RESEARCH AND INDUSTRY SYNERGIES IN DATA QUALITY MANAGEMENT

(Research Paper)

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

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

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

Abstract: Research and practice in data and information quality is characterized by methodological as well as topical diversity. The cross-disciplinary nature of data quality problems as well as a strong focus on solutions based on the fitness for use principle has further diversified the body of knowledge on data and information quality. Although research pluralism is highly warranted, there is evidence that substantial developments in the past have been isolationist. As data quality increases in importance and complexity, there is a need to motivate exploitation of synergies across diverse research communities in order to form holistic solutions that span across organizational, architectural and computational aspects of data quality management. As a first step towards bridging gaps between the various communities, we undertook a literature review of data quality research published in a range of Information System (IS) and Computer Science (CS) publication outlets, and conducted a global survey of data quality management practitioners. In this paper, we present taxonomy of the main research topics contrasted against industry perceptions on the relative importance of those topics. Through the research-industry contrast, we hope to create a better understanding of research industry synergies as well as highlighting areas of high potential gaps and impact for the research community.

Key Words: Discipline Reflection, Research Methods, Literature Survey, Research Impact

BACKGROUND AND RATIONALE

The issue of data quality is as old as data itself. However it is now increasingly exposed at a much more strategic level increasing manifold the stakes for all involved including corporations, government agencies, and communities. Further, the proliferation of shared/public data as on the World Wide Web and growth of the web community has increased the risk of poor data quality usage for individuals as well. This is particularly alarming due to the diversity of the web community, where many are unaware of data sources and data credentials. The situation is further complicated by presence of data aggregations and assimilations e.g. through meta-search engines where source attribution and data provenance can be completely hidden from the data consumers.

One can also observe the changing nature of data quality management over the last decade or more. 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, further 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. Accordingly, a related factor exists that relates to difficulties in defining appropriate data quality metrics.

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. Currently, data quality contributions from research and industry appear to originate from three distinct communities: *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*, who 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.

For the research community to adequately respond to the current and changing landscape of data quality challenges, a unified framework for data quality research is needed. Such a framework should acknowledge the central role of data quality in future systems development initiatives and motivate the exploitation of synergies across diverse research communities. It is unclear if synergies across the contributing communities have been fully exploited. As such a deep and comprehensive analysis of data quality research contributions is warranted.

There have been previous studies that have contributed to the understanding of data quality research impetus by developing frameworks through which data quality research can be characterized, e.g. [12], analogized data quality processes with product manufacturing processes. Some key research aspects such as data quality standardization, metrics/measurements and policy management emerged from these earlier works. More recent studies have also provided valuable means of classification for data quality research. [4], for example, have structured their review of the literature as IQ Assessment, IQ Management and Contextual IQ. The study in [7] classifies the literature into theoretical (conceptual, applied, illustrative) and practical (qualitative, experimental, survey, simulation) aspects. Further, [9] present their classification as a cross-tabulation of framework in [12] the original fitness for use factors as given by [6].

Owing to the cross-disciplinary nature of the data quality research domain, identifying the central themes and topics, and, correspondingly, the associated methodologies, has been a challenge. In academia, several studies have addressed the issue of defining and analysing the scope of data quality research. Recent work by [8] 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. The research has identified four categories of data quality research, each having several sub categories, as follows:Data quality impacts: development of methods, designs and test mechanisms that maximize positive impacts of data quality in organisations while minimising negative impacts; Database related technical solutions: development of database technologies for assessing, improving, and managing data quality, including development of techniques for reasoning about data quality and for designing systems that result in data of high quality; Data quality in the context of computer science and IT: technologies and methods (except for the specific database-related techniques) to manage, ensure, and enhance data quality; and lastly Data quality in curation: selection, preservation, and management of digital information in ways that promote easy discovery and retrieval for both current and future uses of that information.Further they have identified fourteen high level research methodologies that researchers have used to investigate into data quality issues which provide an indication of the span of the studies so far.

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 [1]. As the realm of data quality has grown, so has the scope of its reference disciplines. These include information systems, management studies, databases, statistics, and computer science.

