Journal of Theoretical and Applied Information Technology
15th November 2019. Vol.97. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SYSTEMATIC REVIEW OF DATA QUALITY RESEARCH

[1]M.IZHAM JAYA, [2]FATIMAH SIDI, [3]LILLY SURIANI AFFENDEY, [4]MARZANAH A. JABAR, [5]ISKANDAR ISHAK

[1,2,3,5]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia

[4]Department of Information System and Software Engineering, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia

Corresponding Author: fatimah@upm.edu.my

ABSTRACT

Data quality drawn a major concern when dealing with data especially in the event that insightful outputs is needed. Research in data quality emerged in various topics and diversification in known knowledge and used approach is inevitable. In this paper, we apply systematic review study to explain the landscape of data quality and to identify available research gap by using categorization and mapping. Our search scope is limited to research articles from journals, conference proceedings and magazine published between 2010 until 2016. We defined three types of main categorization to map the selected research articles and to answer our research questions. These categorization focus on research topics, research type and contribution type. On average, fifty-four research articles related to data quality were published every year. This number shows the importance of data quality research in various research topics such as online users, database, web information, sensors and big data. This study also indicates that almost half of the selected articles proposed a novel solution or an essential extension of an existing data quality technique. Moreover, most of the selected research articles belongs to the model type in the contribution category. Our mapping also suggests that obvious contribution disparity happen between contribution in metric type and model type category.

Keywords: *Data Quality, Information Quality, Systematic Review*

1. INTRODUCTION

Research in data quality has been widely applied in many areas such as health, finance, corporate organization and information system. It included the needs to uniquely understand the dimensions of data quality, measurement methods, assessment techniques and improvement process in each domain. Lately, data quality also drawn attention in the big data wave. Although the numbers of collected data in big data are huge, but the quality of these data and the impact of having low data quality are still in debate. Despite the fact that the landscape of data quality research is wide, researchers and practitioners have agreed that high quality data is always referred to the data which is fit for use and meet the criteria set by the data user according to its domain [1]–[5].

Many years of research in data quality has seen a lot of advancement being made in defining data quality dimensions, measurement techniques, assessment models and improvement methods. Diversification in used approach and research domain cannot be avoided and there is a need to synergies the existing knowledge for further expansion. The landscape of data quality research has also changed from years to years to adapt the new coming technology such as big data, crowdsourcing, sensors network and online web data. Such changes opened the opportunity for collaboration among research community.

The categorization and mapping of data quality research in visual representation would help the researcher to find and fill the gap. Mapping is used in this review to ascertain gap and to direct readers towards the extension of existing knowledge in data quality. Previous literature review in data quality research has not adopted mapping technique in visual representation to describe data quality research and explain the gap found. For example, a knowledge diffusion in data quality research using main path analysis [6] and thematic analysis of data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

quality research [7]. Furthermore, both review articles did not categorized and described the research approach and contribution type of each research paper included in the review.

Our study examined the landscape of data quality using systematic review study. We select research articles from journals, conference proceedings and magazine published in 2010 until 2016. This study highlights and critically discuss topics in data quality research, types of research, types of research contribution, the research methods being used in each of selected articles and techniques used to solve the discussed problems. The findings are then categorized and presented in a visual representation using bubble plot.

This paper is organized as follows. We explain the methodology used for this review in Section 2. Then, in Section 3, we present the results of this study. Section 4 discuss the answers to our research questions. Lastly, in Section 5, we conclude our study. In most of the research articles being reviewed, data quality and information quality are always used as synonym to each other. We shall follow this convention in this paper.

2. METHODOLOGY

This study has been conducted based on the guidelines in doing systematic review proposed in [8]. The execution of this study involved eight stages including the formulation of research questions, keywords identification, selection of article resources, searching process, selection of articles, data synthesis and the categorization of data. We summarized the execution of this study in Figure 1.

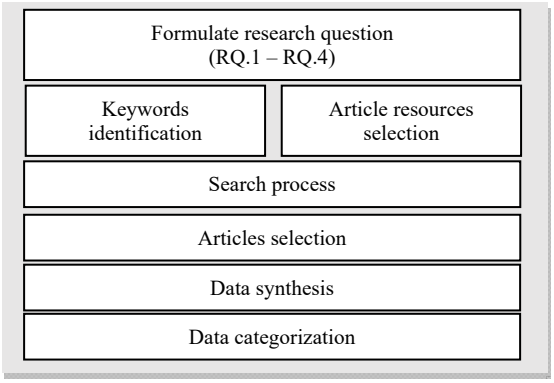

Figure 1: Execution of review protocols

The following section explained the execution phase in detail.

2.1 Research Questions

Our primary aim is to explain the landscape in data quality research and to identify available gap by using categorization and mapping. For this reason, we considered four research questions to be answered:

RQ.1 What are the topics being discussed in data quality research?
RQ.2 What types of research has been carried out in each data quality topics?
RQ.3 What are the types of contribution being proposed so far in each of data quality research topics?
RQ.4 What kind of research methods is being used in data quality research?

2.2 Search Strategy

We developed search strategy based on three factors including keywords identification, article resources selection and search process.

2.2.1 Keywords identification

Keywords for articles searching were built based on these criteria [8]:

(a) Keywords must be identified from the research questions.
(b) Identification of keywords from other relevant studies and review articles.
(c) Synonym, abbreviation and alternative spelling should be considered as keywords.
(d) Usage of the Boolean OR for synonym, abbreviation and alternative spelling.
(e) Usage of the Boolean AND for linkage between keywords.

We identified the following keywords based on the criteria and used during the search process: "data quality" OR "quality data" AND 'information quality" OR "quality information".

2.2.2 Article resources selection

In order to ensure that all the relevant articles were included, we took into consideration previous review articles available in data quality research including [9][10][6][11] and examined the article resources used. As a result, four journals, three conference proceedings and one magazine which considered important in data quality research were selected. Table 1 listed the selected journals, conference proceedings, and magazine. We provided the SCImago Journal & Country Rank (SJR) from Scopus website for each article resources to justify the influences of selected

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

resources as the SJR is calculated based on the citation received.

Table 1: Selected Journal, Conference Proceedings and Magazine

Resource	Type	Acronym	SJR
Journal of Data and Information Quality	Journal	JDIQ	0.318
Decision Support Systems	Journal	DSS	2.262
Information and Management	Journal	IM	1.381
IEEE Transactions on Knowledge and Data Engineering	Journal	DKE	2.087
International Conference on Information Quality	Conference Proceedings	ICIQ	N/A
Very Large Databases	Conference Proceedings	VLDB	1.052
ACM SIGMOD	Conference Proceedings	ACM SIGMOD	0.779
Communication of ACM	Magazine	CACM	1.910

2.2.3 Search process

We performed automatic search in each of the selected article resources using keywords as defined earlier. We scoped the searching process to only include articles being published between 2010 and 2016. Each articles found has been label as 'prospective'.

2.3 Articles Selection

During the searching process, a total of 409 prospective articles has been retrieved. Manual screening has been done on the articles title, abstract and keywords to eliminate any irrelevant and repetitive study. As a result, 381 articles was selected and labeled as 'relevant'. Thereafter, an inclusion and exclusion criteria has been applied to the relevant articles. During the process, a manual screening and a brief study on each of relevant article has been conducted. Consequently, 374 articles has been selected and being labelled as 'selected'.

2.3.1 Inclusion and exclusion criteria

Manual screening of the relevant articles has been done based on the inclusion and exclusion criteria with aim to select articles that reflected our study. For this reason, any articles published between 2010 until 2016 and satisfied the following inclusion criteria were labelled as 'selected'. The inclusion and exclusion criteria used in this study are adopted from [8].

a) Articles published in English language.
b) Articles that discussed on data quality or information quality.
c) Articles which are able to answer at least one of our research questions.

We excluded articles which has the criteria described below:

a) Articles not published in English language.
b) Articles not related to any of our research questions.
c) Duplicate articles. In case of duplicate articles, outdated articles and most incomplete content were excluded.

2.4 Data Synthesis

In this phase, we standardized, organized and consolidated evidences from the selected articles in order to answer our research questions. We listed the criteria used to collect evidence from the articles content in Table 2. The criteria were used to help us in finding precise answers for each research questions and to guide us during data synthesis phase.

Table 2: Criteria for data synthesis

Criteria	Description
Identification number and bibliographic references	Unique identification number, article title, author, year and source
Focus of research	Research topic, research problems and article contribution
Type of research	Type of research conducted in the articles e.g. solution proposal, validation research, evaluation research
Research method	Type of analysis and evaluation used in articles e.g. statistical analysis, experimental evaluation, empirical evaluation
Application domain	Application domain of the research e.g. financial, health

Data collected during this study consisted of qualitative and quantitative data type. Data was tabulated in this study using table, graph and pie chart format. Findings of this study are then being presented in mapping to aid better understanding of data quality landscape. We explained the types of collected data in Table 3.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: Collected data

Quantitative	Qualitative
Number of articles by sources	Research topic
Number of publication by year	Research problem
Number of contribution type by sources	Type of research
Number of type of research by sources	Article contribution
Number of research topic by sources	Research method
Citation number in Scopus and Google Scholar	

Qualitative data of research topic and research problems were synthesized to answer RQ.1. Categorization of the discussed topic in data quality research has been made and a pie chart has been used to visualize the findings in quantitative format. For RQ.2, we synthesized type of research data and further generalized the data into 6 categories and presented the findings in a pie chart and bar graph. As for RQ.3, we categorized articles contribution into 6 main categories by synthesized the article contributions. A pie chart and bar graph was used to summarize the most type of contribution in data quality research. In order to answer RQ.4, we categorized the research method used in the selected research articles according to its related research topic and present the finding in a table. Lastly, we mapped the findings of RQ.1, RQ.2, RQ.3 and RQ.4 into a bubble plot to explain the available gap in data quality research.

We confirm the validity of these finding by doing crosscheck between researchers involved. A random sample of synthesized data has been validated in a group discussion among the researchers. In a case of conflicted findings, discussion had been done and mutual agreement had been achieved.

2.5 Data Categorization

2.5.1 Categorization of data quality research topics

Categorization of data quality research topics are done based on the classification proposed in [12]. The classification anchored by four main topics including impact of data quality, data quality in the context of computer science and IT, technical solution in data quality related to database and data quality in curations. As recent technologies emerged, current domains and techniques in data quality such as web, crowdsourcing, big data and online communities has not been discussed in details by the previous work. We extended the discussion in this study by included current domains and techniques available in data quality research. We categorized data quality research topics into three and eliminated data quality in curations from our research topic categorization. We believed that data quality in curations should be discussed in details in more related subcategory such as measurement and assessment, security and database cleansing.

We thoroughly read each selected articles to examine the topic being discussed. Then, we classified the articles into three main topics as listed in Table 4. Most frequent words such as cost, virtual communities and mobile application in each articles were considered as theme and determined the related sub topics.

