# Root causes affecting data quality in CRM

*Wolfgang Leußer, Klaus D. Wilde*

*Chair of Business Informatics,*
*Catholic University of Eichstaett-Ingolstadt*

## 1 Introduction

An important field of application of business intelligence tools is in Customer Relationship Management (CRM). Especially analyses of a firm's customers' behavior and believes are a prerequisite to develop and realize innovative concepts to initiate and tighten customer relationships. So the quality of available customer data is a decisive factor for companies applying this customer focused strategy. Data quality (DQ) problems often hinder the implementation and cause major negative consequences (Redman 1996, pp. 6-7). To address this problem from a quality management point of view a systematic management, based on a definition of data quality, a continuous quality measurement, and processes to find improvement measures for identified and prioritized problems should be implemented (Wang 1998, pp. 61-65). A root cause analysis is necessary to identify useful and effective improvement measures. Based on a review of literature in quality management we propose a model to identify possible reasons for data quality problems. The evaluation of this model is based on an empirical study.

The remainder of this article is structured as follows: Section 2 provides a brief introduction to data quality management as well as an outline of the relation between CRM and data quality. Section 3 presents a gap model for data quality to systematically derive causes for data quality problems. Section 4 gives a short description of the deduction of the proposed hypotheses. In the following section these are evaluated and the results will be discussed. Finally, Section 6 presents a short discussion of managerial implications and proposes aspects for further research.

## 2 Background

Data or information quality is usually defined out of a user perspective: "fitness for use", which is the ability to satisfy the requirements of users for the intended use in operations, decision making and planning (Juran 1999, p. 2.2). Data quality is considered a manifold construct, consisting of a set of data quality dimensions requir-

ing consumer assessment (Wang and Strong 1996, p. 8). Examples of these dimensions are accuracy, timeliness, reliability, relevancy, objectivity, completeness and consistency (Wang and Strong 1996, p. 16). When the relevant data does not meet the requirements of a user in one or more dimensions this can be defined as data or information quality problem. The expressions "error" or "fault" will be used in this article synonymously. To distinguish data from information the latter is usually seen as data put in a context (English 1999, p. 19). Even though throughout this article the terms "data" and "information" are used interchangeably according to most data or information quality literature.

## 2.1 Relevance of Data in CRM

Data and information quality is a relevant topic in different fields of research, like management, information systems, medicine, pedagogy, legal studies or linguistics, just to name a few examples (Eppler 2006, p. 9). In our study we focus on the field of CRM as a dynamically growing area for data usage to support decision making and operational processes.

This approach follows the definition of CRM by Payne and Frow (2005, p. 168): "CRM is a strategic approach that is concerned with creating improved shareholder value through the development of appropriate relationships with key customers and customer segments. CRM unites the potential of relationship marketing strategies and IT to create profitable, long-term relationships with customers and other key stakeholders. CRM provides enhanced opportunities to use data and information to both understand customers and cocreate value with them."

While operational CRM directly supports customer-related business processes, in analytical CRM customer contacts and reactions are recorded systematically and evaluated for continuous optimization of customer-related business processes (Arndt and Langbein 2002, p. 47). Primarily analytical models are created to segment customers, to classify customers, or to rate them (Berry and Linoff 2005, pp. 9-11).

Customer data from several separate internal and external databases, like marketing information provider or cooperating companies, have to be integrated to realize these usage scenarios (Berry and Linoff 2005, pp. 141-150). Therefore a data warehouse is usually implemented to provide access for users across the company. Tools like data mining and Online Analytical Processing are applied for analytical purposes (Eppler 2006, pp. 125-126).

## 2.2 Importance of Root Cause Analysis

A number of Data Quality Management (DQM) concepts have been developed to improve data quality on a sustained basis. Total Data Quality Management (TDQM) has been proposed by Wang (1998, pp. 61-65) as a theoretical foundation for data quality. Other approaches are Data Quality for the Information Age

(Redman 1996, pp. 114-117), Total Quality data Management (TQdM) (English 1999, pp. 69-82), and Proactive Data Quality Management (Helfert 2002, pp. 100-112).

All these management concepts have in common a root cause analysis of detected DQM-problems as a prerequisite to specify improvement measures. Wang designs a four step ongoing TDQM-Cycle. In a first step information quality dimensions are defined. Then a measurement using DQ metrics is executed to identify DQ problems. In an analytical step, root causes for DQ problems are identified and the impact of poor quality information, as basis for prioritization, is calculated. Finally, the improvement component provides techniques to improve DQ (Wang 1998, pp. 61-65).