In order to provide an overarching coverage of data quality research contributions from various research communities (in particular information systems and computer science), we conducted an analysis of data quality research over the past twenty years [10]. In this study we considered a broad range of Information System (IS) and Computer Science (CS) publication (conference and journal) outlets so as to ensure adequate coverage of organizational, architectural and computational contributions. The main aims of the study were to understand the current landscape of data quality research, to create better awareness of (lack of) synergies between various research communities, and, subsequently, to direct attention towards holistic solutions that span across the organizational, architectural and computational aspects (thus requiring collaboration from the relevant research communities). In addition to providing insights into the major themes, venues, contributors and citations, the analysis also produced a large collection of hierarchically organized keywords. The basic objective for the keywords was to provide a means of searching the bibliography of over 1400 publications that resulted from the analysis. However the set of keywords also served as taxonomy for classifying the last two decades of research contributions.

In this paper, our aim is to relate the main themes of the taxonomy (i.e. characterization of research) to perceptions of relative importance of data/information quality aspects as seen by industry. In doing so, we hope to provide an exposition of research industry synergies and provide direct feedback for the research community on industry pain points and areas of high potential and impact that have not been adequately addressed by the research community.

To the best of our knowledge, this comparison has not been undertaken in previous studies. There have been some industry led initiatives that have attempted to identify key requirements or demands from industry in terms of data quality management [5], [14]. The most relevant and recent of which is a job analysis report published by the International Association for Information and Data Quality (iaidq.org). The report provides data that assists in understanding and establishing the roles of data quality professionals in industry. Additionally, the report also identifies the body of knowledge required by those professionals to provide information/data quality services across various roles of an organization [13].The results from this paper can provide substantial links into research for the identified body of knowledge, thereby presenting a means of exploitation of research-industry synergies.

In the subsequent sections, we present our research methodology, including a brief explanation of the approach taken to construct the taxonomy (via the literature analysis), as well as the approach taken to elicit industry responses against the taxonomy. We then present the key findings from the analysis of the industry responses and highlight the gaps, areas of interest and potential impact.

APPROACH

The study incorporates two separate components, viz. literature analysis and practitioner survey, to enable the contrast of core data quality research themes with main practitioner challenges.

The literature study follows a conceptual analysis approach [11] 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 [3]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 [2].

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).

	Includes	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 (Blackwell), ISJ (Sarasota), JAIS, JISR, MISQ, MISQ Executive	2493

Table 1. Considered Publication Outlets (*Due to space limitation, widely accepted abbreviations have been used, where full names are easily searchable via WWW)

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 [10]. 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 was then later reduced as several did not return search results that were meaningful for data quality research.

