7-15-2012

Practical Significance Of Key Data Quality Research Areas

Vimukthi Jayewardene
School of Information Technology and Electrical Engineering, The University of Queensland, Australia, w.jayawardene@uq.edu.au

Shazia Sadiq
School of Information Technology and Electrical Engineering, The University of Queensland, Australia, shazia@itee.uq.edu.au

Marta Indulska
UQ Business School, The University of Queensland, Australia, m.indulska@business.uq.edu.au

PRACTICAL SIGNIFICANCE OF KEY DATA QUALITY RESEARCH AREAS

Vimukthi Jayewardene, School of Information Technology and Electrical Engineering, The University of Queensland, Australia, w.jayawardene@uq.edu.au

Shazia Sadiq, School of Information Technology and Electrical Engineering, The University of Queensland, Australia, shazia@itee.uq.edu.au

Marta Indulska, UQ Business School, The University of Queensland, Australia, m.indulska@business.uq.edu.au

Abstract

The body of knowledge on data and information quality is highly diversified, primarily due to the cross-disciplinary nature of data quality problems, coupled with a strong focus on fitness for use principle in developing data quality solutions. As a result, research and practice in data and information quality is characterized by methodological as well as topical diversity. Although research pluralism is highly warranted, there is evidence that substantial developments in the past have been isolationist. As a first step towards bridging gaps between various communities involved in data quality research and practice, we undertook a literature review of data quality research published in a range of Information System (IS) and Computer Science (CS) publication outlets and identified the key themes of research from last 20 years. In this paper, we utilize the above results to explore the impact of these themes within the data quality professional community. To that end, we developed an initial model of data quality factors (based on the identified key research themes), and conducted a survey of data quality practitioners to test the model. Our study found that the effective implementation of data quality assessment practices, data quality frameworks, and data constraints and rules, has a significant impact on overall data quality levels in organisations, whereas focus on other factors do not appear to significantly affect data quality. Results from this study can assist organizations in prioritising their data quality initiatives to focus on the factors that have the potential to contribute most significantly to overall data and information quality.

Key Words: Data Quality, Information Quality, Data Quality Factors, Literature Survey

1 INTRODUCTION

The nature of data quality management is constantly evolving as the organisational environment changes. There are a number of reasons behind this change. First, there are clear implications that relate to the sheer volume of data produced by organizations today. Second, recent years have seen an increase in the diversity of data. Such diversity refers to structured, unstructured, semi-structured data, and multi-media data such as video, maps, images, etc. Data also has an increasing number of sources. The use of various technologies, for example, sensor devices, medical instrumentation, RFID readers, increases the amount and diversity of data being collected. More subtle factors also exist - such as the lack of clear alignment between the intention of data creation and its subsequent usage. A prime example of such lack of alignment is the vast amount of data collected from social networks that can then be used, without assessment of quality, as a basis for marketing decisions. As these changes occur, traditional approaches and solutions to data management in general, and data quality control specifically, are challenged. There is an evident need to incorporate data quality considerations into the whole data cycle, encompassing managerial/governance as well as technical aspects.

Contributions to address such challenges, both from the research and the practitioner community, originate from three distinct communities, *viz*. Business Analysts, Solution Architects, and Database Experts (Sadiq *et al,* 2011). Business Analysts, who focus on organizational solutions. That is, the development of data quality objectives for the organization, as well as the development of strategies to establish roles, processes, policies, and standards required to manage and ensure the data quality objectives are met. Solution Architects, work on architectural solutions. That is, the technology landscape required to deploy developed data quality management processes, standards and policies. Database Experts and statisticians, who contribute to computational solutions. That is, effective and efficient IT tools, and computational techniques required to meet data quality objectives. Techniques in this regard can include record linkage, lineage and provenance, data uncertainty, semantic integrity constraints, as well as information trust and credibility.

The three communities have, collectively, explored data quality from a variety of perspectives over several decades, allowing the emergence of a number of core topics of data quality research. For example, (Lee *et al*. 2001) measure information quality in organisations based on the parameters relating to various dimensions of information quality, such as accessibility, believability, completeness, and so on. In our work, we focus on identifying the core data quality research themes, as indicated by two decades of data quality research, and exploring their practical impact. In other words, with this study we aim to identify which of the core research foci have a positive effect on data quality in organisations.

