Data Quality for Credit Risk Management: New Insights and Challenges Proceedings (Completed paper)

Helen-Tadesse Moges[a], Karel Dejaeger[a], Wilfried Lemahieu[a], Bart Baesens[a,b,c,*]

[a] *Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*
[b] *School of Management, University of Southampton, Southampton, SO17 1BJ, United Kingdom*
[c] *Vlerick Leuven Gent Management School, Leuven, Belgium*

Abstract

Recent studies have indicated that companies are increasingly experiencing Data Quality (DQ) related problems as more and more complex data are being collected. In order to address such problems, literature suggests the implementation of a Total Data Quality Management Program (TDQM) that should consist of the following phases: *data quality definition*, *measurement*, *analysis* and *improvement*. DQ is often defined as "fitness for use". Although "fitness for use" captures the essence of quality, it is difficult to measure DQ using this broad definition. Thus, it has long been acknowledged that the quality of data is best described or analyzed via multiple attributes or dimensions. Yet, despite broad discussion in the DQ literature, there is no one definite set and exact definition of DQ dimensions because DQ is context dependent. Therefore, DQ dimensions should be identified and defined in relation to tasks to achieve a suitable level of DQ. This work identifies the important DQ dimensions for evaluating the quality of the data for credit risk assessment. It also explores the key DQ challenges and causes of DQ problems in financial institutions. Findings of a statistical analysis of an empirical study identify nine important DQ dimensions including accuracy and security for assessing the quality of the data in credit risk databases.

Keywords: Data quality, Information quality, Credit risk, Data definition

1. Introduction

The risk of poor DQ increases as larger and more complex information resources are being collected and maintained [19, 24]. Because most modern companies tend to collect increasing amounts of data, good data management is becoming ever more important. As a response, in the last two decades, the issues of DQ have received a lot of attention, both by organizations worldwide and in academic literature. Several studies are exploring DQ challenges, focusing on DQ measurement and improvement [3–10, 15–38].

In practice, decision makers differentiate information from data intuitively, and describe information as data that has been processed. Unless specified otherwise, this paper uses data interchangeably with information.

DQ is often defined as "fitness for use" which implies the relative nature of the concept [4, 17, 25]. Data with quality for one use may not be appropriate for other use. For instance, the extent to which data is required to be complete for accounting tasks may not be required for sales prediction tasks. More general, data that are of acceptable quality in one decision context may be perceived to be of poor quality in another decision context, even by the same individual [25, 29]. This is mainly because DQ is a multi-dimensional concept in which each dimension represents a single aspect or construct of data items [4, 5, 36] and also comprises both objective and subjective aspects. Some aspects are independent while others depend on the type of task and/or experience of the data user [4, 25]. Moreover,

*Corresponding author. Tel. +32 16 32 68 84; Fax +32 16 32 66 24
Email addresses: Helen.Moges@econ.kuleuven.be (Helen-Tadesse Moges), Karel.Dejaeger@econ.kuleuven.be (Karel Dejaeger), Wilfried.Lemahieu@econ.kuleuven.be (Wilfried Lemahieu), Bart.Baesens@econ.kuleuven.be (Bart Baesens)

in the end, it is the user who will decide whether or not data are fit for use. Therefore, quality of data is considered to be task and expertise dependent. Accordingly, studying DQ in the context of a specific task and expertise is a recognized method [10, 21, 23–25, 37, 38].

1.1. Credit Risk Assessment Task

DQ is of special interest and relevance in a credit risk setting because of the introduction of compliance guidelines such as Basel II and Basel III. Since the latter have a direct impact on the capital buffers and hence safety of financial institutions, special regulatory attention is being paid to addressing DQ issues and concerns. Hence, given its immediate strategic impact, DQ in a credit setting is more closely monitored and/or scrutinized, than in most other settings and/or business units [12, 28].

The credit risk assessment task considered in this study is subjected to Basel II regulation which demands complete transparency and traceability of data, and is primarily concerned with quantifying the risk of loss of principal or interest stemming from a borrower's failure to repay a loan or meet a contractual obligation. Thus, financial institutions are obliged to assess the credit risk that may arise from their investment. They may estimate this risk by taking into account information concerning the loan and the loan applicant.

The quality of the credit approval process from a risk perspective is determined by the best possible identification and evaluation of the credit risk resulting from a possible default on a loan. Credit risk can be decomposed into four risk parameters as described in the Basel II documentation [12]. These are Probability of Default (PD), Loss Given Default (LGD), Exposure at Default (EaD) and Maturity (M). These parameters are used to calculate the regulatory capital. Regulatory capital, also referred to as a buffer capital, is the money set aside to anticipate future unexpected losses due to loan defaults.

$$Regulatory\ Capital = f(PD, LGD, EaD, M)$$

Incorrect parameters may result in a loss and even bankruptcy of the institution. Therefore, minimizing the errors when quantifying the credit risk parameters is a crucial process [2, 13]. Improving the quality of the data used for calculating these parameters is one way of improving the precision of the parameters.

1.2. Total Data Quality Management Program (TDQM)

It is argued in the literature that organizations should implement a Total Data Quality Management (TDQM) program which includes *DQ definition*, *measurement*, *analysis* and *improvement*. This enables them to achieve a suitable DQ level [30].

The *DQ definition* phase is the starting point for a TDQM program identifying all the necessary DQ dimensions to be measured, evaluated and analyzed. Next, the *measurement* process is implemented. The results from the measurement process are *analyzed* and DQ issues are detected. These issues will be taken into account during the *improvement phase*. In this phase, the collection of poor quality data cases is thoroughly investigated and improvement actions are suggested. The four phases are iterated in this order over time as shown in Fig. 1. In fact, the primary goal of DQ assurance is the continuous control of data values and possibly, their improvement [4, 34].

Since the identification of DQ dimensions from a user perspective defines the list of important DQ dimensions for the specific task that have to be assessed, analyzed and improved [4, 34], this empirical study explores and measures the important DQ dimensions in order to assess the quality of the data for the credit risk assessment task. Moreover, it reveals the most frequent challenges of DQ and their causes.

The remainder of the paper is structured as follows. The next section explores the related research on the topic. The third section explains the research methodology used. The fourth section elaborates on the key findings. Finally, the paper ends by elucidating the conclusions and indicating the future research ideas.