Table 4: Classification of data quality research topics

Main topics	Sub topics	Description
Data quality impact	Application usage, organization strategy and policy, cost and benefit, IT management, organization change/process, online users, big data[1].	Research articles that discussed the impact of data quality, challenges and data quality related issues
Technical solution in database area	Database cleansing, database integration and data warehouse, entity resolution and record linkage, data provenance, uncertainty and probabilistic data, security[2].	Research articles that provided technical solution in data quality related to database
Technical solution in computer science area	Measurement and assessment, information system, web, sensor[3], protocol and standard,	Research articles that provided technical solution but not related to

[1] Online users and big data are newly added subcategories in data quality impact topic compared to previous subcategories in [12]

[2] Database security is a newly added subcategories in technical solution in database area research topic.

[3] Web and sensors are newly added subcategories in technical solution in computer science area research topic.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Main topics	Sub topics	Description
	network, security.	database area

2.5.2 Categorization of research type

We categorized the selected articles into six related research types including solution proposal, validation research, evaluation research, conceptual proposal, experience paper and opinion paper. In doing the categorization, we read through the methodology and conclusion chapter from each of selected articles. The categorizations are determined by the type of research approach and the methodology used by the author. For a better understanding, in Table 5, we provided the classification criteria as proposed in [13] which are adopted in our categorization of research type.

Table 5: Categorization of research type

Classification criteria	Description
Solution proposal	Data quality research that proposed a novel solution or an essential extension of an existing technique
Validation research	Data quality research that examined solution proposals that have not yet been applied in practice. It can be presented by experiments, prototype, simulations and mathematical analysis.
Evaluation research	Data quality research that examined solutions that have been already applied in practice. Results can be presented in case studies or field studies.
Conceptual proposal	Research articles that represented things that have already existed. It can be presented in taxonomies and theoretical frameworks.
Experience paper	Authors explained the process, achievement and experiences from projects.
Opinion paper	Discussed the suitability or unsuitability of a specific technique or a tool based on personal opinion of author.

2.5.3 Categorization of article contribution

Article contribution has been identified based on the introduction, result, discussion and conclusion in each of selected articles. In order to identify significant contribution of each article, we relate each finding with defined problem statement. The contribution is then categorized into six main categories including tool, model, metric, enhancement, technique and framework. The categorization scheme has been outlined from [14]. However, we added framework as a new type of contribution and widen the classification area to suit data quality research. We explained each type of contributions in Table 6.

Table 6: Types of Contribution in Data Quality Research

Contribution	Explanation
Tool	Research paper that concentrated on data quality tools, data quality assessment model or prototype
Model	Research paper that discussed data quality dimensions, relationship among dimensions, data quality challenges and the evaluations of existing data quality approaches
Metric	Research paper that proposed a data quality metrics to measure data quality dimensions
Enhancement	Hybridization of existing data quality framework, data quality model or data quality methods
Technique	Research paper that proposed a data quality technique
Framework[4]	Research paper that concentrated on data quality framework

3. RESULTS

This section summarized the findings of our study. We first presented our search result including the number of prospective articles, relevant articles and selected articles according to articles resources. Next, we discussed the overview of selected articles.

3.1 Search Results

Table 7 tabulates the result of the search process according to article sources and year published. 409 articles retrieved during the search process and 381 articles were found relevant to this study. We then scrutinize the relevant articles and finally selected 374 articles to be included in this study.

Our result showed that JDIQ published most of data quality journal articles with 67 publications. Nevertheless, only 57 articles are selected in this study. DKE published the lowest number of data quality articles with 16 articles. As JDIQ main subject is data quality and information quality research, the high numbers of data quality publication are expected. The same reason applied to ICIQ conference proceedings with 178 articles compared to the lowest number in ACM SIGMOD with 12 articles.

[4] We included framework as a category for contribution in data quality research due to relevancy reason.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 7: Sources searched for the years 2010 – 2015

Source	Year	Prospective	Relevant	Selected
JDIQ	2010	5	4	3
	2011	7	7	6
	2012	14	14	13
	2013	8	8	7
	2014	9	8	8
	2015	12	10	9
	2016	12	11	11
DSS	2010	2	2	2
	2011	3	3	3
	2012	7	7	7
	2013	5	5	5
	2014	6	5	5
	2015	10	10	10
	2016	7	7	7
IM	2010	1	1	1
	2011	1	1	1
	2012	1	1	1
	2013	5	5	5
	2014	2	1	1
	2015	5	5	5
	2016	2	2	2
DKE	2010	1	1	1
	2011	1	1	1
	2012	1	1	1
	2013	4	3	2
	2014	2	2	2
	2015	5	5	5
	2016	2	2	2
ICIQ	2010	31	30	30
	2011	36	36	36
	2012	27	27	27
	2013	21	21	21
	2014	30	30	30
	2015	11	11	11
	2016	23	23	23
VLDB	2010	13	10	10
	2011	1	1	1
	2012	12	10	10
	2013	13	11	11
	2014	9	6	6
	2015	3	2	1
	2016	2	2	2
ACM SIGMOD	2010	1	1	1
	2011	5	5	5
	2012	0	0	0
	2013	0	0	0
	2014	1	1	1
	2015	3	3	3
	2016	2	2	2
CACM	2010	8	6	6
	2011	6	3	3
	2012	4	3	3
	2013	2	1	1
	2014	2	2	2
	2015	2	1	1
	2016	1	1	1
TOTAL		409	381	374

3.2 Overview of Selected Studies

We further evaluated the result to give overview of publication trend in the area of data quality research. On average, fifty-four selected articles in related to data quality were produced yearly. Sixty-two percent of these articles were published in conference proceedings. On the other hand, forty-one percent of the selected articles were published in journals and magazine. High percentage in conference proceedings was due to the availability of data quality related conference such as International Conference of Information Quality (ICIQ). High acceptance rate for the conference proceedings compared to journal also contributed to the differences. Figure 2 described the source distribution further.

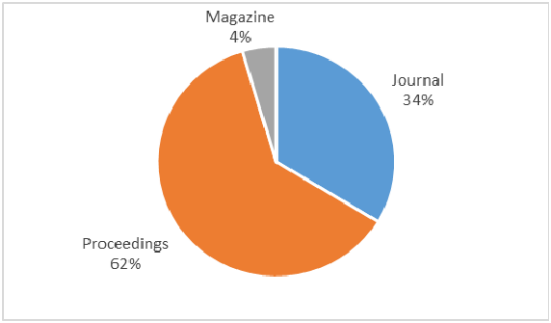

Figure 2: Source distribution of related articles

We also found out that the number of data quality research articles published from 2010 until 2016 are maintaining the same trend as proposed in [6], [9]. The trend shows that data quality research is still significant and retained the same amount of research articles publication each year. Moreover, this trend may indicate the importance of data quality research. We illustrated the trend in Figure 3.

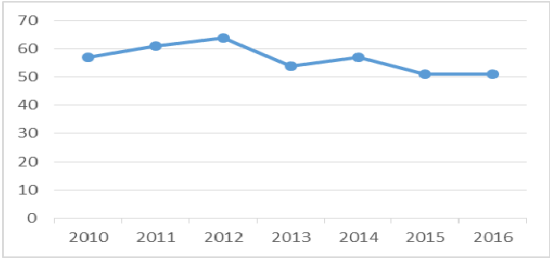

Figure 3: Number of data quality publications by year

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. DISCUSSION

In this section, we discussed the answers to our research questions. At the end of the discussion, we presented a mapping of data quality research landscape and the research gap available for future data quality research.

4.1 RQ.1 What Are the Topics Being Discussed in Data Quality Research?

Data quality research has attracted researchers for many years and thus, the topics of discussion are really broad. We believed that categorization of research topics in data quality are vital to direct researcher attention to the least explored topic. For this reason, we categorized the selected articles according to the topic discussed and classified the research topics into three main areas including data quality impact, technical solution in database area and technical solution in computer science area. As presented in Figure 4, most focus has been given to the technical solution in computer science area. However, only one hundred and twenty-nine research articles focus on data quality impact. It is known that the early stage of data quality research encompassed on the development of data quality knowledge structure [6]. As we are entering the growth stage of data quality research, lower number in data quality impact can be justified. Surprisingly, the numbers of research in technical solution in database area are still low compared to other research areas. We further discussed each topic found within 2010 until 2016 in the next subsection.

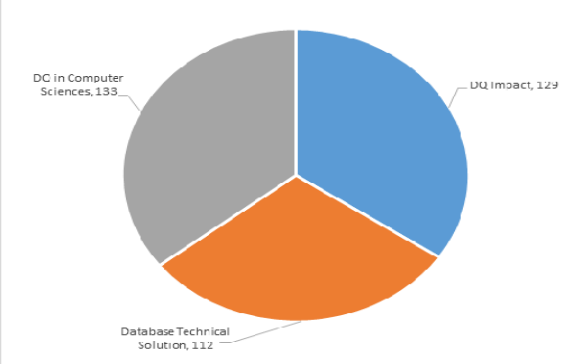

Figure 4: Research topic

4.1.1 Data quality impact

As shown in Figure 4, a total of one hundred and twenty-nine articles analyze and provided explanation to the impact of data quality. The influences can be further categorized into six different subcategories as described below:

- Application usage

Within this subcategory, articles investigate the role of data quality in ensuring success of specific application and proposed ways to measure and improve data quality in specific application including Enterprise Resource Planning (ERP), Business Intelligent System (BIS), Supply Change Management (SCM) and Engineering Asset Management (EAM). From the review, we strongly agreed that high data quality in specific application could increase trust and believability of the application output and further help to ensure success of the specific application usage. For discussion purposes, we highlighted examples in BIS and ERP application in the next paragraph.

A quantitative survey based study was conducted in previous research to understand the correlation between success dimensions in BIS. The research confirmed that information quality played important role in ensuring BIS success and effective data quality management resulted high acceptance of BIS [15], [16]. Data quality capabilities has the same important role in decision environment. The ability of the organization to manage data quality efficiently helps to determine the success of BIS application in decision environment [17] Evidence found from both research justified our earlier suggestion that data quality management is essential in ensuring success of specific application usage. However, in both research, the organizational behavior and its culture in managing data quality is not being fully considered. Level of data quality in organization as we know, reflected by the organization culture towards data quality [18], [19]. Further research could be done to investigate the impact of data quality in different organization culture setup and behavior towards the success of BIS.

In large organization, specific application such as ERP system can be too complex and thus, data quality issues become more difficult to be manage. However, it is still important for the organization to efficiently manage data quality no matter how complex the system is. This is to ensure the success of specific application such as ERP [20]. In ERP, complexity in data production can caused data quality problem even though it has been implemented for a long time in the organization. Complexity in ERP can be reduce by identifying hidden interaction within the ERP application [20]. Additionally, increasing people awareness of the ERP complexity can provide better support for data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

quality management. Yet, the ERP application applied in each company could be differed and thus the challenges in managing data quality problem are varied. We believed that modelling the complexity of specific application in large organization helps to determine the critical factors and relationship among them in order to achieve high data quality. Further analysis could be done to analyze the impact of specific application complexity towards data quality level in organization.