2 BACKGROUND

Several works have outlined the contributions on data and information quality research in the past. Owing to the cross-disciplinary needs of this area, identifying the central themes and topics and correspondingly the associated methodologies has been a challenge. Recent work by Madnick *et al*, (2009) has presented a framework that characterizes data quality research along the two dimensions of topics and methods thereby providing a means to classify various research works. Ge and Helfert (1996) have structured their review of the literature in categories relating to information quality assessment, information quality management and contextual information quality. Lima *et al*., (2006) classify the literature between theoretical (conceptual, applied, illustrative) and practical (qualitative, experimental, survey, simulation) aspects. Further, Neely and Cook (2008) present their classification as a cross-tabulation of Wang's framework (Wang *et al*., 2005) and Juran's original fitness for use factors (Juran, 1962). The above studies provide various angles through which the body of knowledge can be classified and thus provide an essential means of understanding the core topics of data quality. However, understanding the intellectual corpus of a discipline requires not only an understanding of

its core, but also its boundaries (Benbasat and Zmud, 2003). As the realm of data quality has grown, so has the scope of its reference disciplines. With these factors in mind, we focused our study on understanding the interconnections and synergies across the various communities that contribute to data quality, rather than identification of its central themes. We argue that addressing the current challenges in data quality warrants such an understanding so synergies would be better exploited and holistic solutions may be developed.

3 APPROACH

The study incorporates two separate components, *viz*. literature analysis to identify core research themes, and practitioner survey to capture practitioner opinion on the importance of the themes and their level of implementation within the practitioner's organisation.

3.1 Identification of Key Research Themes

The literature study follows a conceptual analysis approach (Smith and Humphreys, 2006) in which material is examined for the presence, and frequency of concepts. These concepts can be words or phrases and may be implied or explicit. To ensure broad coverage of data quality research, we selected well-regarded Information Systems and Computer Science academic publication outlets. The selection is based on journal and conference rankings (See www.aisnet.org and www.core.edu.au) that are now common in many disciplines (Fisher *et al*., 2008), as well as our perception of these outlets. We acknowledge that this is an area of much debate and may vary between researchers. However, we have attempted to minimize any bearing on the outcome through the selection by an expanded scope and as far as possible identifying a well-balanced set of publications for the analysis. We further broaden our perspective through the consideration of both conference and journal publications, to provide a different perspective to the relatively common journal-only literature and citation studies (Chen *et al*, 2007)

Table 1 details the list of considered Information Systems and Computer Science publication outlets, and the respective volume of papers, that has been considered in this study. In particular, we have focused on almost the last two decades of conference and journal publications (1990-2009).

	Acronyms	Totals
CS Conferences	BPM, CAiSE (Workshops), CIKM, DASFAA, ECOOP, EDBT,PODS, SIGIR, SIGMOD, VLDB, WIDM, WISE	7535
IS Conferences	ACIS, AMCIS, CAiSE, ECIS, ER, HICSS, ICIQ, ICIS, IFIP, IRMA, IS Foundations, PACIS	13256
CS Journals	TODS, TOIS, CACM, DKE, DSS, ISJ (Elsevier), JDM, TKDE, VLDB Journal	8417
IS Journals	BPM, CAIS, EJIS, Information and Management, ISF, ISJ (Black-well), ISJ (Sarasota), JAIS, JISR, MISQ, MISQ Executive	2493

Table 1. Considered Publication Outlets (Due to space limitation, widely accepted abbreviations have been used)

Our data set consists of 31,701 articles. Given the large volume of papers considered, we developed a consistent and reproducible full text search strategy prior to commencing analysis (Sadiq *et al*, 2011). In summary, each article was inspected via full text search tools for generic keywords (such as data quality, quality of data, information quality etc.), scrutinized for relevance (e.g. keywords only appeared in bibliographic reference), and then utilized to systematically build the taxonomy. The above task produced 764 papers.

