinstitutes mit einem großen Logistikdienstleister wurden in diesem Projekt zwei Ziele verfolgt: die Definition eines Datenqualitätsmanagementprozesses und die prototypische Durchführung eines ersten Projektes anhand dieses Prozesses.

Das Ziel des ersten Teilprojektes war es, die Fachkenntnisse des Forschungsinstitutes nutzbringend dem Logistikdienstleister zur Verfügung zu stellen und in Zusammenarbeit mit den Fachleuten aus

dem Unternehmen einen Prozess zu definieren, der Datenqualitätsmanagement im Unternehmen verankert.

Mit korrekten Zahlen kann ein Unternehmen strategisch richtig entscheiden. Ziel des zweiten Teilprojektes war es zu versuchen, ohne detaillierte Kenntnisse der Prozesse des Unternehmens von außen herauszufinden, welche Qualitätsprobleme dort verwendete Daten aufweisen. Der Logistikdienstleister stellte hierzu einen ausgewählten Datensatz zur Analyse zur Verfügung und unterstützte beim Verständnis der Daten. Das HPI führte dann eine unabhängige und unvoreingenommene Analyse der Daten durch. Dabei wurde teilweise auf vorhandene kommerzielle Tools zurückgegriffen, sowie auf Eigenentwicklungen des Lehrstuhls.

3 Ergebnisse

Rollen im Datenqualitätsmanagement

Im Rahmen eines Datenqualitätsmanagements sind mehrere Rollen zu unterscheiden. Generell gilt, dass es nicht notwendigerweise natürliche Personen sind, die eine Rolle einnehmen, sondern auch Gruppen, oder Gremien, wie z.B. Arbeitskreise sein können, die diese Rollen ausfüllen. Der **Sponsor** stellt Geld und Beziehungen im Unternehmen zur Verfügung. Er schafft auf den wichtigen Entscheidungsebenen das Bewusstsein für das Thema Datenqualität und setzt die Einführung eines Datenqualitätsmanagements durch. Der **Datenqualitätsbeauftragte** ist unabhängig von den Fachabteilungen, zentral organisiert, hält aber den Kontakt zu den dezentralen Stellen. Er steuert das Datenqualitätsmanagement und führt eine Liste mit möglichen Projekten. Der **Datenverwalter** ist Mitarbeiter einer Fachabteilung und hat vollen Zugriff auf die benötigten Daten. Man kann ihn auch als den Dateneigentümer bezeichnen. Er ist der Ansprechpartner des Datenqualitätsbeauftragten in der Fachabteilung. Auch der **Datenanalyst** ist Mitarbeiter einer Fachabteilung und besitzt genau wie der Datenverwalter Domänenwissen. Hinzu kommt das Wissen um die technischen Feinheiten, also wie die Daten technisch gespeichert werden.

Allgemeines Vorgehensmodell

Datenqualitätsmanagement besteht darin, durch Datenqualitätsprojekte die Datenqualität im Unternehmen schrittweise und andauernd zu steigern, bzw. fest definierte Datenqualitätsziele zu erfüllen. Zu Beginn steht die Definition eben dieser Ziele und die Etablierung eines Vorgehensmodells. Einzelne Komponenten des Vorgehensmodells werden nun im Einzelnen vorgestellt:

- **Unternehmensweite DQ-Ziele**
Unternehmensweite Datenqualitätsziele bestimmen mögliche Projekte. Sie sollten von Sponsor und Datenqualitätsbeauftragten gepflegt und bestimmt werden. Einflüsse können auch aus den Fachabteilungen oder von extern kommen.
 - Abstrakte Ziele (Beispiele):
 - Verbesserung der Kundenzufriedenheit.
 - Verbesserung der verwendeten IT-Systeme, damit diese die Wirklichkeit korrekt abbilden.
 - Erhöhung des Bewusstseins der Mitarbeiter für Datenqualität und die Wirkungen schlechter Datenqualität („Katastrophenszenarien“).
 - Konkrete Ziele (Beispiele):
 - Verringerung falscher Datumsabgaben um X%.
 - Verbesserung der Zuverlässigkeit der Daten aus Land XYZ.
- **Mögliche Projekte**
Eine Liste mit möglichen Projekten wird vom Datenqualitätsbeauftragten gepflegt und richtet sich nach den DQ-Zielen, den Erfahrungen vergangener Projekte und Anforderungen aus den Fachabteilungen. Die Liste wird gespeist aus Ideen aus den Fachabteilungen, von externen Quellen wie z.B. rechtlichen Anforderungen, sowie aus den DQ-Zielen.
- **Auswählen**
Anhand vorher festzulegender Kriterien (z.B. Dringlichkeit, Kosten/Nutzen, Erfolgchance) werden Projekte aus der Liste möglicher Projekte ausgewählt. Die zu untersuchenden Aspekte der Datenqualität (Datenqualitätsdimensionen) werden bestimmt.