2. Related Research

2.1. Identification and Definition of Data Quality Dimensions

DQ problems cannot be addressed effectively without identifying the relevant DQ dimensions. Thus, a first objective of DQ research is to determine the characteristics of data that are important to, or suitable for data consumers [36]. While fitness for use captures the essence of DQ, it is difficult to measure DQ using this broad definition [3, 16].

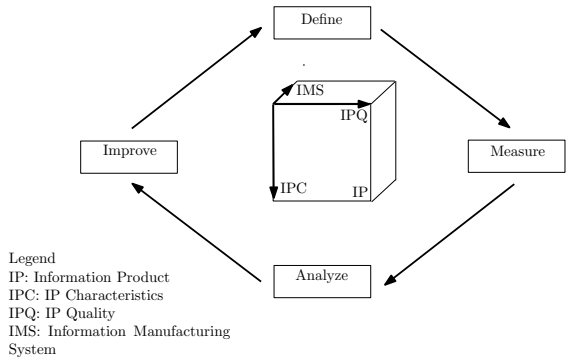


Figure 1: *A schematic overview of the TDQM methodology adopted from Massachusetts Institute of Technology (MIT) [34]*

Thus, it has long been acknowledged that data are best described or analyzed via multiple attributes or dimensions [20, 29, 32]. Yet, despite broad discussion in the DQ literature, there is no one definite set and exact definition of DQ dimensions because DQ is context dependent (see Table 1).

Ref	Intrinsic IQ	Contextual IQ	Representational IQ	Accessibility IQ
[22]	Accuracy, Completeness, Consistency, Validity	Timeliness	Uniqueness	
[36]	Accuracy, Believability, Reputation, Objectivity	Value-added, Relevancy, Completeness, Timeliness, Appropriate amount	Understandability, Interpretability, Concise and consistent representation	Accessibility, Ease of operations, Security
[33]	Correctness, Unambiguous	Completeness	Meaningfulness	
[8]	Accuracy, Precision, Reliability, Freedom from bias	Importance, Relevance, Usefulness, Informativeness, Content, Sufficiency, Completeness, Currency, Timeliness	Understandability, Readability, Clarity, Format, Appearance, Conciseness, Uniqueness, Comparability	Usableness, Quantitativeness

Table 1: *DQ dimensions from the literature ordered according to the framework of Wang and Strong [36]*

Different studies analyzed DQ from a task specific perspective. For example, Zhu and Gauch [40] assessed the DQ of a web page in terms of a DQ framework comprising of six DQ dimensions, namely currency, availability, information-to-noise ratio, authority, popularity, and cohesiveness. They measured the dimensions through the properties of web pages. Similarly, Chien et al. [5] assessed different DQ dimensions in order to evaluate the quality of online product review by customers or other parties. They adopted the definitions of different DQ dimensions for the quality analysis of the online product reviews. For example, they defined objectivity as the extent to which an information item is biased, appropriate amount of data as the extent to which the volume of information in a review is sufficient for decision making, and completeness as the extent to which the information in a review is complete and covers various aspects of a product. Furthermore, they identified objectivity and appropriate amount of information as effective DQ dimensions in identifying product review quality but assessed completeness as a very ineffective DQ dimension to measure the quality of a product review by customers or other parties.

On the other hand, there are a number of studies which identify and define DQ dimensions regardless of the use of the data in order to facilitate the general applicability and comparability of their DQ dimensions across functions. In this regard, Wand and Wang [33] based their definition of DQ on the internal view of information systems (data production and system design processes) because it is context independent. Hence, it supports a set of definitions of DQ dimensions that are comparable across applications. First, they identified different criteria for a real-world system to be properly represented by an information system. Based on these criteria, they defined four deficiencies namely ambiguous representation, incomplete representation, meaningless states, and operation deficiencies. Based on these deficiencies, they summarized different DQ aspects into complete, unambiguous, meaningful, and correct DQ dimensions. In addition, in the same study, they categorized different DQ dimensions from the literature as internal

view (design or operation related) and external view (use or value related), whereby both views were further refined as either system or data related DQ dimensions. Within the internal view, accuracy or precision, timeliness or currency, reliability, completeness and consistency are defined as data related and reliability is defined as a system related DQ dimension. On the other hand, in the external view, timeliness, relevance, content, importance and sufficiency are defined as data related and timeliness, flexibility, format and efficiency are defined as system related DQ dimensions.

Similarly, Wang and Strong [36] analyzed the various DQ dimensions from end users' perspectives but regardless of the use of the data. They conducted a large scale survey to determine and categorize the DQ dimensions. Their analysis began by collecting information from users regarding various DQ descriptors that resulted in over 100 items that were grouped into 20 categories. These were further aggregated into four broad DQ classes: intrinsic (the extent to which data values are in conformance with the actual or true values), contextual (the extent to which data are applicable to the task of the data user), representational (the extent to which data are presented in an intelligible and clear manner), and accessibility (the extent to which data are available or obtainable). Table 1 shows the DQ dimensions considered in different studies and classifies them into the four classes of the Wang and Strong DQ framework [36].

We summarized the most often cited DQ dimensions and their definitions based on a comprehensive literature review and further extended these DQ dimensions based on our pilot survey in Section 3.2. In fact, we adopted the DQ framework of Wang and Strong to classify DQ dimensions [36]. This framework is recognized as the only one that attempts to strike a balance between theoretical consistency and practicability. Furthermore, the framework has been found to be applicable to various domains [9]. The structure of the framework is hierarchical, and it organizes DQ features along fifteen DQ dimensions to comprehend the four broad DQ classes. Table 2 shows the summary of the DQ dimensions. We believe that these DQ dimensions provide a comprehensive coverage of the multidimensional nature of DQ. Hence, in this paper, we used this summary to measure the relevance or applicability of the DQ dimensions for the credit risk assessment task.

2.2. DQ Challenges

As more data are collected and maintained, the risk of poor DQ increases. Multiple data sources, subjective judgment in data production, security/accessibility trade-off, and changing data needs are often mentioned challenges [17]. For example, multiple sources of the same data produce different values for that data. For instance, similar accounting data held in different files are very likely to differ to each other as updating or changing all the files at the same time is not always possible. This is also illustrated by system designers' tendency to avoid having similar data in different files. Similarly, using several different processes is also likely to produce different values for the same information [20]. Like multiple sources of data, subjective judgment of data is also a challenge for DQ. Information production using subjective judgment often produces biased information. Data stored in an organization's database is considered to be a set of facts. However, the process by which these "facts" are collected may involve subjective judgments. For example, the expense codes assigned to indicate different allowances paid to employees by an accountant can be biased by the accountant's knowledge.