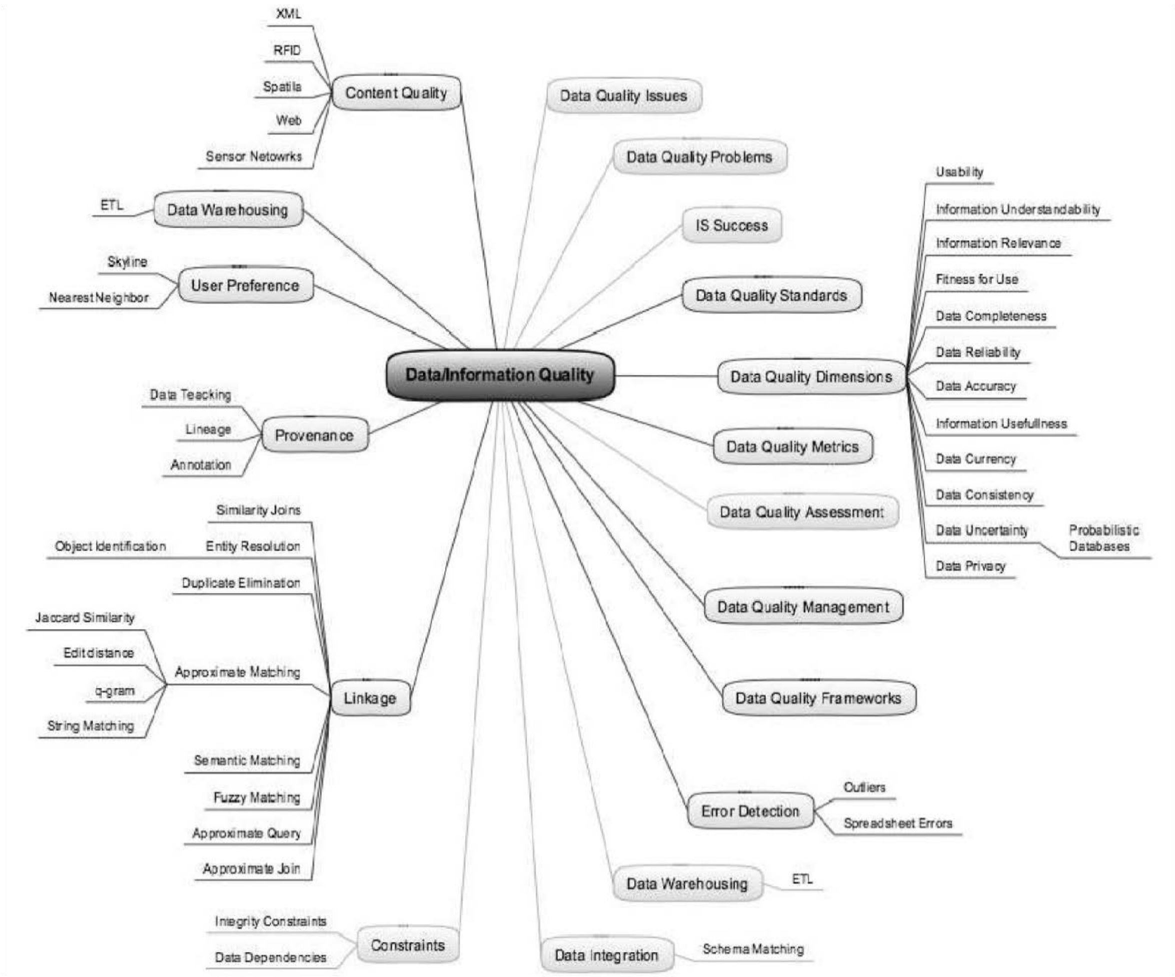

Figure 1. Taxonomy of Keywords

A review of the second level keywords identified that several had 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 (see Figure 1). 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. [7],[4], [8]. A number of augmentations were made to the list, including some further categories of the so-called 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*. Figure 1 presents the developed taxonomy.

The second objective of our study and the prime focus of this paper was to elicit industry response against the taxonomy. Accordingly a survey instrument was designed and pilot tested. The survey instrument was based on the taxonomy derived from the literature analysis stage and 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 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 included questions related to seven key data quality concepts that were identified via the taxonomy through a (keyword) grouping of the main research areas, namely:

1. **Data Quality Assessment**: Includes statistical profiling, error detection, metrics, and methods for cost estimations.
2. **Data Quality Frameworks**: Includes governance, benchmarking, best practices, standards, etc.
3. **Data Modelling and Design**: Includes schema quality, availability of documentation/meta-data, difficulties due to legacy systems etc.
4. **Data Integration and Linkage**: Includes schema matching, duplicate detection/entity resolution, use of master data, different formats, ETL/Data Warehousing etc.
5. **Data Constraints and Rules**: Includes conformance to business rules, data standards, key/id management etc.
6. **Data Lineage**: Includes provenance, data tracking, source attribution, ownership etc.
7. **Data Acquisition and Presentation**: Includes data interfaces, data entry, data collection/upload e.g sensor & RFID data, multimedia data etc.

Although the taxonomy represents a much larger diversity in the research concentration areas, the intention for the above grouping was to reduce the number of questions in the survey while ensuring as broad coverage as possible. Hence, for example, data governance, standards and practices were grouped under *Data Quality Frameworks*. The questions on the concepts were designed to elicit an evaluation of the importance of the concepts within the respective organizations. Further questions then aimed to uncover how successfully these concepts have been implemented (practically used) in their organizational context. Moreover, the participants were asked to recall, in the context of a recent data quality project, the issues and challenges that the organisation faced with respect to the seven identified concepts. Finally, the participants were also asked to identify any further concepts that the provided list of seven concepts did not cover.

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 on SurveyMonkey (www.surveymonkey.com/s/teaching-and-research-data-quality). Responses were however elicited through both print as well as online means. Print versions of the survey were distributed at one local data quality conference to over 100 delegates and 27 responses were collected. Secondly, an invitation was sent to a targeted mail list of 110 experts, practioners /professionals from which an additional 25 responses were collected totalling 52. Lastly, the survey announcement was also posted through the newsletter of the International Association of Information and Data Quality (iaidq.org) resulting inoverall60 responses (at the time of writing this paper).