As with all complex problems, there is no single cause behind DQ problems. Especially the complexity of data quality problems does not allow a restriction to a single cause. So a systematic approach is needed to guide the process of root cause analysis. A "cause" in this article is defined as an event leading as a direct consequence to another event, in our context a data quality problem.

## 3 Causes for data quality problems

In the research area of service management Parasuraman et al. (1985, pp. 44-46) conceptualized a model of service quality identifying gaps responsible for the perceived quality, known as gap model. Guided by this model a list of constructs (causes as we call them in this article) was identified theoretically and during an exploratory study (Zeithaml et al. 1988, pp. 37-46).

In our research on root causes for DQ problems we followed this widely accepted and applied approach. We developed a gap model for data by taking into account the characteristics of data. The main aspects to consider are the designing processes and the realization of the business processes as well as information systems, the processes of data production, storage, maintenance, and usage, and inherent characteristics of data (Heinrich et al. 2009, pp. 3-4; Strong et al. 1997, p. 38). Especially the characteristics of data as an intangible asset, the possibility to store, the obsolescence during time, and the determination of the quality level for the intended use have to be taken into account (Wang 1998, pp. 59-60). In the following we will explain the identified gaps and identify causes for DQ problems.

Regarding production and usage processes two distinct groups can be identified: Users of data and providers of data. The latter encompasses producers and custodians, responsible for data storage and maintenance (Strong et al. 1997, p. 38). Following the definition of data quality as "fitness of use" the first step to define the requirements for data is to capture the explicit and implicit expectations of the users (Wang 1998, p. 61). These may be influenced by the needs to support tasks in daily business, past experience and recommendations by other users.

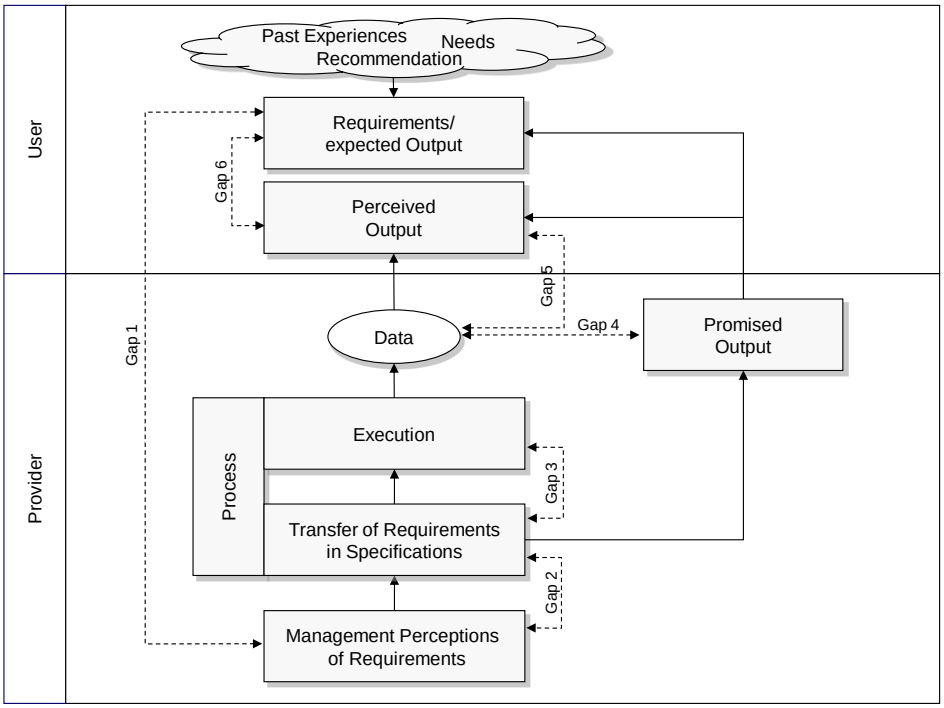


**Figure 1: Gap model for Data**

So the first gap can be defined as follows:

*Gap 1: Difference between user expectations and management's perceptions of user requirements*

This gap can be a result of three causes. As a first step in the development of information systems a systematic requirements management aims to plan and provide data as needed for usage, anticipating future developments.

In practice the collection of those requirements is often *focused on the support of operational processes*. E.g. a sales order has to be processed, so the relevant information to handle this order is taken into account. The need to record other valuable data for analytical purposes, that is available during such processes, is ignored (English 1999, p. 20). Another aspect is the storage of data in text fields instead of a coded format as required for easy analytical use (Strong et al. 1997, pp. 42-43).