It is essential to explore knowledge from the context of specific application usage especially to support data quality initiative within the organization. We believe that more research will be done within this category in the upcoming years as the application such as ERP and SCM will faces new challenges in big data era. As data volumes, data velocity and data variety increase, complexity of specific application as discussed before cannot be avoided.

- Organization strategy and policy

Specific strategy and policies for managing data quality within the organization played important role in ensuring high data quality. Within this category, reviewed articles investigated organizational issues, influences factor, policy and suggest the effective strategy to manage data quality within the organization. Additionally, the organization should be able to strategize employees' task in managing data and should clearly understand its influences towards organizational data quality level. Data quality problem occurred within the organization when employees have different views in data quality dimensions such as timeliness, accuracy and completeness. The differences caused low quality of data product especially when the needs of other employees which tasks are linked cannot be satisfied and being misunderstood. This inability to meet the needs of data user promotes low level of data quality within the organization. Employee empowerment approach [21] resolved the interdependency problem between employee tasks in the organization. The approach proposed that the employee who collected the data should be given responsibility to process the same data. At the same time, organization should encourage collective quality decision to achieve high quality data. Collective quality decision eliminates individual decision and facilitates common understanding of data quality dimensions. In order to further increase the organization data quality level, a clear goal should be setup to each team and employees. Incentives and rewards should be given to the team and employee who achieved the goal. On the other hand, the strategies outlined in this research have never been applied in a real organization environment. Changes in organization policy and modification in the suggested approach might be essential to adapt with a fast changing organization.

Common understanding of data quality facilitates the organization to strategize business efficiently by avoiding flaws data during decision making. Thus, a proper and well managed data quality policy is needed to create employee awareness when managing data. A data quality metadata (DQM) is one of the example that explained data quality level within the organization. On top of that, DQM also describe data quality according to the organization description. Usage of DQM during strategic decision making improves decision accuracy and decision confidence [22]. However, as DQM usage increased decision time, it is not suitable for task that need to be accomplish urgently. Further enhancement such as DQM visualization [23] can be done to address the limitation of using DQM in urgent task.

Organization expenses and investments information such as organization investment plan and costs are highly valuable to the organization. On the other hand, disclosure of such information might influence positive view towards the organization specifically the aspirations to be more productive, competitive and innovative. Research in [24] investigated the impact of IT investment information disclosure including plans, costs and benefits towards market information quality and factors that encourage managers to disclose such valuable information. Information can be disclosed either in quantitative or qualitative format determined by the assessed risk. This research paves the way for organization to review its information disclosure policy and at the same time to gain more economic value. However, the benefit gains and occurred cost when information being disclosed are varied depends on the context of risk assessment.

- Cost and benefit

Data quality could impact the economic outcome either positively or negatively. In this subcategory, we could see various research has been conducted with aim to assess the effect and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

proposed improvement steps to minimize the effect. In supply chain management, data quality influence the economic benefit of the organization [25]. However, improvements towards data quality required the organization to determine which data quality dimensions that should be improve. Proper study related to the information flow within the organization and data quality dimensions that involved is needed. Furthermore, the identification of data quality dimensions and its relationship is important to minimize the economic effect and further to gain more economic benefit. On the other hand, high data quality does not increased the economic net-benefit of the organizations to the optimum level [26]. However, as this finding is only supported by the assessment of data currency and data completeness dimensions, further evidence from other data quality dimensions is needed to strengthen the findings. As mentioned before, data quality could cause negative impact towards the economic outcome if it is not properly managed. For example, the analysis done in previous research shows that inaccurate project data in the organization cause increment in project cost [27]. As a solution, a filter is used to increase data accuracy in distributed project. However, as the proposed filter required variables that is estimated from the data noise, performance of the filter could be jeopardize in high noise environment as variables are dynamic.

- IT management

We believed that attaining high data quality within the organization entail the ability of information technology (IT) employees to understand their role in managing data quality. Research in this area disclosed IT employee roles and factors affecting IT management decision in data quality management.

IT management should be able to exploit the expertise of knowledge workers in order to improve data quality within the organization. Expertise of knowledge workers can be model to enabled task routing during data cleaning [28]. Such routing is important to maximize the accuracy in verifying changes made to data. Modelling the expertise of knowledge workers enable the IT management to determine suitable employee with the right expertise in handling data. However, this approach will not guarantee optimum data quality result as other factors including workload constraint, cost, latency and workers motivations is not considered in the research article. Considering these factors with balanced workload and expertise will help to improve data quality optimally. Another initiative from the IT management that will positively impact data quality is to appoint a Chief Data Officer (CDO) [29]. Result from this empirical study suggested that CDO appointment improved organization strategic growth and increase the organization data quality capabilities. However, as CDO tenure is not considered in this study, we argued that the duration of appointment could affect the CDO performance and determined result in data quality capabilities.

- Organization change/process

Organization employed data quality knowledge to structure and design processes within the organization. This initiative helps to improve data quality and further increase the level of trust to the data they have. Research in this category investigated the relationship between data quality and organization process and its impact.

As example, the development of IT acceptance model within the organizations suggested that information quality dimensions such as usefulness and ease of use are among important factors to determine IT acceptance rate within the organization [30]. Such findings are useful in the organizations process redesign and familiarizing new IT application to the employees. However, the result may not be generalized as the study was done only to nursing staff. Data quality are dependent to the context of the application and for this reason, further study should be conduct on diverse area to generalize the result. In [31], a study to investigate the negative relationship between input accuracy and output accuracy has been conducted. This study examined the effects caused by high input error probability towards the output error probability. From the findings, it is statistically proven that if inputs are incorrect, the output from OR operation are always correct. This finding is important in allocating resources and to design related task in organization. However, we cannot see any evaluation of this finding in real-life settings. Such evaluations will help us to understand the sign of occurrence in real problem.

- Online Users (Web, intranet, mobile application, virtual communities)

We see more articles related to the online users have been published after 2010 compared to before. These articles included data quality influenced on online web users, virtual communities, mobile

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

application and intranet. Articles in this subcategory investigated, measured and highlighted the importance of data quality towards online users. For example, information quality played an important role to determine online users satisfaction compared to interactivity features [32] [33]. Information quality also played important role to influence online users to buy products from social media [34]. However, consideration should also been given to other factors such as users' computer literacy and online users psychological state as these factors could affect online users in making decision. Previous research as mentioned before does not explained the relationship between computer literacy, psychological state, information quality and it influence against online users.

In mobile application, data quality dimensions such as accuracy and timeliness determined user experienced and continuous usage [35], [36]. Nevertheless, user behaviors and tendency to use mobile application is dynamic. Based on this argument, further research can be conducted by investigating user behavior and information quality effect towards continuance usage of mobile application.

Another study was conducted in [37] to evaluate factors that determined virtual community decision on contributing and consuming information in virtual space. The result suggested that information quality dimensions such as information reliability, objectivity and information format are vital to determine virtual community participation. Information from trusted user are considered as reliable to the virtual community and information that doesn't meet participant objective such as commercial comments will be ignored. In order to maintain online user participations, information should be posed in a simple format and more freedom should be given to users in expressing their experience. As this study implemented in a large virtual community, smaller size virtual community may not produce the same result as smaller virtual community may know each other well and build up trust easily. In [38], a model was developed to examined social capital dimensions which influenced information quality within virtual community. This study suggested that increasing numbers of information in virtual community will not increase the quality of information. Information quality in virtual community is influenced by relational capital and cognitive capital such as trust, mutual benefit, shared language and shared vision. Though, this study measured limited information quality dimensions such as reliability, accuracy, timeliness and relevancy.

Data quality also determined user intention in adopting intranet [39]. A model of intranet quality and acceptance has been developed and used in this research to assess collected data. However, this is just a preliminary work in intranet quality and acceptance model. Thus, it may not addressed the real problem in intranet quality and acceptance. More research is needed and should be tested in more than one organization to produce better result.

- Big data

There is a major concern to the quality of massive data available in big data technology. The debate between large volume of data and the uncertainty level of quality that the data have are still continue. In [40], an empirical investigation using data mining classification methods has been conducted to analyze data size effect towards data quality problem. The result proved that as data volume increase, possibility to have data quality problem are larger. Nevertheless, the result can still be argued as the dataset used in this experiment are still small compared to what we have in real big data application.

Big data application makes used data with unusual characteristic such as vast volume, high velocity and high variety. In regards to these characteristic, existing data quality initiatives is incapable to efficiently manage data quality as some data quality dimensions such as data completeness and concise presentation has become more important than before [41]. Thus, extension to the existing approach especially in data quality management model and assessment methods is required to support the impact of such changes in data characteristic.

A case study to identified data quality issues in Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) and Hyperspectral Imaging (HIS) sensor has been conducted in [42]. The findings outlined several challenges in maintaining big data quality level in AVIRIS and HIS usage including the ability of data collection devices to minimize data error and the inability to validate data quality level in each stage of data manipulation. From the case study, big data quality problems were derived from the tools used in data collection, analysis and processing. Skills and experiences of the staff who involved in managing AVIRIS and HIS were also

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

reported as the factor that contribute to the big data quality problem.

4.1.2 Technical solution in database area

We found 112 research articles which proposed and discussed data quality in related to database. These articles measured, analyzed and improved data quality in database area. We divided these articles into several subcategories as follows.

- Database cleansing

Database cleansing improved data quality by detecting and correcting low quality data. A lot of progress has been seen in this topic with most of the articles trying to automate the database cleaning process and reduced user interventions. As example, USHER, a real time feedback system for data quality improvement during data entry process [43]. USHER make used probabilistic model constructed from the user answers in the online form to enable real time feedback. However, as the user answers played important role in constructing the probability model, cross validation towards the answer can be is very crucial. This important feature was not considered in this research. Another example to reduce user interventions during data cleansing is a reasoning based solution trough ontology [44]. In this approach, conditional functional dependencies (CFD) and conditional inclusion dependencies (CIND) have been used to discover data quality problems. However, constructing the ontology and domain knowledge for reasoning are expensive in term of time and expert involvement. Another approach used a machine learning technique for automatic data correction [45]. This approach learns from user feedback in the cleaning process and further refines its learning model. Future research can be done to extend this research especially by using other dependency rules such as CIND, matching dependencies and matching rules. An algorithm to accurately match database records and repairing them based on conditional functional dependency (CFD) rule and matching dependency (MD) rule was proposed in [46]. However, it is difficult for the user to determine the correct rules to be used.

- Database integration and data warehouse

Crowdsourcing can be considered as a new approach in managing data quality in integrated database and data warehouse. In [47], crowdsourcing approach has been proposed to reduce uncertainty in schema matching during database integration. Correspondence correctness question (CCQ) is used to gathered crowdsourcing feedback. However, there is still quality issues in adopting this approach such as the ability of non-domain expert to clearly understand data and the level of trust in adopting crowdsourced answers. There is a wide challenge in adopting crowdsourcing approach unless we can measure the level of trust in crowdsourced answers.

A belief function theory has been adopted in [48] to assess data reliability in database integration process. In order to resolved inconsistent information and evaluate data reliability, the researcher has proposed a merging technique using maximal coherent subset (MCS). Yet, to gain more information in data reliability, additional sources of information such as user feedback can be considered.