In the next section, we now discuss the main results of the survey, including insights gained from both quantitative and where relevant qualitative responses.

SURVEY RESULTS

In our 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 approached that stems from a lack of standardised or best-practice education.

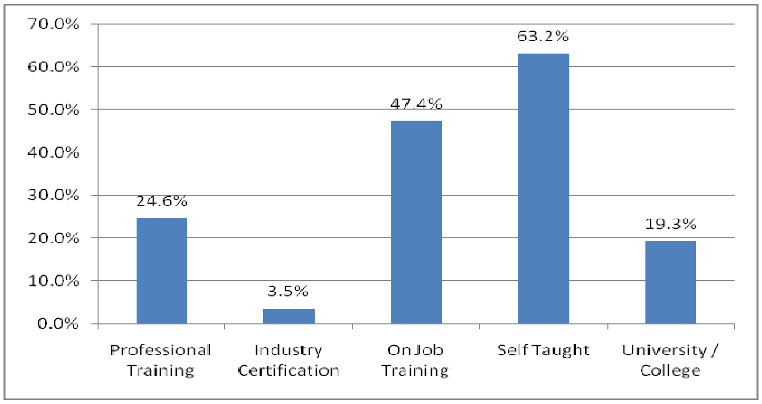

Figure 2: Level of data quality training

While we expected to find significant differences in the sources in which data quality problems are found, the survey results indicate that all sources are quite problematic (see Figure 3). It is understandable to expect that transactional data, external data and legacy data, in particular, are sources of data quality problems. However, an unexpected finding is the indication that data warehouses and business intelligence (BI) data are also problematic (37% and 39% of respondents, respectively). Given the amount of data cleaning required for a data warehouse and BI implementation, these figures are surprisingly high and have implications for the quality of decision making in organizations.

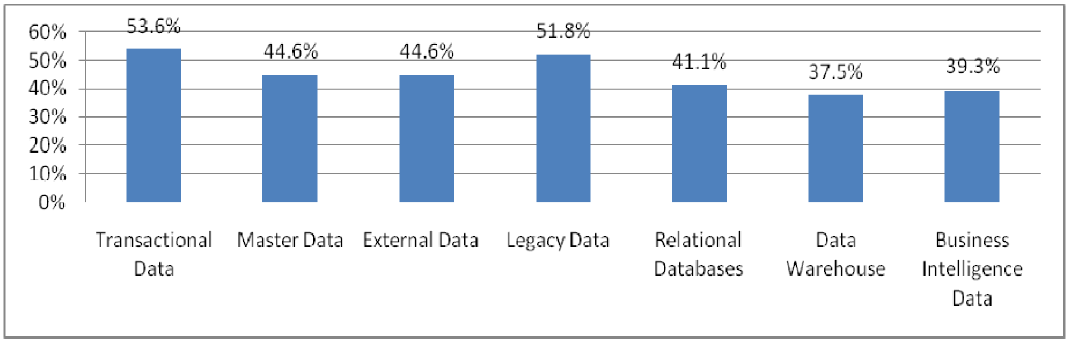

Figure 3: Sources of data quality problems

When asked how data quality problems are typically detected, the respondents indicated that the majority of detections stem from complaints (either by customers or employees), see Figure 4. More worryingly, only 51.8% of respondents agreed that problems were detected by a dedicated data quality management team and almost 43% indicated that problems were found by chance. While in addition to these cases, problems are reported in periodic audits and data migration projects, the results are alarming and indicate problems with organizational approaches to data quality management. A lack of a standardized and systematic approach for data quality problem detection results in situations where the majority of problems are detected by the customer, leading to reputational damage. When we compare responses relating to the data quality training of respondents and responses regarding how data quality problems are detected, we find that, among the professional training category the highest amount of problems (75%) are detected by the dedicated teams on data quality. In all other categories (on job training, self-taught, university college) the highest amount of problems are detected due to complaints by customers and employees.