In fact, DQ improvement actions require the identification of the causes of data errors and their permanent elimination through an observation of the whole process where data are involved [3, 4, 20]. Data are impacted by many processes, most of which affect their quality to a certain degree [20]. Fig. 2 shows different data inputting and manipulation processes as identified by [20]. Measuring the impacts of data inputting and manipulating processes on DQ is necessary for proper DQ improving activities. In this paper, we identify different DQ challenges and their main causes in financial institutions.

3. Research Methodology

3.1. Research Questions

As described above, data of sufficient quality considered appropriate for one task may not be of sufficient quality for another task [32]. Therefore, identifying and defining DQ dimensions which are relevant to assess the DQ of one specific task is a recognized approach [5, 40]. Also, we believe that identifying the most important DQ dimensions for the specific task is a very crucial step for DQ improvement because it gives a clear direction as to how and where to invest the improving actions. Accordingly, we performed an empirical study which identifies and defines DQ

Cat.	DQ dimensions	Definitions	References
Intrinsic	Accuracy (AC)	The extent to which data are certified, error-free, correct, flawless and reliable	[1, 3, 5, 7, 16, 18, 21, 24–26, 28, 29, 31–34, 36, 38]
	Objectivity (OBJ)	The extent to which data are unbiased, unprejudiced, based on facts and impartial	[1, 5, 7, 16, 18, 21, 25, 31, 34, 36]
	Reputation (REP)	The extent to which data are highly regarded in terms of its sources or content	[3, 5, 16, 18, 21, 25, 31, 36]
Contextual	Completeness (COM)	The extent to which data are not missing and covers the needs of the tasks and is of sufficient breadth and depth of the task at hand	[1, 4, 5, 7, 16, 18, 21, 23–25, 28, 29, 31, 33, 34, 36, 38]
	Appropriate-amount (APM)	The extent to which the volume of information is appropriate for the task at hand	[5, 16, 18, 21, 24, 31, 34, 36]
	Value-added (VAD)	The extent to which data are beneficial and provides advantages from its use	[5, 16, 18, 21, 24, 31, 34, 36]
	Relevance (REL)	The extent to which data are applicable and helpful for the task at hand	[5, 16, 18, 21, 24, 31, 33, 34, 36]
	Timeliness (TIM)	The extent to which data are sufficiently up-to-date for the task at hand	[5, 7, 16, 18, 21, 25, 29, 31, 33, 34, 36, 38]
	Actionable (ACT)	The extent to which data is ready for use	pilot survey
Representation	Interpretable (INT)	The extent to which data are in appropriate languages, symbols, and the definitions are clear	[5, 16, 18, 21, 25, 34, 36]
	Easily-understandable (EU)	The extent to which data are easily comprehended	[5, 16, 18, 21, 25, 31, 34, 36]
	Representational-consistent (RC)	The extent to which data are continuously presented in same format	[5, 16, 18, 21, 25, 28, 31–34, 36]
	Concisely-represented (CR)	The extent to which data is compactly represented, well-presented, well-organized, and well-formatted	[5, 16, 18, 21, 25, 28, 31–34, 36]
	Alignment (AL)	The extent to which data is reconcilable (compatible)	pilot survey
Access	Accessibility (ACC)	The extent to which data is available, or easily and swiftly retrievable	[5, 16, 18, 21, 25, 31, 34, 36]
	Security (SEC)	The extent to which access to data is restricted appropriately to maintain its security	[5, 16, 18, 21, 25, 31, 34, 36]
	Traceability (TRA)	The extent to which data is traceable to the source	pilot survey

Table 2: *Mostly cited DQ dimensions (attributes) and their definitions*

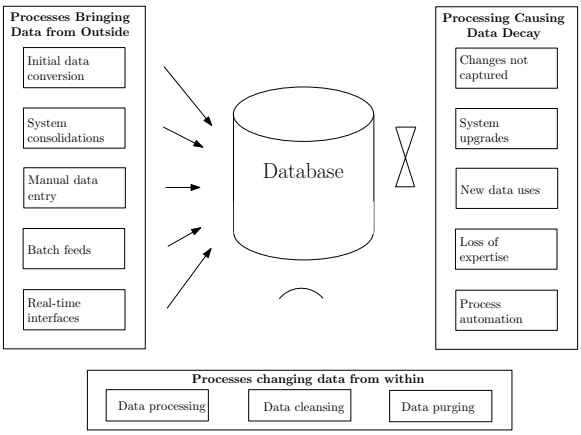


Figure 2: *Different data inputting and manipulating processes as discussed in [20]*

dimensions for the credit risk assessment task by collecting and analyzing data in the form of a survey taken from financial institutions worldwide. The advantage of adopting an empirical approach is that it captures task specific users' requirements [36]. Furthermore, it may reveal characteristics that researchers have not defined as part of a general DQ definition. This empirical study explores and measures the important DQ dimensions in order to assess the quality of the data for the credit risk assessment task. Moreover, it reveals the most frequent challenges of DQ and their causes.

3.2. Pilot study

To test our procedures and determine the clarity of the questions/items, we conducted a pilot study. The pilot study also helped to identify DQ dimensions important to the credit risk assessment task but not shown in Wang and Strong's IQ framework [36]. Subjects were asked to list as many DQ dimensions as they found relevant for their task in addition to the given framework. As a result, "alignment", "actionable" and "traceability" DQ dimensions were identified. The categories and definitions of these four dimensions are also determined by the subjects, Table 2.

The pilot study also served to discover a blueprint ICT architecture desired by financial institutions to enhance DQ as shown by Fig. 3. Data is entered in the system using different interfaces and is checked against validation constraints. Before transferring the data to the analysis warehouse, a staging area is used to verify the completeness, uniqueness, accuracy and consistency of the data. If the data meets the expected quality level, it is directly transferred to the data warehouse. Next, the data can be used to build credit risk models. On the other hand, if the data do not meet the required quality level, the business analyst investigates the problem and traces the data back to the source for corrections. This may then lead to the implementation of new business rules and constraints at the data entry level.