It was evident that the data set may also contain articles in which the chosen generic keywords may not necessarily explicitly appear, but the articles could still be implicitly related to the area and contain valuable outcomes. For example, papers within the database/computer science community that focus on record linkage may not contain any of the aforementioned generic keywords but are still relevant to data quality research. Accordingly, as a next step, we identified a set of 'second level' keywords to further review the literature. To obtain an objective and relevant list, two researchers independently reviewed a sample (5%) of the initial set of articles to obtain further relevant concepts/keywords. The researchers identified the high level main theme(s) of the papers and associated these with terms and/or phrases that are representative of the theme e.g. terms such as entity resolution, record linkage, data profiling, provenance and lineage etc. Through this resource intensive activity, a large number of second level keywords were identified. The results of the two independent researchers were then compared, followed by a discussion to resolve any keyword conflicts. The agreed set of keywords were further reduced as some of them did not return search results that were meaningful for data quality research.

A review of the second level keywords identified synonyms. For example, record linkage had several related techniques such as approximate join, similarity join, fuzzy matching etc. Thus our identification of the second level keywords resulted in the development of keyword taxonomy.

Finally, the identified keywords were also compared with a number of existing studies that have contributed to developing concept maps and various taxonomies for data quality, see e.g. Lima *et al* (2006),Ge and Helfert (1996), Madnick *et al* (2009). A number of augmentations were made to the list, including some further categories of the second (and sometimes further) level keywords in order to ensure wider and more complete coverage. Accordingly, these new keywords were then used to search the data set again. The same strategy was used to prune the returned results as for the general keywords. After this second phase of analysis, a total of 1364 relevant publications were identified. Where there was a large group of publications (>50 papers) within a given keyword, an attempt was made to find sub keywords if possible eg. edit distance, q-gram etc. for approximate matching. Finally 54 keywords were included in the taxonomy; further details of the resultant taxonomy can be found in Sadiq *et al* (2011).

Based on this taxonomy our next step was to identify a set of key themes of data quality by grouping the keywords in the taxonomy. The grouping was done using expert judgement on data quality. First the hierarchical nature of the taxonomy provided a natural grouping. Second, although the taxonomy construction was driven by number of publications (that is >50 papers in a particular topic resulted in further specialization of keywords), the grouping was driven by thematic similarity. Hence topics like Linkage and Integration were grouped together. It is important to further note that although the taxonomy represents a much larger diversity in the research concentration areas, the intention for the above grouping was to develop a more practitioner oriented themes which can be validated through an industry based survey while ensuring as broad coverage of the taxonomy as possible.

The groups were constructed by three experienced researchers, discussed and reviewed through analysis of the paper samples within the topics, and finally consolidated into seven main themes. These themes are briefly described below and are referred to as "Data quality factors" during the rest of the study.

3.2 Practitioner Survey

As the next step of the study, a survey was constructed based on the above data quality factors with the objective to understand:

(1) What is the significance of these factors from industry practitioners' point of view when designing data quality strategies in their respective organizations?

(2) How successfully these factors are implemented in their organizations through their data quality initiatives?

The survey was structured into two sections – *viz*. demographics and data quality related questions. The first section on demographics included questions relating to the individual's role in the organization, his/her education and experience with regards to data quality, number of data quality projects handled by the individual, the industry sector which they operate in and the size of the organization in terms of number of employees. The second section was focused on the identified data quality factors. For each factor two questions were posed. The first was designed to elicit an evaluation of the importance of each of the identified data quality factors within the respective organizations. A second question aimed to uncover how well these factors have been implemented (practically used) in their organizational context and also the organization's satisfaction with their level of data quality. Questions relating to the importance of the data quality factors and the effectiveness of implementation of these factors used a 5-point Likert scale. The survey instrument was also pilot tested with 6 data quality practitioners and researchers prior to the launch of the survey.