Especially in CRM the need for data analysis changes becomes obvious, when marketing and sales concepts have to adapt to changing customer behavior. These changes lead to new or enhanced requirements for provided data. Hence this *changed use of data* is the second source of data quality problems if the requirement management cannot cope with this demand (Strong et al. 1997, pp. 44-45).

The third cause for this gap is a *lack of data quality culture* in companies, for example a low perceived importance of data quality or a lack of responsibility for data. Necessary resources will not be allocated to data quality management, if a lack of sponsorship for data quality exists (Radcliff 2006, p. 4).

Regarding the next gaps, two steps in data production have to be taken into account. Firstly, based on the user requirements, appropriate business processes have to be designed and specifications for supporting IT systems have to be set. Secondly, data production is the realization of the data schemata and the information systems and the use of these according to the design (Helfert 2002, pp. 66-68). Both steps are relevant for the data quality delivered, so gap 2 can be defined as:
*Gap 2: Difference between the management's perceptions of requirements and specifications for output, processes, and systems.*

The following reasons for the design gap causing data quality problems can be identified: *Business processes and processes in IT systems can be designed insufficiently* to support the requirements of data producers and users (Strong et al. 1997, p. 40). If, for example, quality checks in the process of generating e-mail addresses for newsletter campaigns are not integrated in a consistent manner, incorrect addresses will be stored in the customer database, leading later to bounces while using his data.

Derived from business processes, adequate supporting IT systems have to be designed. An *inadequate design of IT systems*, such as redundant data storage or not provided acquisition options for needed data, can lead to problems using this data (Strong et al. 1997, p. 41). Redundant data capture in operative systems as an example may produce inconsistent data in the customer data base and additional work is needed when clarifying differences and building trust (English 2006, pp. 24-25). Redundancy often is not the result of design, but has developed through mergers and acquisitions or changing organizational structures (Radcliffe 2006, p. 4, Berry and Linoff 2000, p. 145).

In CRM *intermediaries* often support the customer communication, e.g. external call center or e-mail service provider. Insufficient integration of these intermediaries in IT processes can cause problems with the captured data, when separate IT systems are designed, and consequently data conversion and transfer are necessary. *Media conversion* is another reason for errors in this context (Strong et al. 1997, pp. 41-42).

Data from different internal and external sources is integrated in a customer data warehouse to realize a holistic view on the customer and its needs and behavior. The *influence on the quality of an external data supply* and the information about the processes behind are limited (Berry and Linoff 2000, pp. 149-150). Finally, the design of the data exchange processes and the data quality assurance while integrating external data, influence the quality perceived later by the data user.

Within a company the *management sets guidelines*, like the availability of financial and human resources and incentive schemes. These guidelines limit resources available for data quality assuring activities, for designing business processes, for IT systems and later during the execution processes (English 1999, p. 71). If e.g. the time, a call center agent can spend on assuring data during a customer contact, is restricted, the possibility to check and refresh data is restrained. Setting guidelines contrarily to the aim to produce a high level of data quality within the design or the execution is another reason for data quality problems.

*Gap 3: Differences between specifications and execution in the process.*

Several reasons causing a gap between the specification to meet user requirements and the realization in the design of IT systems and their use in daily business processes have already been mentioned above. Some more causes for gap 3 can be identified in the data capture process.

Relevant data for CRM is captured both by employees and customers. Recently the use of self service applications has become of growing importance. If customers are suspicious about the use of their data, the willingness to provide information about themselves is often limited (Treiblmaier and Dickinger 2005, p. 194). This can lead to inconsistent or missing data if faulty data is captured deliberately because of security or privacy requirements (Strong et al. 1997, p. 45). Employees can also cause DQ problems if a typing error occurs or they are refraining from storing available information in IT systems. Neither all of these problems can be detected by software checks, nor an elimination of personal judgments can be a solution (Berry and Linoff 2000, p. 180). Thus errors may occur when either the *customers or the employees provide false data* or omit needed information *accidentally or deliberately*.

According to this is a cause that can be described as "knowing-why" (Lee and Strong 2003, pp. 14-16). If the employees are *not aware of the relevance* of the customer information captured, the quality of this data is likely to be inadequate for further usage. An employee may fill an attribute, e.g. the age of the customer, with default data only to fulfill integrity checks, if he is not aware of the relevance of a specific information captured (Berry and Linoff 2000, p. 180). This can often be hardly detected and corrected during data analysis causing wrong analytical results.