The pilot study finally provided feedback as to the usability and clarity of the study instruments and the simplicity and consistency of the procedures. The subjects of the pilot study were credit risk department managers of three major European banks. They took an average answer time of 30 minutes to finish the questionnaire.

3.3. Empirical evaluation

3.3.1. Design of the Study

The study consists of two parts. In the first part, the respondents are asked 20 questions to identify recurring DQ problems and their magnitude, the motivation for DQ initiatives in their department and if there are any DQ improving activities in place. The questions have exhaustive response categories with an "other" response option.

The second part of the study measures the importance of the DQ dimensions defined in Table 2 in the credit risk assessment task context. Subjects are provided with Table 2 and are asked to rate the importance of the DQ dimensions listed on a scale from 0 to 10 for their task (credit risk assessment), where 0 was not important at all, 5 was somehow important and 10 was extremely important.

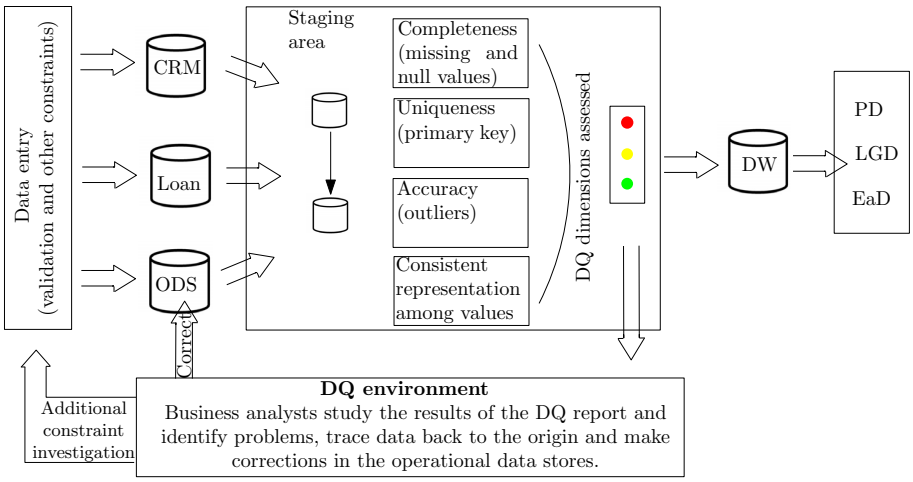


Figure 3: *A blueprint ICT architecture desired by financial institutions to enhance DQ*

3.3.2. Participants

Among a set of 500 financial institutions worldwide determined by multiple business experts, a random subset of 150 were taken. The study subjects are managers of the credit risk department who are responsible for developing or assessing credit risk models and are assumed to have similar experience on the job.

3.3.3. Procedures

A personalized link of the web survey, carrying the company name, was mailed together with a cover letter explaining the nature of the study, the time to complete the study (less than 30 minutes) and the importance of this study. All addresses used were company mail addresses. Finally, of the 150 questionnaires mailed, 64 (an effective response rate of 42.67%) were returned.

3.4. Statistical analysis

In order to test the significance of the obtained results, a number of statistical tests are applied in accordance with the literature. Each of the different tests is assessed at a significance level of 5% unless stated otherwise. Before adopting specific statistical tests, the underlying assumptions made by these tests should be fulfilled. Parametric tests such as Analysis Of Variance (ANOVA) and t-tests both assume the data are normally and IID (Independently and Identically Distributed) [14]. A Jarque-Bera test was adopted to verify the normality of the data. The Jarque-Bera test is a two-sided goodness-of-fit test used to verify the null hypothesis that the data comes from a normal distribution with unknown variance and mean. It has an asymptotic χ^2 distribution with 2 degrees of freedom. The test statistic takes on the following form:

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right)$$

where n represents the sample size, s the sample skewness and k the sample kurtosis.

As the null hypothesis of normality was rejected for ten out of seventeen DQ dimensions at $\alpha = 5\%$, we used non-parametric tests in the remainder of the analysis.

To compare the survey results across DQ dimensions, a Friedman test was adopted which is a non-parametric equivalent to the well known ANOVA test [11]. This test detects differences across all DQ dimensions and is defined as:

$$\chi_F^2 = \frac{12P}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right]$$

with R_j the average rank of DQ dimension $j = 1, 2 \ldots k$ for P banks. Under the null hypothesis, the Friedman test statistic is χ_F^2 distributed with $k - 1$ degrees of freedom, at least when P and k are big enough ($P > 10$ and $k > 5$). In this survey, $P = 33$ and $k = 18$.

Next, since the assumption of equality between all DQ dimensions is rejected, we proceed with a post-hoc Bonferroni-Dunn test. The Bonferroni-Dunn test is a non-parametric alternative to the Tukey test and compares the DQ dimensions with the dimension associated with the highest average rank. The difference between two dimensions is found to be significant if the corresponding average ranks differ by at least the critical difference:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6P}}$$

where q_α is drawn from a Studentized range statistic divided by $\sqrt{2}$. This test also incorporates an additional Bonferroni correction by dividing the confidence level α by the number of comparisons made, $k - 1$, to control for family wise testing, thus resulting in a stronger test.

In case of comparing the sample median between two groups, a (non-parametric) Wilcoxon ranked sum test is used. This test hypothesizes that the data comes from two unknown distributions with equal median [39]. All statistical tests were implemented in Matlab.

4. Results and Discussion

In this section, we present and discuss the key findings of the study. The results of Section 4.1 explain the key DQ challenges, the key causes of DQ problems and the motivations of DQ enhancing activities in financial institutions. In Section 4.2, we present the results of the statistical analysis and identify the important DQ dimensions for the credit risk assessment task.