The target audience of the survey was primarily data quality professionals. The participants were targeted based on their job roles and active participation in data quality related online forums, industry conferences, and professional bodies. The survey was hosted online on Survey Monkey (www.surveymonkey.com/s/teaching-and-research-data-quality). However, responses were elicited through both print and online means. Print versions of the survey were distributed at one local data quality conference to over 100 delegates (27 responses were received). Email invitations to the online survey were also sent to a targeted mailing list of 110 experts and practitioners and an additional 25 responses were collected.

The questionnaire itself provided a definition for each data quality factor to ensure the respondents had a clear understanding of practices included in each factor. A verbal explanation was also provided for the participants recruited through the above-mentioned conference. The offline and online approach resulted in 52 usable responses, which represented a 23% response rate. Last, the survey announcement was included in a regular newsletter of the International Association of Information and Data Quality (iaidq.org). This last inclusion resulted in a further 8 responses, providing a total of 60 usable responses (we cannot report on the relevant response rate from the newsletter announcement as we have no indication of the active readership of the newsletter).

4 KEY DATA QUALITY RESEARCH THEMES

Based on the literature analysis described in section 2, we identified the following core research themes:

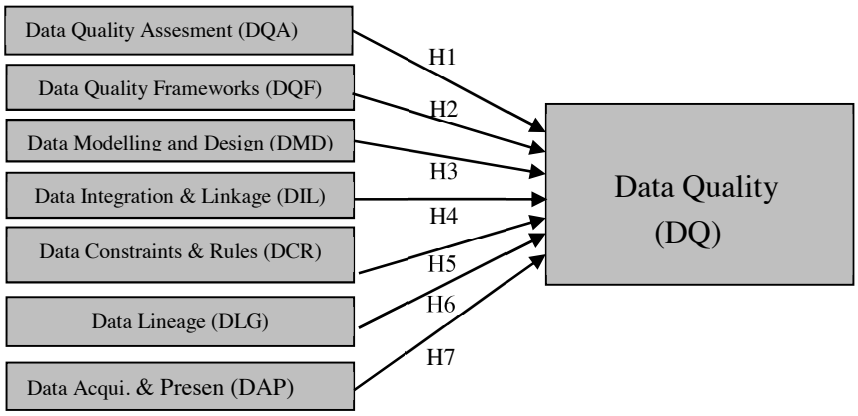
1. **Data Quality Assessment**: Refers to the process of investigating exposing and measuring the data related problems in the organization with the aim of planning and implementing data quality improvement strategies. This includes activities like statistical profiling, error detection, metrics, and methods for cost estimations.
2. **Data Quality Frameworks**: Refers to establishing an organizational level system for managing data quality with clearly defined objectives roles and responsibilities which is compatible with the overall corporate strategy of the organization. This includes data governance, benchmarking, best practices, and introduction of quality standards.

3. **Data Modelling and Design**: Refers to the effective identification of the data requirements and/or conceptualization of current data sets in the organization with the aim of maintaining a holistic and consistent view and understanding of the data across the organization. This includes deliberations on schema quality, maintaining documented meta-data, as well as approaches towards managing legacy systems.
4. **Data Integration and Linkage**: Refers to the technological aspects which ensure the data integrity of organization subjected to various forms of external and/or legacy data. This includes schema matching, duplicate detection/entity resolution, effective use of master data, managing different formats, as well as ETL/Data Warehousing.
5. **Data Constraints and Rules**: Refers to managing of techno-functional activities which ensure the alignment between the business and IT landscape of the organization. This includes data conformance to business rules, data standards, key/id management and various forms of semantic constraints.
6. **Data Lineage**: Refers to the process which ensures the management of the data lifecycle starting from its creation to disposal/archival. This includes provenance, data tracking, source attribution, ownership etc.
7. **Data Acquisition and Presentation**: Refers to the process of establishing effective and efficient mechanisms to acquire and present data in the technical layer. This includes design of suitable data interfaces, data entry controls, data collection/upload e.g sensor & RFID data, multimedia data etc.

Whereas previous studies have also presented various forms of classifications that could be used for such a study (Lima *et al*., 2006), (Ge and Helfert, 1996), we stipulate that the factors above are indicative of the broad and diverse nature of data and information quality research as they are extracted from a body of knowledge that spans disciplinary and methodological boundaries (Sadiq *et al*, 2011).