- Entity resolution, record linkage

Entity resolution and record linkage is a technique to discover data records related to the same entity [49], [50]. Discovery of related data records enabled us to solve data quality problems such as data inconsistency and data redundancy. However, the quality of the outcome of this approach depends on the technique used to discover related data records. In [51], evaluation of learning based and non-learning based technique for entity resolution has been carried out. The outcome showed that the learning based technique produced more quality result but the execution time is higher compared to non-learning technique. Nevertheless, the experimental setting in this research used a bibliographic data and does not enough to reflect the scalability of learning based technique. Entity resolution in e-commerce data and biological data may need more robust experiment setup.

New era of big data introduced challenges in managing high data velocity environment. Data are rapidly coming and record linkage need to accommodate this new data characteristic. Research in [52] adopted incremental record linkage approach to enabled linkage record update during data arrival. This approach fit the challenges in high data velocity environment. However, to enable the application of this research in big data, the proposed approach should be able to handle unstructured data type such as text.

- Data provenance

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Data provenance offered information regarding data sources and history of the data which are important and related to data quality dimensions such as accuracy, consistency, completeness and believability [53]. However, provenance information itself are prone to quality problems such as correctness, completeness and relevancy. Research in [53] adhered this problem with an analysis framework for analyzing the quality of provenance. In this framework, technique to analyze annotation, timestamps and structure of provenance traces has been proposed. In [54], a research was conducted to find minimum recovery for missing events in event data based on index and pruning technique. In data provenance analysis, event data played prominent role and inaccurate event data could jeopardize the analysis result. This approach returned a list of possible event recovery and a human intervention is needed to identify accurate recovery solution. Future research should integrate machine learning technique to automate the recovery process and reduce human intervention.

- Uncertainty and probabilistic data

Data could be uncertain and probably associated with multiple values. Data quality research in uncertainty and probabilistic type of data has been conducted to find the relationship between these types of data and data quality. In [55], probabilistic target model was used to enable automatic deduplication process and further improved decision in uncertain condition. Another research in [56], discussed the problems in deteminizing probabilistic data. A branch-and-bound algorithm has been proposed which aim to find near optimal solution. Experiments conducted in this research proved that the proposed solution produced high quality result.

- Security

Research in this category aims to secure database from unauthorized activity that violates data integrity, confidentiality and availability. In [57], an update certificate has been proposed in order to protect data against improper update. Such approach protected database integrity, control unintentional data update and increase data quality in database. However, in case of multiple update happen, protection from improper update could cause latency in uncertified update detection.

Believability towards data are subjected to a true, real and credible data source [11]. Conversely, data source that are open to attacks, unauthorized alteration and deletion could affect data believability. In order to protect database against intruder and to protect data ownership, a watermarking scheme for relational databases has been proposed in [58]. Although the proposed approach does not degrade data accuracy, this approach is only limited to numerical data type. In a real case, signed data and non-numerical data cannot be excluded. Enhancement to the approach by including signed data and non-numerical data are longed-for.

4.1.3 Technical solution in computer science area

As shown in Figure 3, most of our selected articles proposed data quality technical solution in computer science area. We further categorized the selected articles into nine subcategories as following.

- Measurement and Assessment

Many methods have been proposed to measure and assess data quality. For example, score card index [59], Business Process Modelling Notation (BPMN) [60], probability based currency metric [61] and hybrid assessment approach [62]. The notion of these approaches is to evaluate data quality before any improvement being made. The evaluation process is time consuming and required extensive involvement of domain expert especially in a very large collection of data and heterogeneous data format [48], [63]–[68]. As example, the online financial data. In [63], ontology-based framework has been proposed to assess data quality via ontology mapping. In this approach, rectification of duplicated online financial data and missing value has been addressed by using financial knowledge in the ontology. We argued that usage of domain knowledge may not be dynamic enough to accommodate fast changing environment. Furthermore, to enable the application of this research, other data quality dimension such as reliability, consistency and accuracy should be evaluated as well.

In [69], information volatility has been measured to describe uncertainty in the data. Volatility being measured in percentage using standard deviation helped decision maker to ascertain the reliability of data used. This research has been conducted in the context of health care. Yet, it is still in a grey area to adopt this approach

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

in another context of study unless we can confirm that the measurement technique and impact of information volatility remains regardless of the context of study.

- Information system

Articles in this category proposed techniques and methods to manage data quality problems within information system. Managing data quality within a complex information system is challenging as multiple sources of data either hardware or software involved. Challenges included analysis and processing large data collection to ensure high quality data. Research in [70] proposed the usage of agent-based middleware to manage data validation and data consistency in Data Management System (DMS) architecture. Yet, this approach is designed only for pervasive environment in mobile health. Further research can be done using this approach in non-pervasive environment.

- Web

The web has become dominant resource of information in recent years. Large number of information available in the web increased the user acceptance towards web technology. However, the quality of information contained in the web is doubtful even though the numbers are large. For this reason, data quality research has taking place to investigate, assess and improve the quality level of supplied information. User should be given enough freedom to assess the usefulness of web information by providing enough description regarding the quality of web information [71]. Data fusion method has been applied in [72] which proved the possibility to resolve conflicting web information. Yet, conflicting web information is not just between its sources, but also within data category in a single website. As example, in a health website, data about disease A is accurate but likewise, data about disease B is inaccurate. Such examples are important to improve fusion performance. Empowering web users to estimate the quality of information provided by web are important to avoid inaccurate information being consumed. In [73], Support Vector Regression (SVR) methods has been used to assess information quality in web digital libraries. Ranking features can be considered to assist user selection of high quality information.

- Sensor

Sensors have been widely used to collect data in the latest technology of big data and internet-of-things. However, as any other data sources, sensors data are prone to data quality problems such as accuracy. Data quality research in this subcategory proposed methods and techniques to handle data quality problems in sensors data. A Bayesian inferenced-based framework has been proposed in [74] to clean RFID data. This approach takes advantage on data redundancy problem in RFID to improve data accuracy. Alternately, research in [64] combined statistical analysis using Dynamic Time Wrapping (DTW) and ontology for outliers detection in wireless sensor data stream. In this approach, the ontology is used to represent expert domain knowledge and facilitate in reasoning the outliers. Both research used different technique in managing data quality problem in sensor data. DTW in [64] is used to compute similarity between sensor whereas Bayesian inference used in [74] to exploit sensor redundancy. However, validation of approach proposed in [64] has never been done in a real environment. Experimental evaluation using test dataset and generated outliers in [64] is not sufficient.

- Protocol and standard

Protocol and standard impact data quality level within the organization in many ways. For example, confusion in the data format during data exchange can create data consistency and interpretability problem to the organization. Standards are needed to govern data exchange by specifying consensus understanding between organizations. Research in data quality has to fill in the gap by providing solution to improve standards and quality levels within the organization. The implementation of ISO 8000-1x0 for master data exchange using web services has been reported in [75]. As part of the research, service architecture, I8K, has been proposed which included assessment and certification of data quality level in terms of accuracy and completeness. However, more data quality dimensions could be added to the proposed approach. Data quality dimensions such as consistency and timeliness could help the organization to use the exchanged data efficiently and further reduce the cost for data processing.

On the other hand, research also has been done to assess the quality of the standard used. In [76], a framework for data standard quality assessment has

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

been proposed. The proposed framework has not provided redundancy identification in data standard. Redundancy identification can be used to exploit redundancy for completeness.

Within this category, we also found one article that can be sub categorized into network and five articles in security area. However, as the numbers of related articles is small, further description of these sub categories will not be discussed.

4.2 RQ.2 What Types of Research Has Been Carried Out In Each Data Quality Topics?

In this study, we found 140 articles, proposed a novel solution or at least an essential extension of an existing technique. This enormous number remark the advancement and progress in data quality research within 2010-2016. Meanwhile, 77 articles were examined and evaluated various existing approaches in data quality research. The differences between number of published research articles in these two categories are vast, and supported our claim that data quality research is still relevant and in positive growth. Articles that presented existing techniques, tools, models and conceptual frameworks or validated techniques, tools and models that are still not in practice are lower compared to the first two categories. Only 22 articles are categorized as opinion paper. Opinion papers were mostly found in data quality conference proceedings and discussed data quality techniques or tools based on personal opinion. We presented the findings of RQ.2 in Figure 5.

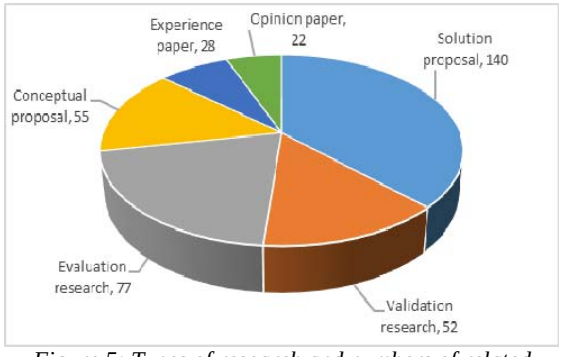

Figure 5: Types of research and numbers of related articles

In order to provide answer for RQ.2, we found that types of research are related to the topics being discussed in the articles. A huge number in solution proposal research type are contributed by articles that discussed technical solution in computer science and database area. Whereas, articles that discussed data quality impact mostly adopted evaluation research type. Assessment of data quality impact and investigation of data quality dimensions relationship using existing models and theory can be the reason of this finding. Existing models and theory are used to support the findings of the investigation or assessment. Surprisingly, we found that conceptual proposals are dominated by articles in data quality impact and technical solution in computer science research topics. We presented the number of articles according to the type of research and research topics in Figure 6.

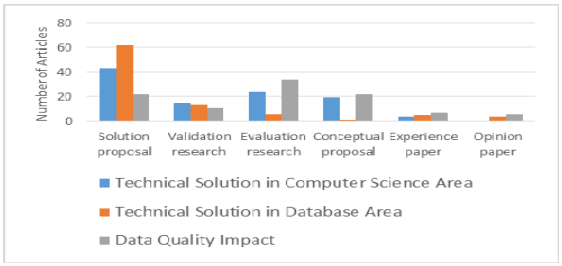

Figure 6: Distribution between types of research and data quality research topic

4.3 RQ.3 What Are the Types of Contribution Being Proposed So Far in Each of Data Quality Research Topics?

Most of the research done within 2010 until 2016 contributed to data quality model. 139 articles investigated data quality dimensions relationship, highlighted the challenges in managing data quality and evaluated existing data quality approaches. Besides, a moderate number of articles contributed to the development of data quality techniques, tools and the enhancement of existing data quality frameworks, models and methods. From this study, we also found 36 articles that proposed a data quality framework. Figure 7 depicted the contribution based on its category.