4.1. DQ issues in financial institutions

Among 64 financial institutions, only 37 are participated in the first part of the survey. In this first part, the respondents were asked to indicate the major DQ challenges or problems that they encounter on a daily basis in financial institutions of which the results are depicted in Fig. 4a. 63% of the respondents indicated that inconsistency (value and format) and diversity of data sources are main recurring challenges of DQ. This indicates that there are many similar data which are kept in different files. Since these data may not be updated or changed at the same time, it is very likely that the data can differ to each other. As a result, decision makers either must rely on their own DQ assessment in order to choose the data source most suited for their decision tasks or must reconcile the different data sources to get one reliable data source. However, we can infer from the results that both processes are not easy. In line with the results in Fig. 4a, Cappiello et al. [4] indicated that mismatches among sources of the same data are a common cause of intrinsic DQ concerns. They identified in their study that mismatches among sources of the same data encourage a subjective DQ assessment by decision makers which gradually affects the intrinsic or objective DQ dimensions. Initially, data consumers do not know the source to which DQ problems should be attributed; they only know that data is conflicting. These concerns initially appear as *believability* problems. Over time, data users assess the *accuracy* of the data for the sources based on experience and personal preferences, which leads to a poor *reputation* for sources considered inaccurate. Hence, less reputable sources are viewed as having little *added value* for the task, resulting in reduced use [4]. However, these less reputable data sources may be of high quality.

In addition to the inconsistency and diversity of data sources, the results in Fig. 4a show that data collection problems and the high costs associated to them are recurring DQ challenges. Data are often produced or maintained by different departments and by different data producers. However, these data are typically also needed by other departments which are not responsible for the production and maintenance of it. This indeed should be facilitated by the system. Yet, collecting all the necessary data for the task is found to be a common challenge as it consumes much of the decision makers' time. Another reported DQ related problem in the results of Fig. 4a are difficulties when making use of the available data. This is related to the relevancy and timeliness DQ dimensions. If data are irrelevant for the task, the task user can't use the data as the data may not have added value. Similarly, if the data are out of date or not timely, the decision makers can't use the data for their decision making activities.

Literature assessed the impact of different data related processes on DQ [17, 20]. We adopted these data related processes shown in Fig. 2 and measured their impact on DQ for financial institutions in the first part of the survey. The results are depicted in Fig. 5. These results indicate that though with different degree, all data related processes have an impact on DQ. Predominantly, manual data entry processes are confirmed to be a major DQ problem cause. This indicates that despite high automation in the institutions, much data enter into databases by people through various interfaces. The most common source of data problems is a person making a mistake while entering data manually. Example of this could be mixing up the age of two customers or not entering any data at all resulting in inconsistent data. This can create a DQ problem which can not easily be identified or explained. These different human manual data entry process problems however can be mitigated by well-designed data entry processes and accompanying instructions [20].

System consolidation and initial data conversion are also confirmed to cause database impurity as shown in the results in Fig. 5. The main common problem in system consolidation is data duplication. Previous research also acknowledges that the data in the consolidated systems often overlap [3]. Similarly, when data are transferred from previous/old systems or paper documents to a new system, data may be lost in the process. This is exacerbated by the fact that there is typically no well recorded metadata or information about the data [3, 17, 20]. In addition to the above identified causes of DQ problems, data mutations taking place internally without being captured by the system and loss of expertise are also indicated as common DQ problem causes as shown in the results in Fig. 5. The changes are known only by those who made the changes and whenever those employees leave, these changes may get lost. This clearly indicates that much of the data exists as tacit knowledge rather than in metadata format. Though very rarely, the respondents also admitted that processes meant to clean impure data in fact caused DQ problems (Fig. 5). Wang and Strong [36] reported that every database has impurity, thus trying to fix one problem may create another one. This finding warns that in order to ensure DQ, the effects of all data related processes need to be taken into account as well.

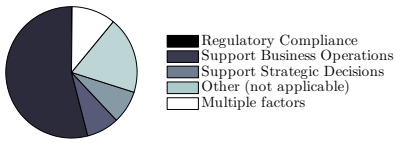
In the first part of the survey, the respondents were also asked to indicate the magnitude of DQ problems in the available credit risk databases. The results in Fig. 6 depict the observed magnitude of poor DQ. More than 10% of the data in credit risk management databases are estimated to be of poor quality. The majority of the institutions estimated that between 10-20% of the data in the databases is subject to errors. However, a large number of institutions do not know the magnitude of the problem. This is clearly shown by the not known answers given by 19% of the total respondents as depicted in the results in Fig. 6. This result indicates that most financial institutions are still unable to develop comprehensive measures and are unable to assess the magnitude of DQ problems. As a consequence, the impact of the existing poor DQ on the decision tasks is hard to assess as well.

In addition, the intrinsic difficulty of accurately measuring DQ might discourage any initiative to improve it. This is confirmed by the results shown in Fig. 4b which indicate that regulatory requirements (e.g. Basel I and II) are cited as the key reasons of any DQ enhancing project. The Basel Accord requires the calculation of detailed loss modeling factors to determine the capital requirement as explained earlier. Accurate quantitative modeling of PD, LGD, and EaD is not only required by this regulation but can become a competitive advantage leading to superior credit performance [2]. However, a competitive advantage is considered as less important to initiate DQ enhancing activities as indicated in the results in Fig. 4b. Because of these regulatory compliance requirements, financial institutions are organizing DQ teams to improve DQ and cross-functional efforts to improve the comparability and applicability of data sources across different business units. However, such efforts are reported to be not matured enough yet.

In the first part of the survey, respondents were also asked whether the DQ ICT architecture implemented in the organization (if any) is similar to the blueprint architecture suggested by the pilot survey. Most respondents indicated the presence of a DQ ICT architecture which was however dissimilar to the suggested one; a minority even did not have any DQ specific ICT architecture in place. Further inquiry revealed the existence of at least three important differences to the DQ ICT architecture proposed as ideal by the respondents of the pilot survey. The first element is the lack of a staging area to check the DQ prior to transfer to the data warehouse. Instead the data are most often directly transferred from the operational data stores to the data warehouses where DQ is then subsequently checked. The second main difference is there is the difficulty of tracing back the DQ errors and/or problems to the source or original operational data stores for correction. Hence, it was indicated that identified DQ problems are typically corrected in the data warehouses itself instead of in the operational data stores as in the ICT architecture suggested in the pilot survey. As a result, the data in the operational stores and data warehouses are likely to be different and thus inconsistent. Finally, the respondents indicated the absence of an independent business analyst to study the DQ problems in the implemented ICT architecture unlike in the suggested ICT architecture in the pilot survey.

Generally, the above explained key findings show that although poor DQ appears to be the norm, rather than the exception, the issue of DQ is currently largely ignored by financial institutions.