We posit that the above themes, when implemented through the various underlying means, have a positive impact on data quality in organisations. We are not interested in the finer detail of the implementation at this stage. For example, Data Quality Assessment may be addressed in organisations through the use of error detection, or cost estimations, or, more preferably a combination of approaches, however we do not aim to determine which implementations are of most benefit to organisations. We aim to study these finer details in our future work for themes that are shown in this study to be significant.

Based on the above themes, we develop an initial model of data quality factors that impact positively on the overall quality of data in an organisation. While the overall data quality is challenging to measure, we focus on the perceptional measure of overall level of data quality (measured on a 5-point Likert scale).



.

Figure 1. Initial model of research-based data quality factors

As shown in Figure 1, we hypothesize the existence of seven relationships based on the interaction of each factor with the overall data quality in organisations.

H1. Implementation of data quality assessment approaches is associated with improved data quality in organisations.
H2. Use of data quality frameworks is associated with higher data improved in organisations.
H3. Conduct of data modelling and design activities,is associated with improved data quality in organisations.
H4. Use of approaches for data linkage/integration is associated with improved data quality in organisations.
H5. Implementation of practices related to ensuring data constraints/rules are conformed to, is associated with improved data quality in organisations.
H6. Identification of data lineage is associated with improved data quality in organisations.
H7. Use ofapproaches for effective acquisition of data is associated with improved data quality in organisations.

Using correlation and multiple regression analysis (Cohen *et al*, 2003) we test for the support of above hypotheses in our collected data. All hypotheses are tested at the 95% confidence limit, unless otherwise stated. The results of the survey are discussed in the next sections, followed by a discussion of implications in a subsequent section.

5 SURVEY RESULTS

In the analysis we considered the responses of the 60 data quality professionals who are currently working in either the government or the private sector. The respondents are employed in various capacities, including directors, managers and executives. Of the 60 respondents, 32%work for large organizations (over five thousand employees); 27% work for medium sized organizations (between1000 and 5000 employees); with the remaining 41% being from organizations with less than 1000 employees. The survey filtering criteria ensured that each participant had conducted at least one data quality project; however the average number of completed data quality projects across the 60 respondents is 13 projects/person. We consider this average number of completed projects to be significant and a good indicator that the respondents have sufficient practical exposure in the domain of data quality to provide valid responses to the survey.

An interesting finding from the demographic questions of the survey is that the majority of the data quality professionals did not receive any formal training in data quality management. Indeed, over 60% indicated that they were self-taught, which was often combined with on the job training. Only a mere 3.5% of the respondents have official industry certification, and 35% have professional or university training that relates to data quality (see Figure 2). The finding has a serious implication with respect to the level of variability in data quality management approaches that stems from a lack of standardised or best-practice education.

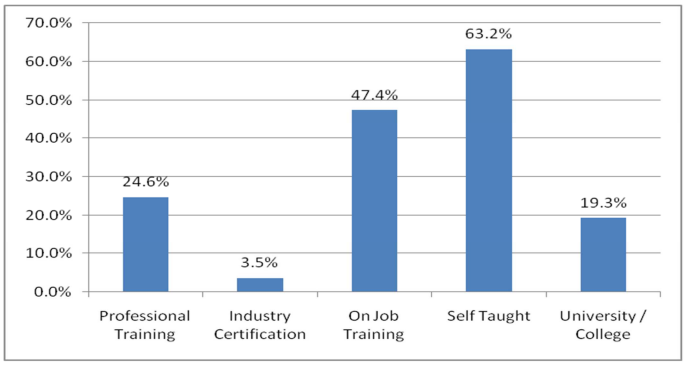


Figure 2. Level of data quality training.

Table 2 summarises the responses for each data quality factor in term of their general importance from practitioner point of view. As per the analysis, approximately 70% of the respondents have rated each concept as either "high" or "very high" with regards to its general importance in their organizational context. Further, over 80% of the respondents agree that all of the identified data quality factors have at least a medium level importance in their respective organizations. Hence, as per the given scores, we can conclude that the practitioners believe that all the identified data quality factors have potential to contribute to improved quality of data in their organizations.