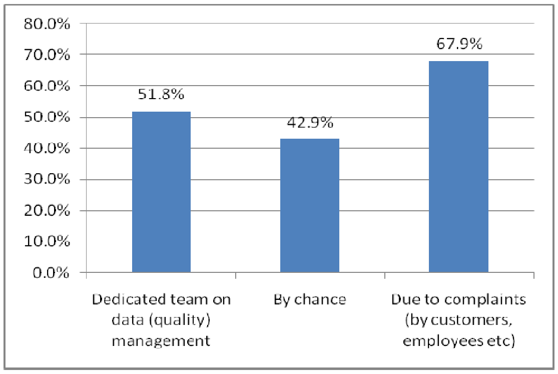

Figure 4: How data quality problems are detected

While the majority of data quality problems are reportedly detected through complaints, a contrast of these responses versus a perceptional judgment of the overall success of the company's data quality management approach shows that such cases clearly do not result in good data quality management (Figure 5). In particular, it is clear that respondents who indicated that data quality problems are detected by complaints are least satisfied with their overall approach to data quality (an average score of 3.8/5). Not surprisingly, respondents whose organizations have dedicated data quality management teams are more satisfied with the overall data quality management approach.

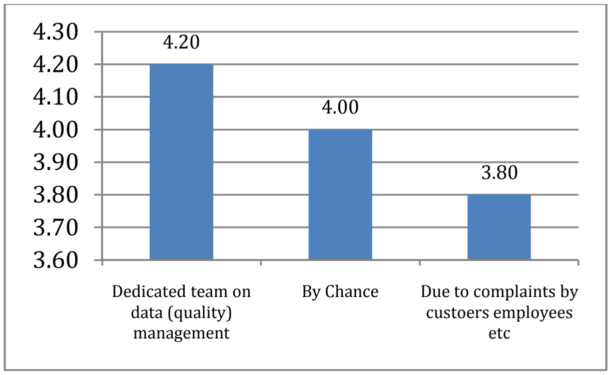

Figure 5: Overall Success of Data Quality Management by type of data quality detection

As discussed in the previous section, the survey was developed based on the data quality taxonomy that was derived from data quality research over the last two decades. The respondents were asked for a perceptional rating (on a Likert scale of 1-5) of the importance of the various data quality management aspects and also the level of effectiveness with which the aspect was addressed in their organization. In the following, we report on the details of these perceptional assessments.

When considering **assessment of data quality**, only 52% of the respondents indicate that it has a very high level of importance from a general data quality management perspective (Table 2).Again, this is a surprising finding given the importance of understanding the overall level of data quality in organizational systems. In addition to the surprisingly low rating of data quality assessment, the actual implementation of data quality assessment is extremely poor, with only 10.9% of participants indicating that data quality assessment is effective in their organization. Even when considering responses that rated the effectiveness at a medium level, the overall respondents are still less than 50%. Almost as many respondents indicated that data quality assessment is done poorly in their organization.

	Very Low/Poor	Low	Medium	High	Very High/Well
General Importance	17.4%	2.2%	8.6%	19.6%	52.2%
How well has this been addressed	31.3%	19.5%	20.9%	17.4%	10.9%

Table 2: Data quality assessment

The respondents were also given the opportunity to explain their perceptional assessments with an open ended question. After coding the text responses. We found that data quality assessment is still a relatively new concept to industry. Hence, lack of knowledge, skills and organizational support has prevented them from having a successful approach to data quality assessment. 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 problems, which in turn explain the surprisingly low ratings in the data quality assessment question. It is also apparent that many organisations spent a large amount of resources for improving specific data sets, rather than investing towards a consistent methodology for data quality assessment or addressing the root causes which cause 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.

When asked about the importance of **data quality frameworks**, 54.3% indicated that they consider having a data quality framework in place as being very important (Table 3). However, at the same time over 70% (26.1%+26.1%+23.9) of respondents indicated that they don’t have an appropriate data quality framework in place. Our further analysis revealed that out of the 54.3% who indicated its high importance, 57.7% indicate that this aspect is not properly addressed in their organisations.