(a) *Major data quality issues in financial institutions*



(b) *Major Data quality initiative motivations in financial institutions*

Figure 4: *The major DQ problems and reasons for improvement actions*

4.2. Basic statistic analysis

In the second part of the survey, the respondents were asked to rate the importance of each of the DQ dimensions given in Table 2 for the credit risk assessment task. The overall results are presented in Table 3. All seventeen DQ dimensions in Table 2 are attributed a score higher than 7/10, indicating the importance of each dimension for credit risk assessment. The results in Table 3 are further analyzed by first performing a Friedman test, which detects if there are statistically significant differences between the scores of all DQ dimensions. The null hypothesis is strongly rejected (p value < 0.001) indicating significant differences exist in the results of the survey. Thus, we proceed with a Bonferroni-Dunn test. The results of the Bonferroni-Dunn test are depicted in Fig. 7. The X-axis in this figure corresponds to the average rank for each of the DQ dimensions. The DQ dimensions are represented by a horizontal line; the more this line is situated to the left, the higher the scores on that DQ dimension. The left end of this line depicts the average ranking while the length of the line corresponds to the critical distance. If the difference in average ranking between a DQ dimension and the 'best' DQ dimension is more than this critical distance, the difference is significant at the 99% confidence level. The 'best' DQ dimension is a DQ dimension which has the lowest average ranking. The dotted, dashed and full vertical lines in the figure indicate the critical difference at respectively the 90%, 95% and 99% confidence level. The scores on a DQ dimension are significantly lower than those of the 'best' dimension if it is located at the right hand side of the vertical line.

Accuracy clearly was attributed the highest score as it is the most left-positioned DQ dimension as shown in the results of the Bonferroni-Dunn test in Fig. 7 and consequently is confirmed to be the most important DQ dimension for the credit risk assessment task. Since accuracy is found to be the best scoring dimension, it is used to compare the average scores of each of the other sixteen DQ dimensions. The scores for security, relevancy, actionability, accessibility, objectivity timeliness, value-added and representational consistency are found to be not significantly different at the 99% confidence level. Based on these results, we can conclude that accuracy, security, relevancy, actionability, accessibility, objectivity, timeliness, value-added and representational consistency are the most important DQ dimensions for the credit risk assessment task. On the other hand, the completeness, interpretability, reputability, traceability, easy understandability, appropriate-amount, alignment and concise representation DQ dimensions are found to be significantly less important (see Fig. 7).

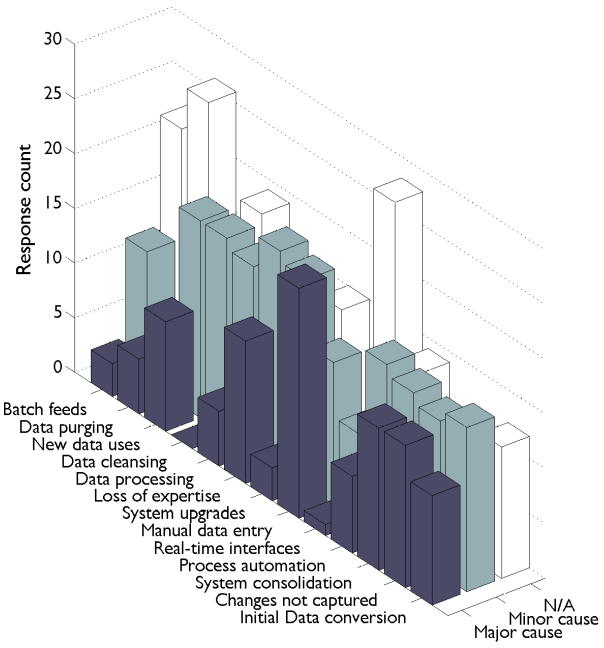


Figure 5: *Different causes of DQ problems in financial institutions*

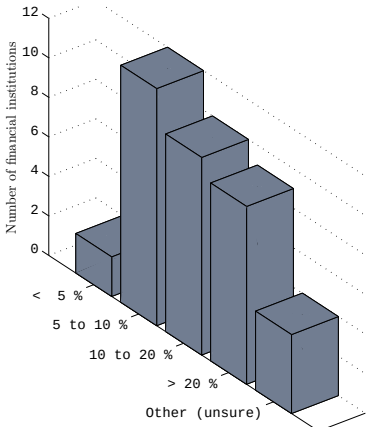


Figure 6: *Magnitude of poor DQ problems measured in financial institutions*

DQ Dimension	Mean	SD	95% C.I.
Accuracy(AC)	9.08	1.54	8.69–9.46
Actionable(ACT)	8.53	1.63	8.12–8.94
Relevancy(REL)	8.52	1.53	8.13–8.9
Security(SEC)	8.47	2.08	7.95–8.99
Accessibility(ACC)	8.41	1.61	8.00–8.81
Timeliness(TIM)	8.28	1.79	7.83–8.73
Value-added(VAD)	8.27	1.94	7.78–8.75
Objectivity(OBJ)	8.19	2.20	7.64–8.74
Representational-consistent(RC)	8.13	2.22	7.57–8.68
Completeness(COM)	8.02	2.31	7.44–8.59
Reputability(REP)	7.89	1.88	7.42–8.36
Interpretability(INT)	7.86	2.05	7.35–8.37
Appropriate-amount(APM)	7.84	1.86	7.38–8.31
Easily-understandable(EU)	7.81	1.93	7.33–8.30
Alignment(AL)	7.75	2.05	7.24–8.26
Traceability(TRA)	7.73	2.23	7.18–8.29
Concisely-Represented(CR)	7.36	2.21	6.81–7.91

Table 3: *Basic statistical description of DQ dimensions (Mean, standard deviation (SD) and confidence interval (C.I.)*

It has been shown that DQ problems increase when there is a large amount of data to be collected and managed [24]. This is typically the case for large institutions. Therefore, we conducted a Wilcoxon-ranked sum test of which the results are shown in Fig. 8 to check whether there is a difference in the importance of DQ dimensions between large and SMEs financial institutions. Financial institutions with total assets of more than and less than 100 billion euros are classified as large and SMEs respectively. While, there are no statistically significant differences among the median scores of DQ dimensions for large and SMEs institutions at the 5% significance level in (Fig. 8), we can see that the medians of most of the DQ dimensions for large institutions are greater than that of SMEs. This inconclusive result may indicate that the importance of each DQ dimensions increases as the amount and complexity of data increase.

