Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2011 Proceedings

European Conference on Information Systems (ECIS)

Summer 10-6-2011

CROSS-DISCIPLINARY COLLABORATIONS IN DATA QUALITY RESEARCH

Shazia Sadiq

Khodabandehloo Yeganeh

Marta Indulska

CROSS-DISCIPLINARY COLLABORATIONS IN DATA QUALITY RESEARCH

Sadiq, Shazia, University of Queensland, St Lucia, Qld, 4072 Australia, shazia@itee.uq.edu.au

Khodabandehloo Yeganeh, Naiem, University of Queensland, St Lucia, Qld, 4072 Australia, naiem@itee.uq.edu.au

Indulska, Marta, University of Queensland, St Lucia, Qld, 4072 Australia, m.indulska@business.uq.edu.au

Abstract

Data Quality has been the target of research and development for over four decades, and due to its cross-disciplinary nature has been approached by business analysts, solution architects, database experts and statisticians to name a few. As data quality increases in importance and complexity, there is a need to motivate the exploitation of synergies across diverse research communities in order to form holistic solutions that span across its organizational, architectural and computational aspects. As a first step towards bridging gaps between the various research communities, we undertook a comprehensive literature study of data quality research published in the last two decades. In this study we considered a broad range of Information System (IS) and Computer Science (CS) publication outlets. The main aims of the study were to understand the current landscape of data quality research, create better awareness of (lack of) synergies between various research communities, and, subsequently, direct attention towards holistic solutions. In this paper, we present a summary of the findings from the study that outline the overlaps and distinctions between the two communities from various points of view, including publication outlets, topics and themes of research, highly cited or influential contributors and strength and nature of co-authorship networks.

Keywords: Literature Survey, Research Framework, Bibliometrics, Citation Analysis, Data Quality

1 Introduction

Deployment of IT solutions, often following from strategic redirections, upgrades, mergers and acquisitions, is inevitably subjected to an evaluation of return on investment (ROI), which includes evaluation of the costs of sizable installations, as well as the cost of changing the culture and work practice of all involved. It is often observed that the results of such analyses frequently indicate a failure to achieve the expected benefits (Carr 2004). A range of factors contributes to dismal ROIs, including significant factors rooted externally to the technological sophistication of the systems and often residing in the quality of the information the system manages and generates.

The issue of data quality is as old as data itself. However it is now exposed at a much more strategic level e.g. through business intelligence (BI) systems increasing manifold the stakes involved for corporations as well as government agencies. For example, the Detroit terror case triggered an overhaul of the nation-wide watch list system, where lack of data propagation/consistency and issues with data freshness can be observed. The issue is equally important for scientific applications where lack of knowledge about data accuracy, currency or certainty can lead to catastrophic results. For example, the hurricane protection system in New Orleans failed because it was "inadequate and incomplete", having been built disjointedly over several decades using out-dated elevation data (New York Times, June 1, 2006). Further, the proliferation of shared/public data as on the World Wide Web and growth of the web community has increased the risk of poor data quality usage for individuals as well. This is particularly alarming due to the diversity of the web community, where many are unaware of data sources and data credentials. The situation is further complicated by presence of data aggregations and assimilations e.g. through meta-search engines where source attribution and data provenance can be completely hidden from the data consumers.

One can also observe the changing nature of data quality management over the last decade or more. First, there are clear implications that relate to the sheer volume of data produced by organizations today. Second, recent years have seen an increase in the diversity of data. Such diversity refers to structured, unstructured, semi-structured data, and multi-media data such as video, maps, images, etc. Data also has an increasing number of sources. The use of various technologies, for example, sensor devices, medical instrumentation, RFID readers, further increases the amount and diversity of data being collected. More subtle factors also exist - such as the lack of clear alignment between the intention of data creation and its subsequent usage. A prime example of such lack of alignment is the vast amount of data collected from social networks that can then be used, without assessment of quality, as a basis for marketing decisions. Accordingly, a related factor exists that relates to difficulties in defining appropriate data quality metrics.

As these changes occur, traditional approaches and solutions to data management in general, and data quality control specifically, are challenged. There is an evident need to incorporate data quality considerations into the whole data cycle, encompassing managerial/governance as well as technical aspects. Currently, data quality contributions from research and industry appear to originate from three distinct communities: *Business Analysts*, who focus on organizational solutions. That is, the development of data quality objectives for the organization, as well as the development of strategies to establish roles, processes, policies, and standards required to manage and ensure the data quality objectives are met. *Solution Architects*, who work on architectural solutions. That is, the technology landscape required to deploy developed data quality management processes, standards and policies. *Database Experts* and *statisticians*, who contribute to computational solutions. That is, effective and efficient IT tools, and computational techniques, required to meet data quality objectives. Techniques in this regard can include record linkage, lineage and provenance, data uncertainty, semantic integrity constraints, as well as information trust and credibility.