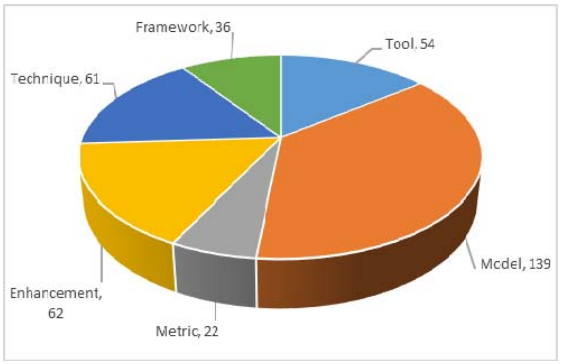

Figure 7: Articles contribution in data quality research and number of related articles

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

We see less attention is given in data quality metrics as only 22 research articles were found related. Metrics were used in data quality research to measure accuracy, timeliness and other data quality dimensions. For example, in [61], a currency metrics based on probability theory has been proposed to assess data currency. In data quality assessment, metrics is used to highlight the root causes of data quality problems and to determine necessary action to improve data quality. Surprisingly, we see most of articles in data quality impact research topics adopted existing metrics in their research.

We found that most of the articles that contributed in model category discussed data quality impact. This are expected as the articles in data quality impact research topics measured, analyzed and explained data quality impact. In order to measure, analyze and explain the impact, investigations of data quality dimension relationship have been conducted. Whereas, articles that discussed technical solution in database area mostly proposed technique or enhancement of existing frameworks, models or methods. The findings are shown in Figure 8.

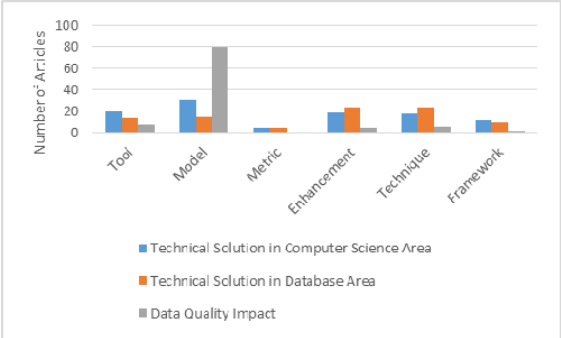

Figure 8: Distribution between articles contribution and data quality research topic

4.4 RQ. 4 What Kind of Research Methods Being Used in Data Quality Research?

Data quality research covered wide topics of discussion in various domains as addressed in RQ.1. On top of that, a number of research methods and techniques have been adopted to suit the needs of the problem being solved in data quality research articles. Table 8 summarized data quality research methods according to data quality research topics.

Table 8: List of research method in data quality research topic

Research Methods	Data Quality Impact	Technical Solution in Computer Science Area	Technical Solution in Database Area
Empirical	/	/	
Qualitative	/	/	/
Case Study	/	/	/
Theory and Formal Proofs	/		
Econometric	/		
Quantitative	/	/	
Mathematical Modelling	/		/
Statistical Analysis	/	/	/
Experimental	/	/	/
Design Science	/	/	
Action Research	/	/	
Ethnography	/		
Delphi Method		/	

In order to give better understanding to the readers, we plotted the percentage value of the technique being used according to data quality research topics in Figure 9. The percentage value is based on the total number of articles in each research topics.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Qualitative, case study, statistical analysis and experimental research methods are common in data quality research that discussed data quality impact, technical solution in computer science area and technical solution in database area. Qualitative research methods has been used to gathered in-depth knowledge about human behavior or phenomenon via qualitative methods such as interviews, focus group or observation and to analyze factors that affect data quality within the organization [12]. In data quality research, qualitative methods are important to gathered knowledge from data users and stakeholders [25] especially when knowledge about data are inadequate. However, the generalization of the resulted studies is highly depended on the participant selection methods being used. Factors such as size of participants and participants' background is important. Whereas, case study is used to analyze data quality problems faced in real environment by the specific organization, group or individual. For example, in [20], case study has been used to investigate data quality problems in multinational manufacturer in China. However, case study limited the generalization of the findings and further implementation of the proposed solution required more attention [20], [42], [77].

The abovementioned limitation, nevertheless, does not limit the knowledge acquisition of data quality problems, issues and challenges from the conducted case study [78]. On the other hand, case study has also been used to evaluate and validate proposed solution in data quality research. For example, in [79], a case study has been used to validate ontology-based data quality framework for data stream application. The validation of proposed solution using case study enabled real-life exploration on limitation and constraint to implement proposed solution. Another common method used in data quality research is statistical analysis. Statistical analysis has been used to find correlation between data quality dimensions, to identify data quality problems and also to improve data quality within the organization. In [80], structural equation modelling (SEM) has been used to validate the model of relationship between information quality, trust and risk perceptions. Statistical analysis also has been used in experimental research to evaluate effects. For example, in [81], analysis of variance (ANOVA) has been adopted to analyze effect of problem complexity in data mining classification algorithms towards data quality.

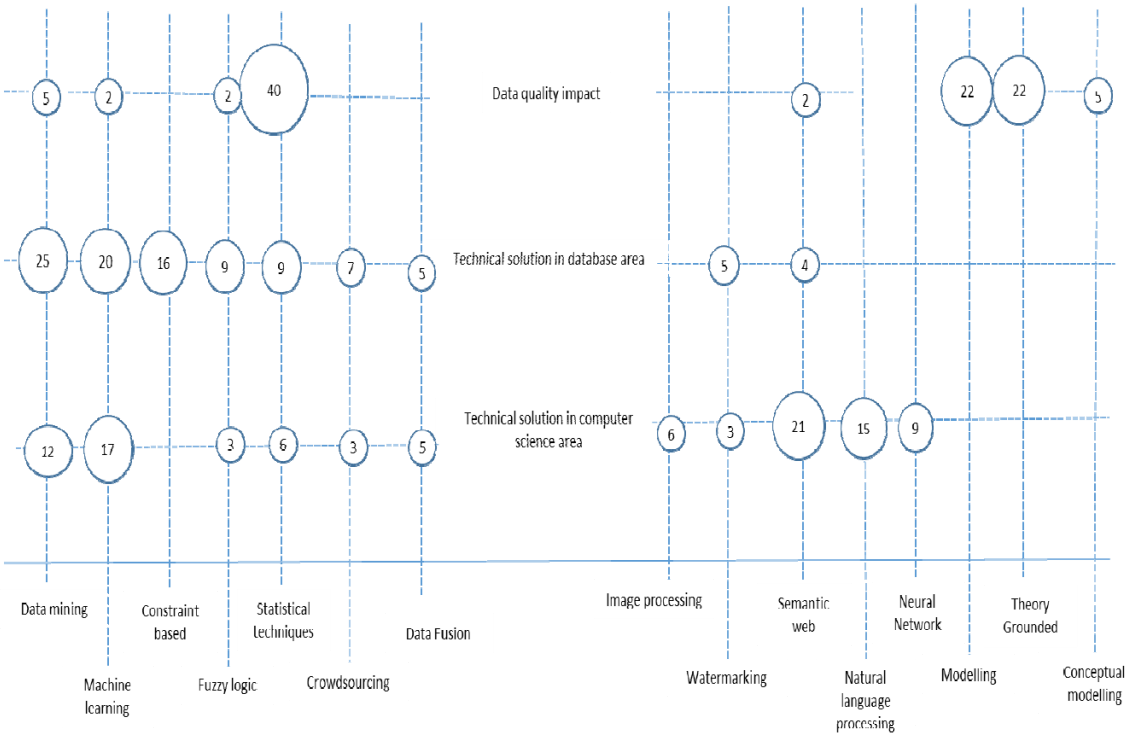

Figure 9: Percentage of techniques used in data quality research according to research topics

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Other research methods such as empirical, quantitative, design science and action research are used in data quality research that discussed data quality impact and technical solution in computer science area. Empirical research methods gathered assessment of contextual data quality problem such as relevancy, timeliness and completeness [3] are needed. Iterative communication with users helps to define data quality problem and solution refinement.

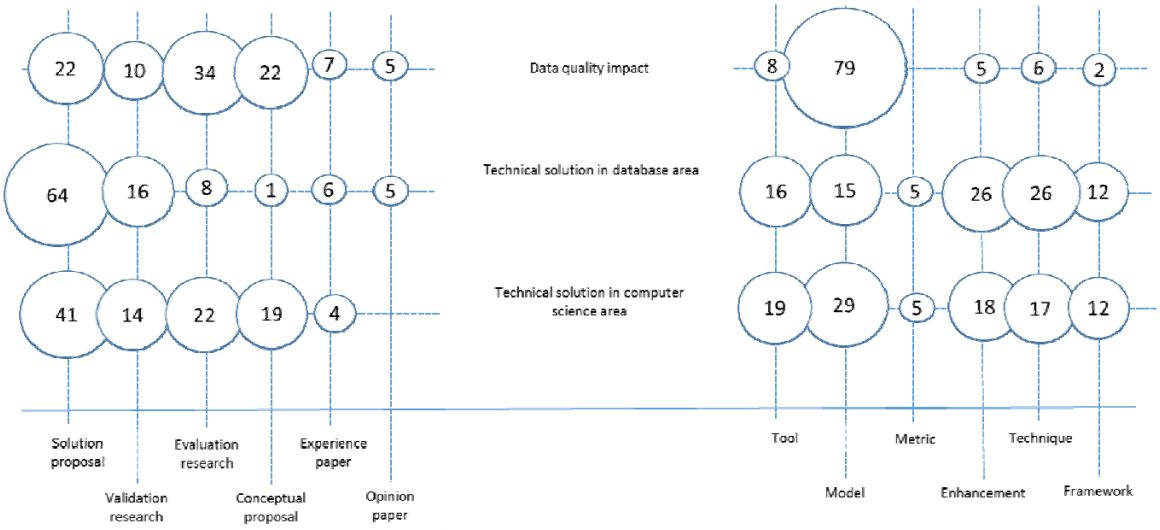

Figure 10: Percentage of research type and type of contribution according to research topic

empirical evidence using observation technique such as experiments, case study, survey and search engine query to gain knowledge about the studied problem. For example, in [82], empirical study to evaluate online health information quality has been conducted using data gathered from search engine query result. Quantitative research in data quality analyzed quantifiable data using statistical analysis and mathematical modelling to identify data quality problems and to probe causes of data quality problems. Quantifiable data in data quality research usually collected from questionnaire and web survey. In [60], a questionnaire has been designed to evaluate data quality dimensions using score. Quantifiable data gathered from the questionnaire is then used to construct a radar plot which explained scores for each data quality dimensions. On the other hands, design science has been adopted in [83] to build quality model for semantic IS standards. In this approach, the model has been evaluated and improved using survey results in repeated cycles. Design science research approach involved rigorous processes with users' evaluation to improve proposed solution [84]. Action research has been adopted in [85] to design and evaluated a process for Total Information Risk Management (TIRM). Three research phase including process design, process testing and process refinement in real environment and lastly, the application of the process in a case study. Both research methods are suitable for data quality research especially when

Whereas, theory and formal proofs, econometric, mathematical modelling and ethnography has been used in investigating data quality impact. For example in [86], an error dominance theory has been proposed to explain data accuracy effects in decision making. Mathematical techniques have been used to provide proof in the proposed theory. Whereas, econometric approach has been used in [26] to assess the effect of data quality management towards economic outcomes such as utility, cost and net-benefit. Econometric approach makes used statistical and mathematical techniques to answer questions regarding economic. Knowledge of the relationship between data quality and economic value are important to the organization in determining suitable action to be taken to manage data quality and to minimize impact to the organization cost. In [21], mathematical modelling has been used to proof the proposed guidelines for organizational data quality policies setup. Mathematical modelling is one of the methods available in quantitative research approach. Research in data quality also adopted ethnography approach to study behavior in specific environment or context. For example, in [87], ethnography approach has been used to analyze issues in forming master data management (MDM) initiatives.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A small number of articles that discussed technical solution in computer science area adopted Delphi method during the validation process. In [88], a Delphi method has been used to validate information quality dimensions ranking. Delphi method is a group communication process involving experts and included multiple iterations of expert involvement and feedback [89]. In data quality research, user and domain expert feedback are important as expectation regarding same data quality dimensions could be differed.