Another source of data quality problems can be an *insufficient knowledge about the correct operation* of IT systems, the "knowing-how" (Lee and Strong 2003, pp. 14-16). Data quality errors may result from a lack of knowledge about the procedures involved in work processes when operating IT systems.

Another cause for a gap between design and execution can occur while handling *operational or analytical processes in IT systems*. Faulty data transfer or data processing are typical examples for errors during this processing (Redman 1996, pp. 159-159. This may be results of an inadequate design, unreliable or unavailable resources, but also of accidental mistakes by operating staff.

Even though data is captured correctly as designed, the data quality can be perceived as insufficient by data users. This may be due to differences between the expectations resulting from promised output and the data provided.

*Gap 4: Difference between outcome and promised outcome.*

By definition relevant semantics are needed to generate useful information from data (English 1999, p. 19). Hence the use of data in analytic and operational processes depends on the *documentation of the meaning of data* and the processes conducted to collect and process these data. Data about data – metadata – provides these essential facts. If this information is missing, incorrect, or outdated the data is not fit for usage (English 1999, p. 409, Berry and Linoff 2000, p. 179).

An incorrect analytical result may also occur if the understood meaning of an attribute by a person, capturing the data, differs from the meaning, which is assumed by a data analyst (Strong et al. 1997, pp. 40-41). These problems often occur in, but are not limited to, intercultural contexts and distributed systems. Thus *insufficient exchange of* information between data users, data collectors and the management may cause data quality problems.

*Gap 5: Difference between provided and perceived outcome.*

As living conditions are subject of constant changes, keeping stored data up-to-date is demanding. The originally correctly stored information becomes incorrect by and by (Heinrich et al. 2009, p. 5). This is an inherent characteristic of data, esp. master data. Transaction data usually stays stable after entry. So the *obsolescence of data* may cause differences between the quality of data provided and the perception of quality perceived by the data user at a later point (Gap 5).

The perceived data quality is defined in our model as the difference between the perceptions and user expectations (Gap 6). The quality level depends on the size of the five gaps identified above.

Across all gaps a lack of implementation and keeping alive a *systematic data quality management* may be an additional reason of DQ problems. Missing or inadequate implementing elements like defining DQ requirements, continuous measurement of DQ level, analysis of identified and prioritized problems and the implementation of counter measures can lead to inadequate data (English 1999, p. 71).

## 4 Deduction of hypotheses

To evaluate the list of possible causes for DQ problems described above, an online survey was conducted. It was addressed to data users in companies in German speaking countries fulfilling analytical and operational tasks in CRM. User like marketing professionals, analytical experts, call center managers, and sales representatives took part in the study. In addition, experts from relevant IT departments as well as DQ managers were included. To retrieve homogenous data only those respondents were considered for evaluation, who were engaged in B2C commerce and with at least 30.000 B2C customers in their database. The final set of participants, after filtering, included 143 experts. The experts were asked to rate the level of quality of their available customer data in their specific CRM scope and to indicate the level of agreement to the causes for the before identified DQ problems. All measures were conducted on 5-point Likert scales. Respondents could add additional reasons. Except differing terminology, additional causes have not been mentioned.

Several identified causes are more or less associated with each other, as an analysis of the correlations show. So it can be valuable to identify the dimensions behind the cause's attributes, later forming our hypotheses. The Kaiser-Meyer-Olkin measure of sampling adequacy for all 17 surveyed accounts for 0.871. Kaiser

recommends suitable data for factor analysis if this value is greater than 0.8 (Kaiser 1970, p. 405). Thus an exploratory factor analysis, a principal component analysis with varimax rotation, was applied. Based on the Kaiser's criterion and judgment on interpretability four components with eigenvalues greater than 1 have been selected. These four components account for a total of 56.0% of the variance.