	Very Low/Poorly	Low	Medium	High	Very High/Well
General Importance	6.5%	8.7%	10.9%	19.6%	54.3%
How well has this been addressed	26.1%	26.1%	23.9%	15.2%	8.7%

Table 3: Data quality frameworks

An analysis of the open-ended question that seeks to understand the respondents’ ratings showed that most of the organizations have made some attempt to establish a data quality framework but need further guidance to resolve the conceptual level issues and implementation challenges to make it a success. “No proper framework” was a frequent feedback implied by respondents (around 48% of responses to the open-ended question). Other issues varied from operational level day-to-day concerns, through to high level governance issues.

Ownership and responsibility issues with regards to data, long term practical viability of the current frameworks, employee involvement in defining implementing and maintaining frameworks, lack of consensus or confidence about existing frameworks, were some of the concerns. Interestingly, the respondents who indicated that they had proper frameworks in place, also commented about the financial savings obtained as a result of good data/information governance in their organizations.

Data modelling and design is another aspect of concern, with over 50%ofrespondents indicating it to be an activity of very high importance in their organization (Table 4). Similar to the previous finding, however, the results indicate that only 8.9% are satisfied that this aspect has been addressed very well in their organization. In their text responses seven participants quoted "No metadata ", "No documentation "as an explanation to the disparity of their rating.

Out of all those who have given explanations, 55% of them quoted that their meta data documentation is not complete. The main reason seems to be the unavailability of metadata of legacy systems while lack of proper enterprise level architecture, lack of conceptual data models, not following software development/implementation life cycle practices were some of the other reasons for the disparity quoted by the respondents. Three responders mentioned that the limited capabilities in their modelling tools resulted in gaps in their data models leaving some thoughts on the modelling tools available in the market. Further around 40% indicated that they are giving high priority to data modelling activities in their current quality initiatives.

	Very Low/Poorly	Low	Medium	High	Very High/Well
General Importance	4.4%	8.9%	20.0%	15.6%	51.1%
How well has this been addressed	11.1%	37.8%	28.9%	13.3%	8.9%

Table 4: Data modelling & design

Over 40%of respondents indicated that **data integration & linkage** is of very high importance in their organisation (Table 5). Altogether95.5% indicated this to be at least a moderately important aspect in their organisation. However, again, the ratings of how well integration & linkage are addressed in the respective organisations are low. Only just over 20% of respondents feel that their organisation has addressed the issue well or very well. The respondent comments indicate that many organizations have issues regarding master data integration with legacy systems. They also face issues regarding ETL (Extract Transform Load) and data warehousing. Conflicts in the organizational information landscape (due to frequent structural changes) and unstructured legacy systems appear to be a root cause for some of these issues.

	Very Low/Poorly	Low	Medium	High	Very High/Well
General Importance	4.4%	0.0%	26.7%	24.4%	44.4%
How well has this been addressed	15.9%	38.6%	25.0%	9.1%	11.4%

Table 5: Data integration & linkage

In response to the importance of **data constraints and rules**, over55% respondents consider this aspect as very important (Table 6). While the implementation of this aspect of data quality management appears to be less problematic (with over 37% of respondents indicating it is addressed well or very well in their organisation), the majority of respondents still have a relatively low assessment of how this aspect is addressed in their organisation. Based on their text responses around 60 % of professionals agreed that they have not yet reached the full potential regarding their initiatives for a variety of reasons, e.g. non alignment of IT and business teams.

	Very Low/Poorly	Low	Medium	High	Very High/Well
General Importance	4.4%	2.2%	15.6%	22.2%	55.6%
How well has this been addressed	20.0%	15.6%	26.7%	31.1%	6.7%

Table 6: Data constraints & Rules

As far as **data lineage** is concerned, problems regarding the ownership and privacy concerns about data have been raised. Just over 67% (30.2% + 37.2%)of the respondents indicated the high to very-high level of importance of data lineage yet only 26% (11.9% + 14.3%) are satisfied with establishing data lineage in their organisation (Table 7).

One respondent commented "I believe it is more about people and processes than techniques and tools when it comes to data lineage". Based on the text responses, data ownership/responsibility issues appear to be the key barriers in establishing data lineage (37.5% respondents). Some organizations have implemented the concept only for their key data elements (8.3% respondents).