Looking further, computer science techniques such as data mining, machine learning and fuzzy logic were mostly used to assess and improved data quality in the articles that proposed technical solution in computer science and database area. Whereas, constraint based technique such as currency constraint and conditional functional dependencies [90], [91] was adopted only in research that discussed technical solution in database area. Constraint based technique were mostly utilized in database cleaning. For example, constraint such as relaxed functional dependencies (RFDS) has been used to identified missing values and data redundancy [92]. It is clearly shown in Figure 9, twenty-five percent of articles that discussed technical solution in database area and twelve percent of articles that discussed technical solution in computer science area adopted data mining technique. Data mining technique unveiled important information from dataset especially during database data cleaning using techniques such as blocking strategy [93], [94] , affinity propagation and affinity scores [95], [96] and association rule mining [97]. Nevertheless, research in [81], [98], [99] adopted data mining technique to assess data quality in software engineering and financial data. We also found that, twenty percent of articles discussed technical solution in database area and seventeen percent articles that discussed technical solution in computer science area adopted machine learning technique for database cleansing [100] and entity resolution [49]. Additionally, machine learning techniques such as Bayesian network [43] and hidden Markov model [101] also been used for database cleansing. Machine leaning also used to assess information quality in web [73]. In [102], Support Vector Machine (SVM) was used to categorized online product review according to review quality. Furthermore, machine learning technique such as Bayesian inference was used in [74] for sensor data cleansing. Machine learning technique in data quality allowed automated data quality management which reduced human intervention for result verification. However, the effectiveness of machine learning technique is depended to the classification and learning technique used.

Within 2010-2016 we could see new techniques such as crowdsourcing and semantic web technologies being used to improve data quality in database and generally in computer sciences domain such as sensors, web and information system. Knowledge gathered from crowdsourcing technique used to improve data quality in data warehouse [65], [103]. Whereas, crowdsourcing also been proposed in [104] as a method to improve information quality. However, usage of crowdsourcing technique has attracted researcher to assess and measure the quality of knowledge gathered. Seven percent of the articles that discussed technical solution in database area and three percent of articles from technical solution in computer science area used crowdsourcing technique. Semantic web technology such as ontology and Resource Description Framework (RDF) enabled automated reasoning during database cleansing as proposed in [44], [63], [105]. In sensors, ontology adoption allowed domain expert to provide knowledge for effective outlier detection process [64]. Percentage of articles that employed semantic web technology is higher compared to crowdsourcing.

Twenty-one percent of the articles that discussed technical solution in computer science area adopted semantic web technology in their research. The percentage is higher than other techniques such as machine learning, natural language processing, data mining, neural network, statistical techniques, image processing, data fusion, fuzzy logic, crowdsourcing and watermarking. Natural language processing is important in assessing the quality of unstructured data such as names and online review. For example in [106], sentence level LDA (SLDA) has been adopted to address inconsistency issue in online review data. Another example is the adoption of phonetic and string matching technique in [107]. Natural language processing technique is used in this research to suggest possible substitute for misspelled name.

Topics on Semantic web adoption is the lowest in articles that discussed technical solution in database area. Data mining technique, as discussed before, were mostly adopted by the articles in this research topic. Followed by machine learning,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

constraint-based, fuzzy logic, statistical technique, crowdsourcing, data fusion and watermarking.

As articles that discussed data quality impact does not focus on technical solution, high percentages in statistical techniques are expected. Yet, other techniques such as modelling and theory grounded were also used in forty-four percent of the articles that discussed data quality impact. In assessing data quality impact, model such as Information System (IS) success model, technology acceptance model (TAM), decision support model, elaboration likelihood model (ELM), utility model and Toulmin's model of argumentation has been discussed and adopted.

Research in data quality impact topic also discussed and employed existing theories such as culture dimension theory, resource-based theory, management theory, unconscious thought theory, information processing theory, flow theory, contingent resource-based theory, theory of human conduct and game theory.

A small percentage of articles adopted other techniques such as data fusion in [72], [108], image processing in [109], watermarking in [58], and neural network in [110] to assess and improve data quality in various computer science area.

4.5 Mapping and Available Research Gap

In this section, we presented and discussed the mapping of findings in RQ.1, RQ.2, RQ.3 and RQ.4 (refer to Figure 9 and Figure 10) to illustrated current landscape in data quality research. During this discussion, we highlighted several research gap for potential future research in data quality. We also presented 10 highly cited data quality articles in Scopus and Google Scholar in Table 9. From the citation numbers, most of the articles proposed a novel solution or an essential extension of an existing technique in database area compared to other category. We also see that research in data quality metrics are not listed in the top 10.

Research in data quality produced highest number of articles that discussed technical solution in computer science area. In this research topic, most of the articles are categorized in solution proposal research type. Citation rate provided in Table 9 supported our claimed as 53 citation (ranked third) recorded for research in [102]. Suprisingly, the highest contribution within this research topic is model compared to technique as third highest type of contribution. Technical solution in computer science area comprised of wide subcategories including web and sensors. Thus, we believe that the resulted numbers were due to the needs in understanding challenges, effects and comparisons between available techniques. From Figure 9, we could see high percentage of articles using semantic web technologies such as RDF and ontology in this research topic. Highest adoption percentage in semantic web technologies remarked the needs of repositories contained domain expert knowledge that provide automated reasoning for data quality management.

Highest number in model contribution that discussed data quality impact can be related to the high percentage of research articles in the same research topic that used statistical techniques, modelling and theory grounded in Figure 9. The techniques mentioned before helped researcher to find correlation between data quality dimensions. In Table 9, highly cited articles within this category contributed model. The citation rate denote data quality research community attention. Most of articles within data quality impact research topics were categorized as evaluation research type.

Lowest publication of data quality research topic is technical solution in database area. Within this research topic, most of the published articles categorized as solution proposal research type. Highest number of articles in this research topic contributed technique and enhancement. Machine learning and data mining technique are mostly adopted in this research topic as presented in Figure 9. Machine learning adoption were mostly motivated by the needs to minimize human intervention during data quality assessment and improvement process.

From the mapping, obviously we can see contribution disparity between data quality metric and model. There is also inbalanced number between articles that contribute model and other contribution such as tool, enhancement, technique and metric in data quality impact research topic. Moreover, less consideration is given by the research community to framework contribution in this topic. Within this research topic, there is potential future research to assess data quality impact using crowdsourcing and natural langguage processing technique.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Less consideration is given in validation research type, evaluation research type and conceptual proposal research type within technical solution in database area research topic. Besides metric and framework, there is still potential future research in tool and model contribution. As unstructured data becoming important recently, technique such as crowdsourcing, semantic web technologies and natural langguage processing has huge potential to be explored in future research.

Nevertheless, within technical solution in computer science area research topic, validation research and conceptual proposal research type has potential for future research. Framework and data quality metric for data fusion and crowdsourcing technique would have future potential development.

4.6 Limitations of This Study

During this study, 374 data quality research articles published between 2010 and 2016 has been selected and reviewed. The selection was done based on the ability of the article to answer at least one research question and fulfilled our inclusion criteria. We also considered four journals, three conference proceedings and one magazine as articles sources in order to include all important data quality research in this study. However, there is still a possibility to miss out important articles especially articles that is not published in English and published in not selected journal, conference proceedings or magazine.

We did not include quality assessment in our article selection process. This may bias the number of selected articles but we believe by collecting as much articles as we can, could help us to get a wide view of data quality research. We excluded any unrelated articles and minimize bias by imposed inclusion and exclusion criteria. For this reason, we believe that quality assessment during articles selection is not necessary.

5. CONCLUSIONS

This paper applied systematic review study to explain the landscape in data quality research. Categorization and mapping is used in this review to highlight available research gap in data quality. We supported our review with 374 published data quality research articles. The result of this study indicates a significant trend in data quality research publication. On average, fifty-four research articles related to data quality were published every year. This number shows the importance of data quality research in various research areas such as online users, database, web information, sensors and big data. This study also indicates the following points: (i) Almost half of the articles included in this review proposed a novel solution or an essential extension of an existing data quality technique, (ii) Most of the selected research articles belongs to the model type in the contribution category, (iii) Obvious contribution disparity happen between contribution in metric type and model type category.

Additionally, our mapping suggested that most of the research articles that discussed technical solution in computer sciences belonging to the model type in contribution category. This shows that data quality research in this area were still discussing the challenges, effects and comparisons between available techniques. We also found that research articles that discussed technical solution in database area were mostly belonging to the technique and enhancement type in contribution category. Research articles in this area were dominated by machine learning and data mining technique. We considered that our aim has been achieved and each research questions has been answered.

Table 9: Overview Of High Cited Articles According To Scopus Citation Rate

Rank	Paper	Source	Year	Times cited		Topics	Research Type	Contribution Type
				SC	GS			
1	[35]	DSS	2013	86	169	A	Solution proposal	Model
2	[15]	DSS	2012	64	139	A	Solution proposal	Model
3	[102]	DSS	2011	53	127	C	Solution proposal	Technique
4	[111]	VLDB	2010	49	104	B	Solution proposal	Enhancement
5	[45]	VLDB	2011	46	99	B	Solution proposal	Framework
6	[37]	DSS	2013	45	113	A	Solution proposal	Framework
7	[112]	VLDB	2010	38	55	B	Solution proposal	Technique
8	[32]	IM	2010	33	88	A	Evaluation research	Model

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

9	[46]	SIGMOD	2011	34	78	B	Solution proposal	Framework
10	[81]	JDIQ	2011	34	49	C	Conceptual proposals	Model

*** Topic A : Data quality impact**
*** Topic B : Technical solution in database area**
*** Topic C : Technical solution in computer science area**

*** SC : Scopus**
*** GS : Google Scholar**

ACKNOWLEDGEMENT

The work reported here is funded by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS 03-12-10-999FR). Mygrants Reference Code is FRGS/2/2014/ICT07/UPM/02/1. This support is gratefully acknowledge.

REFERENCES:

[1] Y. W. Lee and D. M. Strong, "Knowing-Why About Data Processes and Data Quality," *Journal of Management Information Systems*, vol. 20, no. 3, pp. 13–39, Jan. 2003.

[2] A. V. Levitin and T. C. Redman, "Data as a resource: properties, implications, and prescriptions," *Sloan Management Review*, vol. 40, pp. 89–101, 1998.