**Table 1: Principal components analysis with varimax rotation**

| Rotated Component Matrix | | | | |
|---|---|---|---|---|
| | Component | | | |
| | 1 | 2 | 3 | 4 |
| Focus of IT systems on operational | .697 | .396 | -.018 | -.110 |
| Change of data use over time | .271 | .539 | .147 | .179 |
| Missing enterprise data quality culture | .088 | .615 | .425 | .278 |
| Insufficient processes and quality checks | .520 | .376 | .351 | .049 |
| Inadequate design of IT systems | .646 | .350 | .235 | .070 |
| Media conversion and intermediaries | .729 | -.070 | .367 | .166 |
| Limited influence on external data supply | -.016 | .169 | .716 | .037 |
| Management guidelines | .051 | .731 | .041 | -.009 |
| Deliberate false or missing capture of data | .198 | .149 | .556 | -.372 |
| Accidentally false or missing capture of data | .247 | .133 | .715 | .039 |
| Inadequate knowledge on relevance | .163 | .339 | .550 | .350 |
| Lack of knowledge on correct operation | .294 | .246 | .352 | .239 |
| Operation of IT systems | .731 | .180 | .013 | .174 |
| Insufficient documentation | .347 | .512 | .335 | -.106 |
| Insufficient exchange of information | .275 | .587 | .326 | .111 |
| Obsolescence of data | .150 | .068 | .042 | .807 |
| Lack of a systematic DQM | .396 | .609 | .198 | -.102 |
| | | | | |
| Eigenvalues | 2.91 | 2.89 | 2.54 | 1.20 |

Examining the loadings from the factor analysis, we identified four components: The first component is related to the design and operation of business processes and IT systems. Items like the inadequate design of IT systems, insufficient processes, focus on support of operational processes, and operation of IT systems load high on this factor. Hence we propose hypothesis 1 as follows:

*H1: The quality level of data does depend on the design and operation of business processes and IT systems.*

Aspects of data quality management form the second component, with aspects such as insufficient documentation and exchange of information, a lack of data

quality culture as well as a changing use of data over time. Thus we propose hypothesis 2 as follows:

*H2: The data quality level does depend on the implementation and application of the elements of data quality management.*

All items loading high on component 3 have in common the aspect of data capture. Deliberate or accidentally false or missing data capture, inadequate knowledge on the relevance of data captured for later processes, and the limited influence on supplied external data are part of the data recording process. Hence we propose hypothesis 3 as follows:

*H3: The data quality level does depend on the adequacy of data capture.*

Only one item is loading high on the fourth component. The obsolescence of data is an inherent attribute of data and thus is leading to hypothesis 4:

*H4: The data quality level does depend on the obsolescence of data.*

The next section will provide a test for these four proposed hypotheses. For this purpose the values of the components identified have been calculated using regression analysis.

## 5 Hypothesis tests

A $\chi^2$ test for independence between the proposed hypotheses – respectively the corresponding alternative hypotheses as the test method requires – and the data quality level was used for testing the hypotheses. To conduct this test, the data quality level assessed by the participants on a 5-point Likert scale and the calculated factor values were classified in three nearly equal classes.

**Table 2: Results of the $\chi^2$ test for independence (alternative hypotheses tested)**

| Chi-Square Tests | | | | |
|---|---|---|---|---|
| | Value | Degrees of freedom | Asymptotic Significance (2-sided) | Support |
| Hypothesis 1 | 12.407 | 4 | .015 | - |
| Hypothesis 2 | 12.776 | 4 | .012 | - |
| Hypothesis 3 | .384 | 4 | .984 | + |
| Hypothesis 4 | 4.253 | 4 | .373 | + |

As shown in the table above the alternative hypotheses H1 on the independence of design and operation of business processes and IT systems and H2 on the independence of the implementation and application of data quality management could be rejected with a asymptotic significance of 1.5% and 1.2%, respectively. The alternative hypotheses H3 and H4 on adequacy of data capture and obsolescence of data could not be rejected.

The hypothesis test showed clearly that the data quality level as assessed by data users is not independent of the first two identified groups of causes with a low probability of error. So we can assume that a dependency, an influence of these two groups of causes does exist.

Our hypotheses on the adequacy of data capture and obsolescence of data are not rejected. Apparently these topics are not relevant to the overall data quality level of the participants of our study. At this point it has to be mentioned, that a sample bias is possible. Since participation for this survey was voluntary, experts especially interested in the topic of data quality and in an advanced state of DQM took part in the survey. More than 67% of the respondents rated the overall data quality level as very satisfied or somewhat satisfied. This could be seen as a sign for this bias. These advanced companies in DQM may have trivial easy-to-cure causes, like data capture and also obsolescence of data, under control. More advanced sources of DQ problems, like the items of our first two hypotheses, may have more relevance for these companies and thus a higher share of variance in the survey data.

In addition it can be argued, that the obsolescence of data is an inherent attribute of data. Hence measures can only be taken against the aftermath, but not positioned on the root cause. This may be an explanation to dismiss of dependence between this cause and the data quality level.