The Friedman test is also used to test if there are significant differences among the aggregated scores of the four DQ classes defined in the framework of Wang and Strong [36]. The results depicted in Table 4 indicate that there is a statistically significant difference among the representation and the other three DQ classes. This indicates the fact that most of the DQ dimensions under the representation DQ class are significantly less important for credit risk assessment task as shown above.

Data quality Classes	Median	Standard deviation	Intrinsic	Contextual	Represen.	Access
Intrinsic	9	1.94				
Contextual	9	1.86				
Represen.	8	2.09				
Access	9	2.01				

Table 4: *The results of Wilcoxon ranked sum test show statistically significant difference among some of the DQ Classes(dark grey and light grey cells indicate significant and non significant difference at α = 5% respectively)*

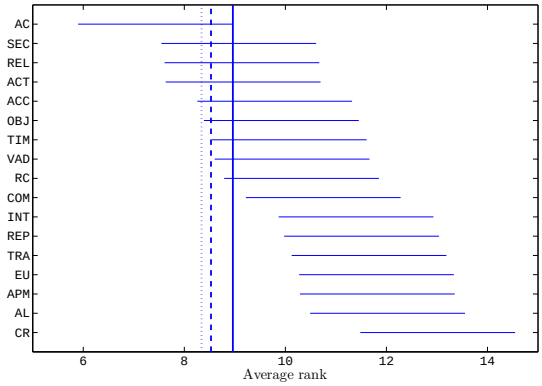


Figure 7: *Bonferroni-Dunn plot of the DQ dimensions scores assessed by the credit risk managers in financial institutions*

	AC	OBJ	REP	COM	APM	VAD	REL	TIM	ACT	INT	EU	RC	CR	AL	ACC	SEC	TRA
AC																	
OBJ																	
REP																	
COM																	
APM																	
VAD																	
REL																	
TIM																	
ACT																	
INT																	
EU																	
RC																	
CR																	
AL																	
ACC																	
SEC																	
TRA																	

p-value ≥ 0.1 — *White*
0.05 ≤ p-value < 0.1 — *Light grey*
0.01 ≤ p-value < 0.05 — *Dark grey*
p-value < 0.01 — *Black*

Table 5: *The results of Spearman's rank correlation test which show the significance of correlation between DQ dimensions*

The correlation between DQ dimensions was also investigated using the Spearman's rank correlation, ρ. This is a non-parametric correlation measure which investigates the monotonic relationship between any two DQ dimensions. ρ is defined as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

with n the sample size and d_i the difference between the ordinal ranks assigned to each of the observations. The significance of the Spearman's rank correlation measure is given in Table 5. These results show that most of the DQ dimensions are correlated with each other. The black, dark grey and light grey cells in the results of Spearman test depicted in Table 5 show the significance of the correlation among the DQ dimensions at 99%, 95% and 90% confidence level respectively. The white cells indicate that there is no correlation between the DQ dimensions.

As the results in Table 5 indicate, the majority of the DQ dimensions are positively correlated to each other. Accuracy is correlated with the majority of other DQ dimensions, clearly illustrating the business analyst's tendency of equalizing it with the total DQ requirements. In fact, the problem of inaccuracy can be related to many of the DQ dimensions. For example, a null value for the age of a customer can be both associated to completeness and accuracy DQ dimensions. Accuracy can also relate to the representational consistency DQ dimension. For example, a birthdate value of a person represented in DDMMYY and MMDDYY format can indicate both inaccuracy and inconsistency problems.

The strong positive correlation observed in the results of Table 5 are also supported by the literature. Lee et al. [18] also found high correlation between a number of DQ dimensions. They reported a significant correlation at the 95% confidence level between the accessibility DQ dimension and the appropriate amount, believability, completeness, concise representation, consistent representation, free-of-error, interpretability, relevance, reputation, security, timeliness and understandability DQ dimensions [18]. Hence, it can be concluded that improvement action on one DQ dimension has a positive effect on the others DQ dimensions.

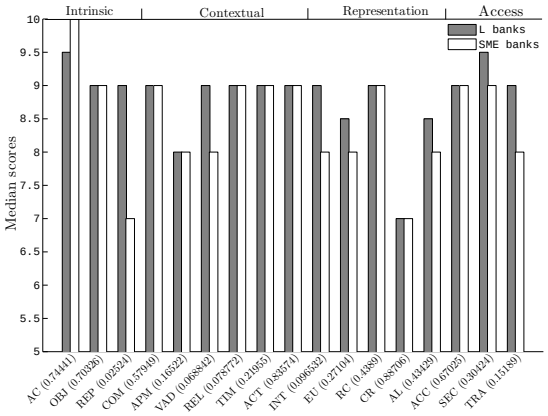


Figure 8: *The results of Wilcoxon ranked sum test, comparing the medians of DQ dimensions for large and SMEs financial institutions as assessed by the credit risk managers with their p-values between brackets*

5. Conclusion and Future Research

This paper explored the important DQ dimensions in the context of the credit risk assessment task and identified different DQ challenges and their possible causes. We started with a literature review of the different DQ dimensions, especially focussing on the framework of Wang and Strong [36]. Based on the results of the pilot survey, this framework was extended with three additional DQ dimensions (i.e. "alignment", "actionability" and "traceability"), totalling seventeen DQ dimensions. The importance of this extended framework has been assessed by credit risk managers. These decision makers rated the DQ dimensions on a scale from 0-10. The results were analyzed using a Friedman test which indicated a significant difference among the scores of the DQ dimensions. The results of the post-hoc Bonferroni-Dunn test confirmed that accuracy is the most important DQ dimension for the credit risk assessment. Also, security, relevancy, actionability, accessibility, objectivity, timeliness, value-added and representational consistency are found to be important DQ dimensions for this task. Moreover, it was found that the majority of DQ dimensions are correlated, implying that DQ, although intrinsically a multidimensional concept, is often perceived from a single perspective.

Finally, the paper identified different DQ challenges and their causes in financial institutions. The results indicated that inconsistency and diversity of data sources are among the most recurring challenges. Likewise, manual data entry processes are found to cause the majority of the DQ problems. Although DQ problems are endangering the effectiveness of the task, only little DQ enhancement activities are currently in place. Moreover, these activities are mostly instigated by regulatory authorities, rather than by internal considerations. Surprisingly, creating a competitive advantage was not found to be an important stimulus in any DQ improving activity.