[3] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data Quality in Context," *Communications of the ACM*, vol. 40, no. 5, pp. 103–110, May 1997.

[4] R. Wang and D. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, vol. 12, no. 4, pp. 5–33, 1996.

[5] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM*, vol. 41, no. 2, pp. 58–65, Feb. 1998.

[6] Y. Xiao, L. Y. Y. Lu, J. S. Liu, and Z. Zhou, "Knowledge diffusion path analysis of data quality literature: A main path analysis," *Journal of Informetrics*, vol. 8, no. 3, pp. 594–605, Jul. 2014.

[7] S. Sadiq, N. Yeganeh, and M. Indulska, "20 years of data quality research: themes, trends and synergies," in *Proceedings of the Twenty-Second Australasian Database Conference*, 2011, vol. 115, pp. 153–162.

[8] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Keele University and University of Durham, 2007.

[9] S. Sadiq, "Prologue: Research and Practice in Data QualityManagement," in *Handbook of Data Quality*, S. Sadiq, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1–11.

[10] 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*, vol. 7, no. 4, pp. 623–640, 1995.

[11] C. Batini and M. Scannapieca, *Data Quality*. Springer Berlin Heidelberg, 2006.

[12] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, "Overview and Framework for Data and Information Quality Research," *Journal of Data and Information Quality*, vol. 1, no. 1, pp. 1–22, Jun. 2009.

[13] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: A proposal and a discussion," *Requirements Engineering*, vol. 11, no. 1, pp. 102–107, 2006.

[14] S. Salehi, A. Selamat, and H. Fujita, "Systematic mapping study on granular computing," *Knowledge-Based Systems*, vol. 80, pp. 78–97, May 2015.

[15] A. Popovič, R. Hackney, P. S. Coelho, and J. Jaklič, "Towards business intelligence systems success: Effects of maturity and culture on analytical decision making," *Decision Support Systems*, vol. 54, no. 1, pp. 729–739, Dec. 2012.

[16] K. Hartl and O. Jacob, "The Role of Data Quality in Business Intelligence – An empirical study in German medium-sized and large companies," in *ICIQ 2016*, 2016, p. 4:1-4:10.

[17] Ö. Işık, M. C. Jones, and A. Sidorova, "Business intelligence success: The roles of BI capabilities and decision environments," *Information & Management*, vol. 50, no. 1, pp. 13–23, Jan. 2013.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[18] A. Haug, J. Stentoft Arlbjørn, F. Zachariassen, and J. Schlichter, "Master data quality barriers: an empirical investigation," *Industrial Management & Data Systems*, vol. 113, no. 2, pp. 234–249, Mar. 2013.

[19] K. S. Ryu, J. S. Park, and J. H. Park, "A Data Quality Management Maturity Model," *ETRI Journal*, vol. 28, no. 2, pp. 191–204, Apr. 2006.

[20] L. Cao and H. Zhu, "Normal Accidents: Data Quality Problems in ERP-Enabled Manufacturing," *Journal of Data and Information Quality*, vol. 4, no. 3, pp. 1–26, May 2013.

[21] V. C. Storey, R. M. Dewan, and M. Freimer, "Data quality: Setting organizational policies," *Decision Support Systems*, vol. 54, no. 1, pp. 434–442, Dec. 2012.

[22] H.-T. Moges, V. Van Vlasselaer, W. Lemahieu, and B. Baesens, "Determining the use of data quality metadata (DQM) for decision making purposes and its impact on decision outcomes — An exploratory study," *Decision Support Systems*, vol. 83, pp. 32–46, Mar. 2016.

[23] G. Shankaranarayanan and B. Zhu, "SPIDEQ – A Prototype for Decision Making with Quality Metadata," in *ICIQ 2016*, 2016, p. 20:1-20:12.

[24] J. W. Kim and J.-H. Lim, "IT investments disclosure, information quality, and factors influencing managers' choices," *Information & Management*, vol. 48, no. 2–3, pp. 114–123, Mar. 2011.

[25] A.-C. Le Dû and F. de Corbière, "IQ as an Enabler of the Green and Collaborative Supply Chain," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, pp. 1–14.

[26] A. Even, G. Shankaranarayanan, and P. D. Berger, "Evaluating a model for cost-effective data quality management in a real-world CRM setting," *Decision Support Systems*, vol. 50, no. 1, pp. 152–163, Dec. 2010.

[27] N. R. Joglekar, E. G. Anderson, and G. Shankaranarayanan, "Accuracy of aggregate data in distributed project settings," *Journal of Data and Information Quality*, vol. 4, no. 3, pp. 1–22, May 2013.

[28] U. ul Hassan, S. O'Riarin, and E. Curry, "Towards Expertise Modelling for Routing Data Cleaning Tasks Within a Community of Knowledge Workers," in *Proceeding of the 17th International Conference on Information Quality (ICIQ) 2012*, 2012, pp. 58–69.

[29] F. Xu, H. Y. Zhang, and W. Huang, "Do CDOs Matter ? Assessing the Value of CDO Presence in Firm Performance," in *Proceedings of the 19th International Conference on Information Quality (ICIQ-2014)*, 2014, pp. 164–170.

[30] T. T. Moores, "Towards an integrated model of IT acceptance in healthcare," *Decision Support Systems*, vol. 53, no. 3, pp. 507–516, Jun. 2012.

[31] I. Askira Gelman, "GIGO or not GIGO," *Journal of Data and Information Quality*, vol. 2, no. 2, pp. 1–27, Feb. 2011.

[32] C.-W. Chen, "Impact of quality antecedents on taxpayer satisfaction with online tax-filing systems—An empirical study," *Information & Management*, vol. 47, no. 5–6, pp. 308–315, Aug. 2010.

[33] M. Aljukhadar, S. Senecal, and J. Nantel, "Is more always better? Investigating the task-technology fit theory in an online user context," *Information & Management*, vol. 51, no. 4, pp. 391–397, Jun. 2014.

[34] J. V. Chen, B. Su, and A. E. Widjaja, "Facebook C2C social commerce: A study of online impulse buying," *Decision Support Systems*, vol. 83, pp. 57–69, Mar. 2016.

[35] T. Zhou, "An empirical examination of continuance intention of mobile payment services," *Decision Support Systems*, vol. 54, no. 2, pp. 1085–1091, Jan. 2013.

[36] Z.-J. Chen, D. Vogel, and Z.-H. Wang, "How to satisfy citizens? Using mobile government to reengineer fair government processes," *Decision Support Systems*, vol. 82, pp. 47–57, Feb. 2016.

[37] Y. Zheng, K. Zhao, and A. Stylianou, "The impacts of information quality and system quality on users' continuance intention in information-exchange virtual communities: An empirical investigation," *Decision Support Systems*, vol. 56, no. 1, pp. 513–524, Dec. 2013.

[38] Y. Lu and D. Yang, "Information exchange in virtual communities under extreme disaster conditions," *Decision Support Systems*, vol. 50, no. 2, pp. 529–538, Jan. 2011.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[39] S. J. Barnes and R. Vidgen, "User acceptance and corporate intranet quality: An evaluation with iQual," *Information & Management*, vol. 49, no. 3–4, pp. 164–170, May 2012.

[40] P. Woodall, A. Borek, J. Gao, M. Oberhofer, and Andy Koronios, "An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics," in *Proceedings of the 19th International Conference on Information Quality (ICIQ-2014)*, 2014, pp. 24–33.

[41] M. Helfert and Mouzhi Ge, "Big Data Quality - Towards an Explanation Model in a Smart City Context," in *ICIQ 2016*, 2016, p. 2:1-2:8.

[42] D. Becker, T. D. King, B. Mcmullen, and D. A. Fahsi, "Big Data Quality Case Study Preliminary Findings: Hyperspectral Imaging (HSI) Using the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," in *Proceedings of the 18th International Conference on Information Quality (ICIQ)*, 2013.

[43] K. Chen, H. Chen, N. Conway, J. M. Hellerstein, and T. S. Parikh, "Usher: Improving Data Quality with Dynamic Forms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 8, pp. 1138–1153, Aug. 2011.

[44] O. Curé, "Improving the Data Quality of Drug Databases using Conditional Dependencies and Ontologies," *Journal of Data and Information Quality*, vol. 4, no. 1, pp. 1–21, Oct. 2012.

[45] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas, "Guided data repair," *Proceedings of the VLDB Endowment*, vol. 4, no. 5, pp. 279–289, Feb. 2011.

[46] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu, "Interaction between record matching and data repairing," in *Proceedings of the 2011 international conference on Management of data - SIGMOD '11*, 2011, vol. 1, no. 1, p. 469.

[47] C. J. Zhang, L. Chen, H. V. Jagadish, and C. C. Cao, "Reducing uncertainty of schema matching via crowdsourcing," *Proceedings of the VLDB Endowment*, vol. 6, no. 9, pp. 757–768, Jul. 2013.

[48] S. Destercke, P. Buche, and B. Charnomordic, "Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 92–105, Jan. 2013.

[49] R. Nuray-Turan, D. V Kalashnikov, and S. Mehrotra, "Adaptive Connection Strength Models for Relationship-Based Entity Resolution," *Journal of Data and Information Quality*, vol. 4, no. 2, pp. 1–22, Mar. 2013.

[50] M. Yakout, M. J. Atallah, and A. Elmagarmid, "Efficient and Practical Approach for Private Record Linkage," *Journal of Data and Information Quality*, vol. 3, no. 3, pp. 1–28, Aug. 2012.

[51] H. Köpcke, A. Thor, and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems," in *Proceedings of the VLDB Endowment*, 2010, vol. 3, no. 1–2, pp. 484–493.

[52] A. Gruenheid, X. L. Dong, and D. Srivastava, "Incremental record linkage," in *Proceedings of the VLDB Endowment*, 2014, vol. 7, no. 9, pp. 697–708.

[53] Y.-W. Cheah and B. Plale, "Provenance Quality Assessment Methodology and Framework," *Journal of Data and Information Quality*, vol. 5, no. 3, pp. 1–20, Dec. 2014.

[54] J. Wang, S. Song, X. Zhu, X. Lin, and J. Sun, "Efficient Recovery of Missing Events," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 2943–2957, Nov. 2016.

[55] F. Panse, M. van Keulen, and N. Ritter, "Indeterministic Handling of Uncertain Decisions in Deduplication," *Journal of Data and Information Quality*, vol. 4, no. 2, pp. 1–25, Mar. 2013.

[56] J. Xu, D. V. Kalashnikov, and S. Mehrotra, "Query Aware Determinization of Uncertain Objects," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 1, pp. 207–221, Jan. 2015.

[57] S. Chen, X. L. Dong, L. V. S. Lakshmanan, and D. Srivastava, "We challenge you to certify your updates," in *Proceedings of the 2011 international conference on Management of data - SIGMOD '11*, 2011, p. 481.

[58] M. Kamran, S. Suhail, and M. Farooq, "A Robust, Distortion Minimizing Technique for Watermarking Relational Databases Using Once-for-All Usability Constraints," *IEEE Transactions on Knowledge and Data*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Engineering, vol. 25, no. 12, pp. 2694–2707, Dec. 2013.