## 6 Conclusion and Limitations

Based on research in service management and process quality we developed a gap model for data quality. By applying this model we derived a list of responsible causes for the respective gaps in the context of CRM. The basis for the evaluation constituted a survey, whose respondents were users of data in CRM. Four groups of causes for data quality problems have been identified during analysis: Design and operation of business processes and IT systems, implementation and application of data quality management, adequacy of data capture and obsolescence of data. A dependency with the level of data quality of the first two groups has been confirmed, for the last two hypotheses this has been refused.

The proposed model intends to support practitioners during root cause analysis. An analysis of the reasons along the described gaps for a specific data quality problem can help in specifying efficient and effective improvement measures. The specification of a combined mix of reactive and proactive measures (Heinrich

2007, pp. 546-550) will be guided by a root cause analysis and supported by the model described beforehand.

Finally, a number of limitations need to be taken into consideration. Data quality problems are often related to several quality dimensions. Errors in the data processing can e.g. cause problems in the criteria of correctness, completeness, and relevancy (Eppler 2006, p. 34). Hence the specific dimensions of data quality have to be taken into account to identify reasons for faulty data. Not only dimensions of data quality demand for different levels of data quality, but also the type of data in a specific scenario of use has to be regarded. In our present study we followed an abstract view, e.g. we surveyed an overall data quality rating. Future research on the causes for data quality problems is needed to differentiate our findings for specific dimensions of data quality and for different data types further.

Discrete causes for DQ problems have been identified in this article. Interdependencies exist between several causes as shown by the correlations observed in the empirical study. E.g. guidelines like monetary and human resources set by the management have an influence on the design of processes and IT systems. These dependencies should be a part of future research to clarify the interconnection between causes for DQ problems, and to identify starting points for improvement.

A root cause analysis is a prerequisite to specify improvement measures. Consequently, the next step will be to use the model proposed in this article to analyze the existing improvement measures to lever data quality at the identified points.

## References

Arndt D, Langbein N (2002) Data quality in the context of customer segmentation. In: Fischer C, Davidson B (eds.) Proceedings of the 7th International Conference on Information Quality. Cambridge.

Berry MJA, Linoff GS (2000) Mastering data mining: The art and science of customer relationship management. Wiley, New York.

English L (1999) Improving data warehouse and business information quality: methods for reducing costs and increasing profits. Wiley, New York.

Eppler M (2006) Managing information quality. Springer, Berlin.

Heinrich B (2007) Der effiziente Einsatz proaktiver und reaktiver Datenqualitätsmaßnahmen. Die Betriebswirtschaft 67(76):539-562.

Heinrich B, Klier M, Kaiser M. (2009) A procedure to develop metrics for currency and its application in CRM. ACM Journal of Data and Information Quality 1(1):Article 5.

Helfert M (2002) Proaktives Datenqualitätsmanagement in Data-Warehouse-Systemen. Qualitätsplanung und Qualitätslenkung. Logos, Berlin.

Juran JM (1999) How to think about Quality. In: Juran JM, Godfrey BA (eds.) Juran's quality handbook. McGraw-Hill Professional, New York.

Kaiser HF (1970) A second generation little jiffy. Psychometrika 35(4):401-415.

Lee YW, Strong D (2003) Knowing-Why about data processes and data quality. J Manag Inform Syst 20(3):13-39.

Parasuraman A, Zeithaml VA, Berry LL (1985) A conceptual model of service quality and its implications for future research. J Mark 49(4):41-50.

Payne A, Frow P (2005) A strategic framework for customer relationship management. J Mark 69(4):167-176.

Radcliffe J (2006) Creating the single customer view with customer data integration. Gartner Research, Stamford.

Redman TC (1996) Data quality for the information age. Artech House, Norwoord.

Strong DM, Lee YW, Wang RY (1997) 10 potholes in the road to information quality. Computer IEEE 30(8):38-46.

Treiblmaier H, Dickinger A (2005) Potenziale und Grenzen der internetgestützten Datenerhebung im Rahmen des Customer Relationship Management. Wirtschaftinformatik Proceedings.

Wang, RY (1998) A product perspective on total data quality management. Comm ACM 41(2):58–65.

Wang RY, Strong DM (1996) Beyond accuracy: What data quality means to data consumers. J Manag Inform Syst 12(4):5-34.

Zeithaml VA, Berry LL, Parasuraman A (1998) Communication and control processes in the delivery of service quality. J Mark 52(2):35-48.