- **Messen**
Die zu untersuchenden Daten und Prozesse werden den Datenanalysten zur Verfügung gestellt. Die Daten werden durch den Analysten untersucht. Probleme, Unstimmigkeiten und Inkonsistenzen sowie sonstige Probleme mit der Datenqualität in den Daten werden identifiziert und dokumentiert.
- **Analysieren**
Anhand der erhobenen Daten und identifizierten Probleme wird versucht, Ursachen für die Probleme zu finden.
- **Verbessern**
In diesem Schritt werden die im Analyseschritt festgelegten Verbesserungsmaßnahmen durchgeführt, um die Ursachen der Probleme zu beseitigen.
- **Berichten**
Die Ergebnisse des Projektes werden dokumentiert und zusammen mit allen erstellten Dokumenten im DQ-Repository aufbewahrt.
- **DQ-Repository (Metadatenrepository)**
Im DQ-Repository werden die Erfahrungen und Berichte der durchgeführten Projekte aufbewahrt. Es handelt sich um das Gedächtnis des Unternehmens in Bezug auf die Datenqualität.
- **Aktualisieren**
In regelmäßigen Abständen findet eine Überprüfung der unternehmensweiten Ziele auf ihre Gültigkeit statt. Genauso wird die Liste möglicher Projekte regelmäßig aktualisiert, es werden Projekte ergänzt, durchgeführte Projekte entfernt, etc.

Untersuchung eines Datenbestandes

Der zur Verfügung gestellte Datenbestand wurde nach den folgenden Gesichtspunkten untersucht. Wir nennen jeweils kurz die Fragen, die unter dem jeweiligen Punkt geklärt werden sollten.

Metadatenebene:

- **Dokumentation:** Wie sind die zur Verfügung gestellten relationalen Daten dokumentiert? Gibt es eine Beschreibung des Schemas und damit auch des Inhalts der einzelnen Felder? Wie aktuell ist die Dokumentation? Wird beschrieben woher die Daten kommen, wie sie entstanden, bzw. erhoben worden sind?
- **Primärschlüssel:** Welche Spalten enthalten eindeutige (unique) Werte bzw. sind Primärschlüssel oder Kandidaten für Primärschlüssel?
- **Abhängigkeiten:** Welche Abhängigkeiten gibt es in und zwischen den Tabellen? Welche Spalten sind als Fremdschlüssel definiert bzw. sind mögliche Kandidaten für Fremdschlüssel? Gibt es eventuell redundante Spalten, also Spalten, die gleiche Informationen enthalten?

Datenebene:

- **Fehlende Werte:** In welchen Spalten enthält die Tabelle keine Werte? Durch die Untersuchung fehlender Werte und deren Verteilungen in Tabellen gewinnt man erste Anhaltspunkte für Probleme. In einem zweiten Schritt muss überlegt werden, was der Grund für fehlende Werte in einem Attribut ist. Eine Aufschlüsselung nach Werten in anderen Feldern (Pivoting, z.B. Aufschlüsselung nach Ländern) liefert genauere Informationen über Probleme, z.B. wenn sich fehlende in einem Land häufen.
- **Häufige Werte:** Welches sind die häufigsten Werte einer Spalte? Hierdurch bekommt man einen ersten Eindruck über typische Werte und im Falle von nur wenigen unterschiedlichen Werten in einer Spalte auch einen Eindruck über mögliche Fehler.
- **Minimale, maximale, mittlere Werte:** Welches sind die minimalen, maximalen und mittleren Werte einer Spalte? Negative Werte bei Spalten in denen nur positive Werte erlaubt oder zu erwarten sind (z.B. Gewichte oder andere Größen) fallen hier besonders auf. Auch besonders große oder besonders kleine Werte lassen auf Probleme schließen. Gibt es generell einen erlaubten Wertebereich für Spalten? Wird dieser eingehalten?

- Offensichtlich falsche Werte: Welche offensichtlich falschen Werte gibt es in den Attributen? Besonders einfach ist dies, wenn fehlerhafte Werte durch den Importprozess in einer ausgezeichneten Spalte vermerkt werden, oder Flags im Fehlerfall gesetzt werden.
- Längenverteilung der Attributwerte: Welche Längenverteilungen gibt es in den Attributwerten? Wie viel Prozent der Werte bestehen aus 2, 3 oder 4 Zeichen?
- Muster in Attributen: Welche Muster gibt es in den Attributwerten? Bestehen die Werte nur aus Zahlen, nur aus Zeichen? Lässt sich ein regulärer Ausdruck finden, der die Daten beschreibt?
- Regeln in den Daten: Welche Regeln der Form „Wenn Wert X in Spalte A, dann Wert Y in Spalte B“ lassen sich aus den Daten ableiten?

Prozessebene:

- Unschärfe Dubletten: Welche Spalten enthalten möglicherweise unscharfe Duplikate? Welche der in den Tabellen beschriebenen Objekte (z.B. Sendungen) enthalten möglicherweise unscharfe Duplikate? Mit dem Begriff unscharfe Duplikate werden unterschiedliche Repräsentationen ein und desselben Objektes bezeichnet, die sich - im Gegensatz zu exakten Duplikaten - in ihren Attributwerten leicht voneinander unterscheiden können. So können z.B. die verschiedenen Zeichenketten „BMW AG“ und „Bayrische Motorenwerke“ ein und dieselbe Firma bezeichnen und bilden somit ein unscharfes Duplikat.
- Ereignisse: Welche Probleme gibt es mit den Sendungen zugeordneten Ereignissen? Wird die durch den Prozess bestimmte zeitliche/logische Reihenfolge (z.B. Abfahrt vor Ankunft) auch eingehalten? Gibt es verpflichtende Ereignisse und sind diese vorhanden?
- Flags: Werden alle zu setzenden Flags auch gesetzt? Kann man überprüfen ob sie richtig gesetzt werden, z.B. durch das Vorhandensein bestimmter Werte in anderen Spalten?