Data Quality concept	Very Low	Low	Medium	High	Very high
Data Quality Assessment (DQA)	17.4%	2.2%	8.6%	19.6%	52.2%
Data Quality Frameworks (DQF)	6.5%	8.7%	10.9%	19.6%	54.3%
Data Modelling and Design (DMD)	4.4%	8.9%	20.0%	15.6%	51.1%
Data Integration and Linkage (DIL)	4.4%	0.0%	26.7%	24.4%	44.4%
Data Constraints and Rules (DCR)	4.4%	2.2%	15.6%	22.2%	55.6%
Data Lineage (DLG)	4.7%	9.3%	18.6%	30.2%	37.2%
Data Acquisition and Presentation (DAP)	6.10%	2.00%	14.20%	20.40%	57.30%

Table 2. General importance of the concepts

Table 3 summarises the responses for each data quality concept in terms of how well it has been implemented in the respective organizations. On average, approximately 20%-30% respondents report that the organisation has implemented each factor "well" or "very well". Further around 50% of the respondents agree that their organisations have at least implemented these factors at a medium level.

Data Quality concept	Very Poor	Low	Medium	Well	Very Well
Data Quality Assessment (DQA)	31.3%	19.5%	20.9%	17.4%	10.9%

Data Quality Frameworks (DQF)	26.1%	26.1%	23.9%	15.2%	8.7%
Data Modelling and Design (DMD)	11.1%	37.8%	28.9%	13.3%	8.9%
Data Integration and Linkage (DIL)	15.9%	38.6%	25.0%	9.1%	11.4%
Data Constraints and Rules (DCR)	20.0%	15.6%	26.7%	31.1%	6.7%
Data Lineage (DLG)	21.4%	26.2%	26.2%	11.9%	14.3%
Data Acquisition and Presentation (DAP)	17.00%	16.00%	34.60%	24.40%	8.00%

Table 3. How well the concepts have been implemented in respective organizations

.

Hence, as per the above scores, organizations, in general, are still struggling to implement the concepts relating to the core factors of data quality.

Our next concern is to identify the level of correlation between the overall success in data quality management with each of the above mentioned data quality factors individually. A high correlation would indicate a strong relationship/dependency, Low correlation, on the other hand, would indicate a weak relationship/dependency between the factors and overall data quality (Table 4).

DQ Concept	Correlation Coefficient (R)	Coefficient of Determination (R^2)
Data Quality Assessment (DQA)	0.795	0.632
Data Quality Frameworks (DQF)	0.809	0.656
Data Modelling and Design (DMD)	0.637	0.406
Data Integration and Linkage (DIL)	0.652	0.425
Data Constraints and Rules (DCR)	0.731	0.534
Data Lineage (DLG)	0.622	0.387
Data Acquisition and Presentation (DAP)	0.694	0.481

Table 4. Correlation between overall DQ success & DQ factors

As indicated by the results, the Data Quality Frameworks factor has the highest correlation (0.809) with overall DQ success with 65.6% of the variation of DQ success explained by Data Quality Frameworks. The Data Quality Assessment factor has the second highest correlation (0.759) with overall DQ success, explaining 63.2% of the variation. Data Constraints and Rules have the third highest correlation (0.731) with overall DQ success explaining 53.4% of the variation.

Data Acquisition and Presentation, Data Integration and Linkage, Data Modelling and Design, and Data Lineage are correlated to success of overall data quality with correlation coefficients of 0.694,

0.652,0.637 and 0.622 respectively. These factors explain the variation of success in overall data quality by 48.1%, 42.5%, 40.6% and 38.7%respectively.It should be noted that Data lineage has the least correlation with overall data quality success (with a coefficient of 0.622) explaining only 38.7% of the variation.