[59] H.-T. Moges, K. Dejaeger, W. Lemahieu, and B. Baesens, "A multidimensional analysis of data quality for credit risk management: New insights and challenges," *Information & Management*, vol. 50, no. 1, pp. 43–58, Jan. 2013.

[60] K. Dejaeger, B. Hamers, J. Poelmans, and B. Baesens, "A Novel Approach to the Evaluation and Improvement of Data Quality in the Financial Sector," *Proceedings of the 15th International Conference on Information Quality*, 2010.

[61] B. Heinrich and M. Klier, "Metric-based data quality assessment — Developing and evaluating a probability-based currency metric," *Decision Support Systems*, vol. 72, pp. 82–96, Apr. 2015.

[62] P. Woodall and A. K. Parlikad, "A Hybrid Approach to Assessing Data Quality," in *Proceedings of the 15th International Conference on Information Quality*, 2010.

[63] J. Du and L. Zhou, "Improving financial data quality using ontologies," *Decision Support Systems*, vol. 54, no. 1, pp. 76–86, Dec. 2012.

[64] L. Gao, M. Bruenig, and J. Hunter, "Semantic-based Detection of Segment Outliers and Unusual Events for Wireless Sensor Networks," in *Proceedings of the 18th International Conference on Information Quality (ICIQ-13)*, 2013, pp. 102–119.

[65] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye, "KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, 2015, pp. 1247–1261.

[66] C. Baillie, P. Edwards, and E. Pignotti, "QUAL: A Provenance-Aware Quality Model," *Journal of Data and Information Quality*, vol. 5, no. 3, pp. 1–22, Mar. 2015.

[67] N. Martin, A. Poulovassilis, and J. Wang, "A Methodology and Architecture Embedding Quality Assessment in Data Integration," *Journal of Data and Information Quality*, vol. 4, no. 4, pp. 1–40, May 2014.

[68] S. Geisler, S. Weber, and C. Quix, "Ontology-Based Data Quality Framework for Data Stream Applications," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, pp. 145–159.

[69] M. C. Tremblay, A. R. Hevner, and D. J. Berndt, "Design of an information volatility measure for health care decision making," *Decision Support Systems*, vol. 52, no. 2, pp. 331–341, Jan. 2012.

[70] J. O'Donoghue and J. Herbert, "Data Management within mHealth Environments: Patient Sensors, Mobile Devices, and Databases," *Journal of Data and Information Quality*, vol. 4, no. 1, pp. 1–20, Oct. 2012.

[71] A. Marotta and A. Delgado, "Data Quality Management in Web Warehouses using BPM," in *ICIQ 2016*, 2016, p. 18:1-18:10.

[72] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, "Truth Finding on the Deep Web: Is the Problem Solved?," *Proceedings of the VLDB Endowment*, vol. 6, no. 2, pp. 97–108, Dec. 2012.

[73] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Automatic Assessment of Document Quality in Web Collaborative Digital Libraries," *Journal of Data and Information Quality*, vol. 2, no. 3, pp. 1–30, Dec. 2011.

[74] W.-S. Ku, H. Chen, H. Wang, and M.-T. Sun, "A Bayesian Inference-Based Framework for RFID Data Cleansing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp. 2177–2191, Oct. 2013.

[75] I. Caballero, I. Bermejo, L. Parody, G. Teresa, Mª López, R. M.Gasca, and Mario Piattini, "I8K: An Implementation of ISO 8000-1x0," in *Proceedings of the 18th International Conference on Information Quality (ICIQ)*, 2013.

[76] H. Zhu and H. Wu, "Assessing the quality of large-scale data standards: A case of XBRL GAAP Taxonomy," *Decision Support Systems*, vol. 59, no. 1, pp. 351–360, 2014.

[77] B. Otto and A. Schmidt, "Enterprise master data architecture: Design decisions and options," in *15th International Conference on Information Quality (ICIQ 2010), Little Rock*, 2010.

[78] B. Flyvbjerg, "Five Misunderstandings About Case-Study Research," *Qualitative Inquiry*, vol. 12, no. 2, pp. 219–245, Apr. 2006.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[79] S. Geisler, S. Weber, and C. Quix, "Ontology - Based Data Quality Framework for Data Stream Applications," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, pp. 145–159.

[80] A. I. Nicolaou, M. Ibrahim, and E. Van Heck, "Information quality, trust, and risk perceptions in electronic data exchanges," *Decision Support Systems*, vol. 54, no. 2, pp. 986–996, 2013.

[81] R. Blake and P. Mangiameli, "The Effects and Interactions of Data Quality and Problem Complexity on Classification," *Journal of Data and Information Quality*, vol. 2, no. 2, pp. 1–28, Feb. 2011.

[82] B. Kitchens, C. a. Harle, and S. Li, "Quality of health-related online search results," *Decision Support Systems*, vol. 57, no. 1, pp. 454–462, 2014.

[83] E. Folmer and J. Van Soest, "Towards a quality model for semantic IS standards," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, pp. 130–144.

[84] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, Dec. 2007.

[85] A. Borek, A. K. Parlikad, and P. Woodall, "Towards a Process For Total Information Risk Management," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, no. 1, pp. 477–491.

[86] I. A. Gelman, "Setting priorities for data accuracy improvements in satisficing decision-making scenarios: A guiding theory," *Decision Support Systems*, vol. 48, no. 4, pp. 507–520, Mar. 2010.

[87] R. Vilminko-heikkinen and S. Pekkola, "Organizational Issues in Establishing Master Data Management Function," in *Proceedings of the 17th International Conference on Information Quality (ICIQ-12)*, 2012, no. Mdm, pp. 1–13.

[88] Y. Yiliyasi and D. Berleant, "World Oil Reserves Data: Information Quality Assessment Analysis," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, pp. 534–547.

[89] C.-C. Hsu and B. A. Sandford, "The Delphi Technique: Making Sense Of Consensus," *Practical Assessment Research & Evaluation*, vol. 12, no. 10, Oct. 2007.

[90] W. Fan, F. Geerts, N. Tang, and W. Yu, "Conflict resolution with data currency and consistency," *Journal of Data and Information Quality*, vol. 5, no. 1–2, pp. 1–37, Sep. 2014.

[91] G. Beskales, I. F. Ilyas, and L. Golab, "Sampling the repairs of functional dependency violations under hard constraints," *Proceedings of the VLDB Endowment*, vol. 3, no. 1–2, pp. 197–207, Sep. 2010.

[92] L. Caruccio, V. Deufemia, and G. Polese, "Relaxed Functional Dependencies - A Survey of Approaches," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 147–165, Jan. 2016.

[93] M. T. de Mello, G. C. Frainer, J. G. C. de Souza, and L. C. R. Junior, "Towards a High Performance Merge Solution for Large-scale Dataset," in *International Conference of Information Quality*, 2010.

[94] F. Panse, W. Wingerath, S. Friedrich, and N. Ritter, "Key-Based Blocking of Duplicates in Entity-Independent Probabilistic Data," in *Proceedings of the 17th International Conference on Information Quality (ICIQ-12)*, 2012, pp. 278–296.

[95] X. Fan, J. Wang, X. Pu, L. Zhou, and B. Lv, "On Graph-Based Name Disambiguation," *Journal of Data and Information Quality*, vol. 2, no. 2, pp. 1–23, 2011.

[96] F. Kobayashi and J. R. Talburt, "Improving the Quality of Entity Resolution for School Enrollment Data through Affinity Scores," in *Proceedings of the 19th International Conference on Information Quality (ICIQ-2014)*, 2014, pp. 69–80.

[97] H. Müller, J.-C. Freytag, and U. Leser, "Improving data quality by source analysis," *Journal of Data and Information Quality*, vol. 2, no. 4, pp. 1–38, Feb. 2012.

[98] "Credit Risk Analysis of Taiwan's Financial Sector with Data Mining Method," pp. 272–276, 2014.

[99] L. Yu, S. Ramaswamy, and A. Nair, "Using Bug Report as a Software Quality Measure," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, pp. 277–286.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[100] M. Yakout, A. K. Elmagarmid, J. Neville, and M. Ouzzani, "GDR: A System for Guided Data Repair," *Proceedings of the 2010 international conference on Management of data - SIGMOD '10*, p. 1223, 2010.

[101] J. Cao, X. Diao, N. Zhang, and T. Wang, "An Approach Using Hidden Markov Model for Estimating and Replacing Missing Categorical Data," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, no. 427, pp. 427–434.

[102] C. C. Chen and Y.-D. Tseng, "Quality evaluation of product reviews using an information quality framework," *Decision Support Systems*, vol. 50, no. 4, pp. 755–768, Mar. 2011.

[103] H. Park, H. Garcia-Molina, R. Pang, N. Polyzotis, A. Parameswaran, and J. Widom, "Deco: a system for declarative crowdsourcing," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 1990–1993, 2012.

[104] P. Wichmann, A. Borek, R. Kern, P. Woodall, A. K. Parlikad, and G. Satzger, "Exploring the 'Crowd' as Enabler of Better Information Quality," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, pp. 302–313.

[105] C. Civili, M. Console, G. De Giacomo, D. Lembo, M. Lenzerini, L. Lepore, R. Mancini, A. Poggi, R. Rosati, M. Ruzzi, V. Santarelli, and D. F. Savo, "MASTRO STUDIO : Managing Ontology-Based Data Access applications," in *Proceedings of the VLDB Endowment*, 2013, no. 6 (12), pp. 1314–1317.

[106] F. Liu and N. Wu, "Your Words Count : Investigating Inconsistency in Reviews from Multiple Online Sources Via Topic Modeling," in *Proceedings of the 19th International Conference on Information Quality (ICIQ-2014)*, 2014, pp. 204–215.

[107] C. Varol and C. Bayrak, "Hybrid Matching Algorithm for Personal Names," *Journal of Data and Information Quality*, vol. 3, no. 4, pp. 1–18, 2012.

[108] X. L. Dong, B. Saha, and D. Srivastava, "Less is More: Selecting Sources Wisely for Integration," *Proceedings of the VLDB Endowment*, vol. 6, no. 2, pp. 37–48, Dec. 2012.

[109] M. L. Jensen, J. K. Burgoon, and J. F. Nunamaker, "Judging the Credibility of Information Gathered from Face-to-Face Interactions," *Journal of Data and Information Quality*, vol. 2, no. 1, pp. 1–20, Jul. 2010.

[110] S. Kaya, M. Milanova, J. Talburt, B. Tsou, and M. Altynova, "Subjective Image Quality Prediction based on Neural Network," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, 2011, pp. 259–266.

[111] T. Neumann and G. Weikum, "x-RDF-3X: Fast Querying, High Update Rates, and Cinsistency for RDF Databases," *Proceedings of the VLDB Endowment*, vol. 3, no. 1–2, pp. 256–263, Sep. 2010.

[112] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu, "Towards certain fixes with editing rules and master data," *Proceedings of the VLDB Endowment*, vol. 3, no. 1, pp. 172–184, Apr. 2010.