Die hier vorgeschlagenen Analysen wurden teils mit existierenden Tools zur Datenanalyse durchgeführt als auch durch einfaches Ausführen von SQL-Anfragen direkt auf dem Datenbestand. Der Abschlussbericht bestand aus einer ausführlichen Zusammenfassung der Ergebnisse sowie aus den Rohdaten in Tabellenform.

4 Empfehlungen

Mit erfolgreicher Beendigung des Projektes lassen sich die folgenden Empfehlungen aussprechen, um ein solches Projekt zum Datenqualitätsmanagement durchzuführen.

- Datenqualität ist ein Thema, das in jeder Firma immer wichtiger wird. Hinter jedem nicht korrekten Datum verbergen sich potentielle, aber auch reale Verluste. Datenqualität wird von vielen Unternehmen bereits jetzt als wichtiges Thema angesehen, Probleme werden jedoch oft nicht proaktiv sondern immer nur reaktiv angegangen. Ein definierter Prozess hilft dabei Datenqualität strukturiert anzugehen und als eigenen Prozess im Unternehmen zu verankern. Dabei ist in besonderem Maße darauf zu achten, wer im Unternehmen angesprochen wird, um Datenqualitätsmanagement erfolgreich zu etablieren.
- Bei der Durchführung eines Projektes zur Datenanalyse ist darauf zu achten, dass der untersuchte Ausschnitt der Daten nicht zu klein, aber auch nicht zu groß ist. Ist die Datenmenge zu klein, sind gefundene Fehler nicht unbedingt repräsentativ, ist die Datenmenge zu groß, ist eine schnelle Analyse nicht gegeben. Gibt es eine zeitliche Komponente lohnt es sich u.U. statt eines ganzen Monats, ein Sample eines Jahres zu untersuchen. Die hängt jedoch stark von der Art der zu untersuchenden Daten ab und bringt die weitere Schwierigkeit der Auswahl des geeigneten Samples mit sich.
- Wird die Analyse wie in diesem Fall durch einen fachfremden Datenanalyseexperten durchgeführt, ist unbedingt ein fachlicher Ansprechpartner erforderlich, um in relativ kurzer Zeit ein gutes Verständnis der zu analysierenden Daten zu bekommen. Hintergrundwissen um die Prozesse im Unternehmen (was wird da repräsentiert und wie werden die Daten erhoben) ist dabei genauso wichtig wie simple Erklärungen, was welche Abkürzungen und Begriffe bedeuten.

- Abschließend ist eine Präsentation der unabhängig ermittelten Ergebnisse unbedingt erforderlich. So können die Experten aus dem Unternehmen, die regelmäßig mit den Daten arbeiten von den Ergebnissen profitieren und schlussendlich das Projekt evaluieren. In den meisten Fällen werden sich dabei unter den Ergebnissen bereits bekannte Probleme, aber oft auch neue, noch nicht erkannte Probleme finden. In jedem Fall ist ein wichtiges Ergebnis die Quantifizierung auch der bisher schon bekannten Probleme.

5 Literatur

[Redman 2000] Thomas C. Redman, „*Data Quality – The Field Guide*“, digital press, Boston, 2000

[Olson 2002] Jack E. Olson, „*Data Quality: The Accuracy Dimension*“, Morgan Kaufmann, San Francisco, 2002

[Naumann 2002] Felix Naumann, „*Quality-Driven Query Answering for Integrated Information Systems*“, Springer, Berlin, 2002

[Mielke 2008] Knut Hildebrand, Michael Mielke, Marcus Gebauer, „*Daten- und Informationsqualität*“, Vieweg+Teubner, Mai 2008

6 Kontakt

Prof. Dr. Felix Naumann
Fachgebiet Informationssysteme
Hasso-Plattner-Institut für Softwaresystemtechnik
Prof.-Dr.-Helmert-Str. 2-3
D-14482 Potsdam

Tel.: 0331/5509-280
Fax: 0331/5509-287
E-Mail: Felix.Naumann@hpi.uni-potsdam.de
WWW: <http://www.hpi.uni-potsdam.de/naumann>

Jens Bleiholder
Fachgebiet Informationssysteme
Hasso-Plattner-Institut für Softwaresystemtechnik
Prof.-Dr.-Helmert-Str. 2-3
D-14482 Potsdam

Tel.: 0331/5509-282
Fax: 0331/5509-287
E-Mail: Jens.Bleiholder@hpi.uni-potsdam.de
WWW: http://www.hpi.uni-potsdam.de/naumann/mitarbeiter/jens_bleiholder.html