As per the above correlation analysis, it is clear that overall data quality is positively correlated to the implementation status of seven data quality factors. The next step is to verify the nature of this correlation (i.e linear vs. non-linear). For this purpose we examine scatter diagrams between each independent variable and the dependant variable. We omit the full set of these plots from the paper due to lack of space. The plots in general indicate, however, that the relationships between the factors of concern are linear and hence it is justifiable to use the multiple linear regression model for further analysis (Cohen *et al*. 2003). Accordingly, using Microsoft Excel & SPSS, a regression analysis was performed based on the survey responses to observe the goodness of fit to the multiple regression model. The confidence level was selected as 95%.The results of the multiple regression analysis are discussed next.

Before applying multiple regression, we calculated Variance Inflation Factors (VIF) for each of the independent variables (data quality factors) and found all the VIF values are well below 10 which means that there is no significant co-linearity between the independent variables, assuring that each factor does not load onto other factors in its measurement (Klienbaum *et al*., 1998).

Multiple R of 0.873 (Table 5) indicates that the positive correlation between the dependent variable (DQ) and the set of independent variables (DQA … DAP) is high. R square of 0.76 indicates that it is statistically correct to say that 76% of the variation in DQ is explained by the independent variables DQA.....DAP.

Regression Statistics	
Multiple R	0.873
R Square	0.763
Adjusted R Square	0.719
Standard Error	0.864

Table 5. Correlation statistics-Multiple regression

As per Table 6, significance F (the associated P-value) of $7.4506649 \times 10^{-10}$ (i.e. $p < 0.05$) indicates that we can accept the regression model at the 95% confidence level.

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	89.524	12.789	17.098	7.4506649E-10
Residual	37	27.675	0.747		
Total	44	117.2			

Table 6. Analysis of variance for multiple regressions

Based on the P values, (Table 7) out of the seven data quality factors, the statistically significant factors (where p is <=0.05) in this model are IV1,IV2 and IV5 , which implies that hypotheses H1 , H2 and H5 are supported at the 95% confidence level. Thus, we conclude that the effective

implementation of aspects relating to these three factors has a significant impact on the overall data quality of an organization.

	Coefficients	*Standard Error*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.86	0.33	0.01	0.17	1.54
DQ Assesment (DQA)	0.33	0.16	0.05	-0.005	0.68
DQ framework (DQF)	0.63	0.20	0.00	0.20	1.05
Data Modeling & Design (DMD)	-0.12	0.20	0.53	-0.53	0.27
Data Integration & Linckage (DIL)	-0.12	0.21	0.55	-0.55	0.30
Data Constraints & Rules (DCR)	0.43	0.18	0.02	0.05	0.80
Data Lineage (DLG)	-0.12	0.16	0.45	-0.45	0.21
Data Acquisition& Presentation (DAP)	0.07	0.19	0.70	-0.32	0.47

Table 7. Multiple regression statistics

This finding has several implications. First this gives the data quality practitioners on how to prioritise their data quality projects in such a way it delivers expected outcome (data quality) more quickly by giving more emphasize to the aspects relating to the above three factors i.e Data quality frameworks, Data quality assessments and Data constraints and rules. Secondly it paves the way for a long term data quality strategy in organizations by encourages senior management to invest further on data quality projects based on a results driven approach. To understand the key barriers for implementing these aspects we analysed open-ended responses given by the respondents with regards to each factor. After coding the text responses, we found that data quality assessment and data quality frameworks are seemingly immature within industry, with lack of knowledge, lack of suitable skills and competencies and lack of organizational support identified as the key reasons that have prevented effective deployment of data quality assessment initiatives and data quality frameworks.

The comments indicate that 83% organizations are making some effort towards doing data quality assessments, but have not reached expected levels due to the above mentioned limitations. It is also apparent that many organisations spend significant resources for improving specific data sets, rather than investing towards a consistent methodology for data quality assessment or addressing the root causes for poor quality data. In other words, many organisations still opt for expensive quick-fixes of problems instead of focusing on the underlying problems or ongoing monitoring.

Convincing senior management to invest in resources on data quality is a significant concern of many professionals (32% respondents indicated this concern explicitly while a vast majority mentioned it implicitly). Since data quality aspects and the benefits are generally not well established among business executives, there is less of a tendency to invest in long-term solutions to address data quality issues. Hence, organizations do not tend to commit resources to implementing data quality frameworks.