	Very Low/Poorly	Low	Medium	High	Very High/Well
General Importance	4.7%	9.3%	18.6%	30.2%	37.2%
How well has this been addressed	21.4%	26.2%	26.2%	11.9%	14.3%

Table 7: Data Lineage

Over 67% (30.2% + 37.2%) of the respondents consider **data acquisition and presentation** to be of high to very-high importance in their organisation, while 25% (11.9% + 14.3%) are satisfied with their organizational approach to data acquisition/presentation (Table 8). Based on the text responses, some organisations (16.33% respondents) use BPM (Business Process Management) technologies, and best-of-breed applications, to manage effective data acquisition. 53% of the respondents believe that they have not reached the required potential in acquisition and presentation of data, despite the use state of the art technologies (including automation).

	Very Low/Poorly	Low	Medium	High	Very High/Well
General Importance	4.7%	9.3%	18.6%	30.2%	37.2%
How well has this been addressed	21.4%	26.2%	26.2%	11.9%	14.3%

Table 8: Data acquisition and presentation

Finally the respondents were asked to rank the relative importance of the above seven data quality aspects.The survey results (see Figure 6) indicate that data quality assessment is considered to be the most important aspect, followed by data quality frameworks. Data lineage appears to be the least important. Despite the importance of data quality assessment and data quality frameworks, there is a clear indication from the above responses that these aspects of data quality management are still poorly addressed in organizations.

An analysis of the open-ended survey questions relating to the overall ranking, hurdles and outcomes indicates that there are some common organizational and technical hurdles that affect success of data quality projects in organizations. For example, convincing senior management to invest resources on data quality is a significant concern of many professionals (32% respondents indicated this issues). 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. IT/Business alignment appears to be another hurdle that affects data quality initiatives (14% respondents). In particular, systems developed without a long term vision subsequently encounter limitations in facilitating quality of data. Inappropriate software/modelling tools and legacy systems are other mentioned hurdles that are of concern to data quality initiates.

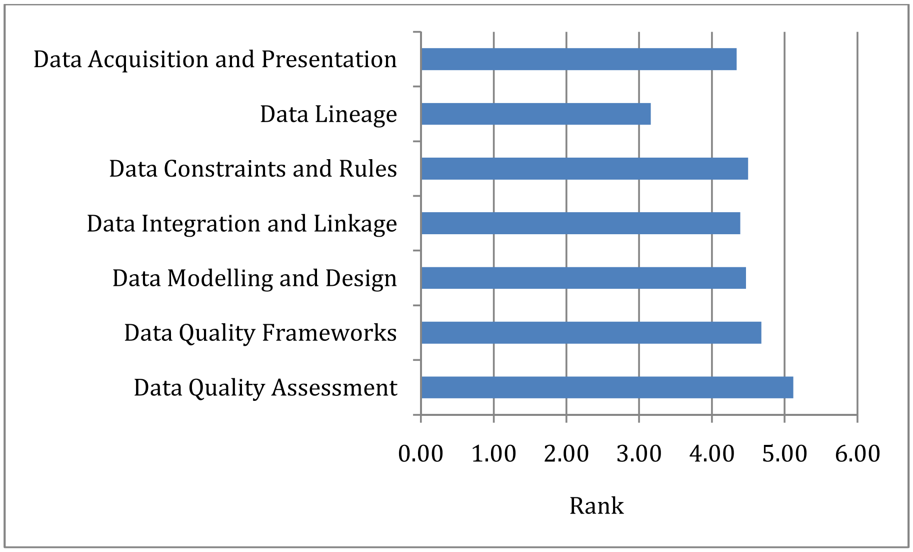

Figure 6: Relative importance of the concepts

We also related the perceived success of data quality projects and initiatives with the size of the organization. Overall, larger organizations appear to have most success in managing data quality (see Figure 7). Organizations with more than 5000 employees seem to have managed data quality more effectively. Not surprisingly they have more resources to dedicate on data quality initiatives and also since they have a greater necessity to manage data quality due to the larger scale of business operations (resulting in large data volumes) and large number of software systems in place.

Small organizations are least satisfied with the effectiveness of their approach. On one hand this is surprising as small organizations are more likely to have a smaller number of systems in place and perhaps low volume of data. However, on the other hand, small organizations may be more unlikely to have the appropriate budget and dedicated personnel for data quality management.