Although the implementation of a TDQM program involves *DQ definition*, *measurement*, *analysis* and *improvement* phases, this paper only focused on the DQ definition phase. In the definition phase, the identification of various DQ dimensions relevant to credit risk assessment was considered. This helps to invest resources for improving the appropriate DQ dimensions. However, the three other phases (*measurement*, *analysis* and *improvement*) for the identified DQ dimensions are left as a topic for future research.

It is also confirmed in this paper that the majority of financial institutions are unaware of the magnitude of their DQ problems. This implies that they are still unable to develop comprehensive measures to these DQ problems. This is a clear indication of the need for comprehensive DQ measuring metrics.

Finally, although the pilot survey identified the blueprint DQ ICT architecture which should be implemented by financial institutions in order to enhance the DQ of data stores, the effectiveness and viability of this architecture needs further investigation.

Acknowledgment

This research was supported by the odysseus program (Flemish Government, FWO) under grant G.0915.09.

[1] B. Baesens, C. Mues, D. Martens, and J. Vanthienen. 50 years of data mining and or: upcoming trends and challenges. *Journal of the Operational Research Society*, 60:16–23, 2009.

[2] Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards. *Technical report, Bank of international settlements*, 2006.

[3] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*, pages 20–50. Springer, New York, 2006.

[4] C. Cappiello, P. Giciaro, and B. Pernici. Hiqm: A methodology for information quality monitoring, measurement, and improvement. *ER Workshops*, LNCS 4231:339–351, 2006.

[5] C.C. Chen and Y.D. Tseng. Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, in press, 2010.

[6] I.N. Chengalur-smith, D.P. Ballou, and H.L. Pazer. The impact of data quality information on decision making: An exploratory analysis. *IEEE Transactions of Knowledge and Data Engineering*, 11(6), 1999.

[7] K. Dejaeger, B. Hamers, J. Poelmans, and B. Baesens. A novel approach to the evaluation and improvement of data quality in the financial sector. In *Proceedings of the 15th International Conference on Information Quality*, Little Rock, USA, 2010.

[8] W.H. Delone and E.R. McLean. Information systems success: The quest for the dependant variables. *Information Systems Research*, 3(1):60–95, 1992.

[9] M.J. Eppler and D.Wittig. Conceptualizing information quality: A review of information quality frameworks from the last ten years. In *Proceedings of the 2000 Conference on Information Quality*, 2000.

[10] C.W. Fisher and D.P. Ballou. The impact of experience and time on use of data quality information in decision making. *Information Systems Research*, 14(2):170–188, 2003.

[11] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92, 1940.

[12] T.V. Gestel and B. Baesens. *Credit Risk Management*. Oxford University Press Inc. New York, 2009.

[13] M.B. Gordy. A comparative anatomy of credit risk models. *Journal of Banking and Finance*, 24:119–149, 2000.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. Springer, 2001.

[15] M. Jarke and Y. Vassiliou. Data warehouse quality: A review of the DWQ project. In *Proceedings of the Confrenece on Information Quality*, pages 299–313, Cambridge, MA, 1997.

[16] B.K. Kahun, D.M. Strong, and R.Y. Wang. Information quality benchmarks: Product and service performance. *Communications of the ACM*, 45(4), 2002.

[17] Y.W. Lee, L.L. Pioino, J.D. Funk, and R.Y. Wang. *Journey to Data Quality*, pages 67–108. The MIT Press, London, 2006.

[18] Y.W. Lee, D.M. Strong, B.K. Kahun, and R.Y. Wang. A methodology for information quality assessment. *Information and Management*, 40:133–146, 2002.

[19] S. Madnick and H. Zhu. Improving data quality through effective use of data semantics. *Data & Knowledge Engineering*, 59:460–475, 2006.

[20] A. Maydanchik. *Data Quality Assesment*, pages 5–30. Technics publications, 2007.

[21] C. Moraga, M.A. Moraga, C. Calero, and A. Caro. Towards the discovery of data quality attributes for web portals. *ICWE*, LNCS 5648:251–259, 2009.

[22] A. Paul P. Cykana and M. Stern. Dod guidelines on data quality management. In *Proceedings of the Conference on Information Quality*, pages 154–171, Cambridge, MA, 1996.

[23] F. Panse and N. Ritter. Completeness in databases with maybe tuples. *ER workshops*, pages 202–211, 2009.

[24] A. Parssian and V.S. Jacob. Assessing data quality for information products: Impact of selection, projection, and cartesian product. *Management Science*, 50(7):967–982, 2004.

[25] L.L. Pipino, Y.W. Lee, and R.Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4), 2002.

[26] S. Raghunathan. Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis. *Decision Support Systems*, 26:275–286, 1999.

[27] E. Rahm and H.H. Do. Data cleaning: problems and current approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2000.

[28] T.C. Redman. The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 1998.

[29] G. Shankaranarayanan and Y. Cai. Supporting data quality management in decision-making. *Decision Support Systems*, (42):302–317, 2006.

[30] G. Shankaranarayanan, M. Ziad, and R.Y. Wang. Managing data quality in dynamic decision environments: An information product approach. *Journal of Database Management*, 14(4):14–32, 2003.

[31] D.M. Strong, Y.W. Lee, and R.Y. Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[32] G.K. Tayi and D.P. Ballou. Examining data quality. *Communications of the ACM*, 41(2), 1998.

[33] Y. Wand and R.Y. Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 1996.

[34] R.Y. Wang. A product perspective on data quality management. *Communications of the ACM*, 2, 1998.

[35] R.Y. Wang, V.C. Storey, and C.P. Firth. A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, 7(4), 1995.

[36] R.Y. Wang and D.M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 1996.

[37] J.E. Ware and B. Gandek. Methods for testing data quality, scaling assumptions, and reliability. *Journal of Clinical Epidemiology*, 51(11):945–952, 1998.

[38] S. Watts, G. Shankaranarayanan, and A. Even. Data quality assessment in context: A cognitive perspective. *Decision Support Systems*, 48:202–211, 2009.

[39] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[40] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wode web. In *Proceedings of the 3rd Annual Internation ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 288–295, 200.