IT/Business alignment appears to be the major hurdle that affects data constraints and rules. Approximately 14% respondents explicitly stated that their IT teams and business teams are not aligned properly, while most respondents mentioned this issue implicitly. In particular, systems developed without a long term vision subsequently, though not surprisingly; encounter limitations in facilitating data constraints and rules. Inappropriate software/modelling tools and legacy systems are other mentioned hurdles that are of concern to effective management of data constraints and rules.

6 CONCLUDING DISCUSSION

In this study we explored, through survey-based research, practitioner views on seven data quality factors that were derived based on literature analysis. Seventy per cent of the participants indicated a “high” or “very high” importance of these factors in achieving good data quality in organizations.

However, only 20-30% indicated they were satisfied with the current implementation status of the seven factors in their respective organizations. This finding paved the way to explore the most significant factors that contribute directly to improved data quality in organizations. Our data analysis indicates that Data Quality Assessment, Data Quality Frameworks and Data Constraints and Rules are respectively the three factors that contribute most significantly towards the achievement of good data quality within organizations. Hence in designing the organizational data quality strategies, practitioners need to pay more attention towards implementing aspects related to the three factors. With this focus, positive impacts on data quality will be obtained sooner, thus providing management with evidence that will encourage further investment in data quality management practices.

In our future research we plan to investigate the moderating factors for our statistical model for improved data quality and develop a more comprehensive model towards achieving data and information quality with objective measurements. And also we hope to investigate into implementation strategies for the above three main factors, and explore through case studies the characteristics of successful implementations. The objective is to share knowledge where available on practical approaches utilized to successfully overcome the above mentioned hurdles.

References

Benbasat, I. And Zmud, R.W. (2003) The identity crisis within the IS discipline: Defining and communicating the discipline's core properties. *MIS Quarterly*, 27(2). pp.183-194Chen, C., Song, I.Y. and Zhu, W. (2007) Trends in conceptual modelling: Citation analysis of the ER conference papers (1979-2005). T*he 11th International Conference on the International Society for Scientometrics and Informatics*. pp 189-200.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences, 3rd Ed*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fisher, J., Shanks G. and Lamp J.A (2008) ranking list for information systems journals. *Australasian Journal of Information Systems*, 14(2). pp 114-125.

Ge. M., andHelfert, M. (1996) A Review of Information Quality Research. *The 12th International Conference on Information Quality, MIT, Cambridge, Massachusetts, USA*. pp 1-9.

Juran, J.M. *Quality control handbook*. (1962) McGraw-Hill Publishing Co

Klienbaum, D., Kupper, L., Muller, K., Nizam, A., (1998). *Applied Regression Analysis and Other Multivariable Tools*. Duxbury Press, Pacific Grove, CA, pp. 240– 249.

Lee, Y., Strong, D., Kahn, B., & Wang, R.(2001). AIMQ: a methodology for information quality assessment. Information Management, 2, 133-146.)

Lima, L.F.R., Macada, A.C.G. and Vargas L.M.(2006) Research into information Quality: a study of the state of the art in IQ and its consolidation. *11th International Conference on Information Quality,*MIT, Cambridge, Massachusetts, USA.

Madnick S.E., Wang R.Y, Lee Y.W., and Zhu H. (2009) Overview and Framework for Data and Information Quality Research. Journal of Data and Information Quality (JDIQ),

Neely, M.P. and Cook, J. A (2008) Framework for classification of the data and Information Quality literature and preliminary results (1996-2007). *AMCIS*.

Sadiq,S. , Yeganeh, N.Y. andIndulska, M. (2011) An Analysis of Cross-Disciplinary Collaborations in Data Quality Research. *European Conference on Information Systems (ECIS2011), Helsinki, Finland*

Smith, A.E. and Humphreys, M.S. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping .*Behavior Research Methods*, 38(2). pp:262-279, .

Wang, R.Y., Storey, V.C. and Firth, C.P. (2005) A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, 7(4). pp 623-640.