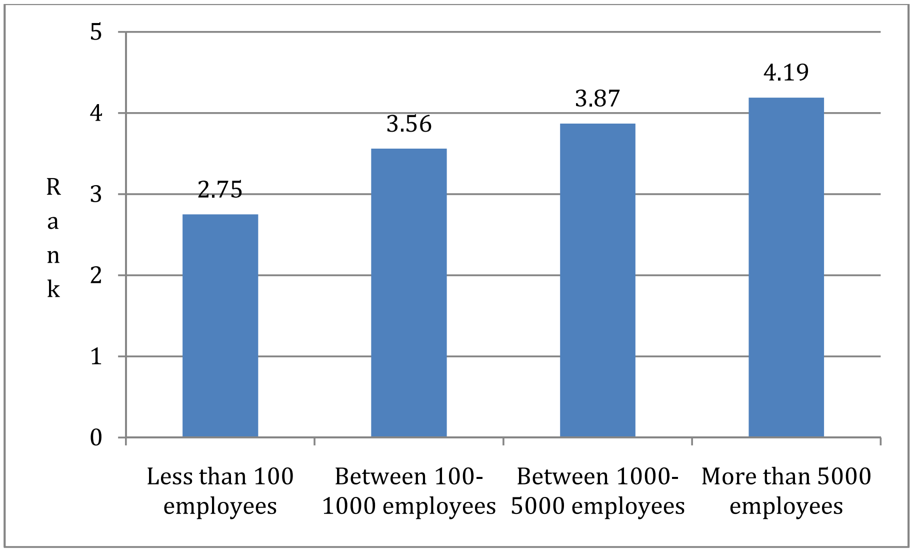

Figure 7: Effectiveness of data quality management by organization size.

CONCLUSIONS AND OUTLOOK

Data quality research has generated a substantial body of knowledge over the last two decades and beyond. These contributions have emerged from both technical and management venues. Due to the cross-disciplinary nature of data quality research it has been a challenge to identify the central themes and associated methodologies, and particularly difficult to exploit the synergies between disparate but complementary contributing communities. In our study, we have attempted to present a representative taxonomy of data quality research over last two decades spanning organizational, architectural and computational aspects. Additionally, we have contrasted the identified themes and topics represented in the taxonomy with industry perceptions and feedback.

Our findings indicate that on average around 70% of the respondents agreed on each of the data quality concepts mentioned in the taxonomy as having importance in their current organizational context. However, even where there have been the most significant contributions from research, for example data quality assessment and data quality frameworks emerged has the most widely studied topics [10], these remain to be a major concern in industry. Although there are many reasons for the low ranking by industry on the effectiveness of these data quality aspects, one reason that perhaps warrants attention from academia is the lack of mechanisms for raising awareness of research results. In a final question of our survey, we asked respondents for their feedback on the need for professional education on data and information quality management. The question has a resounding 97% positive response, which is an indication for the need for extended and targeted educational initiatives led by research and academia.

REFERENCES

[1] Benbasat, I. andZmud, R.W. "The identity crisis within the IS discipline: Defining and communicating the discipline's core properties". *MIS Quarterly*, 27(2). 2003. pp.183-194.

[2] Chen, C. , Song, I.Y. and Zhu, W. "Trends in conceptual modeling: Citation analysis of the ER conference papers (1979-2005)". T*he 11th International Conference on the International Society for Scientometrics and Informatrics*. 2007. pp 189-200.

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

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

[5] Harte-Hanks Trillium Software 2005/6 Data Quality Survey, 2006**,** http://infoimpact.com/Harte-HanksTrilliumSoftwareDQSurvey.pdf.

[6] Juran, J.M. *Quality control handbook*. McGraw-Hill Publishing Co, 1962.

[7] Lima, L.F.R., Macada, A.C.G. and Vargas L.M.. "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, 2006.

[8] Madnick, S.E., Wang, R.Y., Lee, Y.W. and H. Zhu. "Overview and Framework for Data and Information Quality Research".*Journal of Data and Information Quality (JDIQ*), 1(1).2009 pp.1-22.

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

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

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

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

[13] Yonke, C.L., Walenta, C. and Talburt, J. R.," The Job of the Information/Data Quality Professional", *International Association for Information and Data Quality*,2011.

[14] 12th Annual Global CEO Survey Redefining Success, Price Waterhouse Coopers, 2009, http://www.pwc.com/ceosurvey**.**