For the research community to adequately respond to the current and changing landscape of data quality challenges, a unified framework for data quality research is needed. Such a framework should acknowledge the central role of data quality in future systems development initiatives and motivate the exploitation of synergies across diverse research communities.

It is unclear if synergies across the three contributing communities have been fully exploited. We argue that a unified framework for data quality management should bring together organizational, architectural and computational approaches proposed from the three communities respectively. As a first step towards bridging gaps between the various research communities, we undertook a comprehensive literature study of data quality research published in the last two decades (Sadiq, Yegenah and Indulska, 2011). In this study we considered a broad range of Information System (IS) and Computer Science (CS) publication (conference and journal) outlets so as to ensure adequate coverage of organizational, architectural and computational contributions. The main aims of the study were to understand the current landscape of data quality research, to create better awareness of (lack of) synergies between various research communities, and, subsequently, to direct attention towards holistic solutions that span across the organizational, architectural and computational aspects (thus requiring collaboration from the relevant research communities).

In this paper, we present a summary of the findings from the study that specifically relate to topical and contributor connectivity. That is, the level and nature of collaboration across IS and CS research communities working on data quality. In the following section we first discuss some related studies. We then present the methodology employed for the study, followed by a discussion of the key results.

2 Related Studies

A number of studies have addressed the issue of defining and analysing the scope of data quality research in the past. Owing to the cross-disciplinary needs of this area, identifying the central themes and topics and correspondingly the associated methodologies has been a challenge. Recent work by (Madnick and Wang, 2009) has presented a framework that characterizes data quality research along the two dimensions of topics and methods thereby providing a means to classify various research works. Previous works have also assisted by developing frameworks through which data quality research could be characterized, including a predecessor framework by the above group, (Wang and Storey, 1995) that analogized data quality processes with product manufacturing processes. Some key research aspects such as data quality standardization, metrics/measurements and policy management emerged from these earlier works.

Other more recent studies have also provided valuable means of classification for data quality research. (Ge and Helfert, 1996) have structured their review of the literature as IQ Assessment, IQ Management and Contextual IQ. (Lima and Macada, 2006) classify the literature between theoretical (conceptual, applied, illustrative) and practical (qualitative, experimental, survey, simulation) aspects. (Further Neely And Cook, 2005) present their classification as a cross-tabulation of Wang's framework (Wang and Storey, 1995) and Juran's original fitness for use factors (Juran, 1962).

The above studies provide various angles through which the body of knowledge can be classified and thus provide an essential means of understanding the core topics of data quality. However, understanding the intellectual corpus of a discipline requires not only an understanding of its core, but also its boundaries (Benbasat, 2003). As the realm of data quality has grown, so has the scope of its reference disciplines. With these factors in mind, we focused our study on understanding the interconnections and synergies across the various communities that contribute to data quality, rather than an identification of its central themes. We argue that addressing the current challenges in data quality warrants such an understanding so synergies would be better exploited and holistic solutions may be developed.

In this paper, we have accordingly attempted to identify the overlaps and segregations within data quality research. Towards this end, we have identified a corpus of publication outlets that spans across the Information Systems (IS) and Computer Science (CS) disciplines. In the next section we present details of our selection and the methodology employed to conduct the analysis.

3 Methodology

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

Table 1 details the list of considered Information Systems and Computer Science publication outlets, and the respective volume of papers, that has been considered in this study. In particular, we have focused on almost the last two decades of conference and journal publications (1990-2009). This span varies for publications that were incepted after 1990 and in such cases included all of the publications since inception. Collections vary depending on the span of the conference/journal, and was sometimes prohibited by the unavailability of high quality digital copies.

Our data set consists of 31,701 articles. Given the large volume of papers considered, we set out to develop a consistent and reproducible full text search strategy prior to commencing analysis. As a first step and where required, each article was inspected and prepared for a full text search. Following which a series of searches using the high level keywords of ``data quality'', ``quality of data'', ``information quality'', and ``quality of information'' were conducted. An article was selected as potentially relevant to our study if the keyword(s) occurred 3 times or more within the body of the text.

This process resulted in four (one per each keyword) large, overlapping, sets of articles, which were then cleaned by removing duplicates to arrive at a total of 961 unique documents. Once this set of unique articles was obtained, we started the process of elimination of irrelevant papers. The removal of such papers was necessary for the following reasons: (1) The theme of the article was irrelevant to data quality research; or (2) The keywords only appeared in the bibliographic references of the article, authors bio, or footnote; or (3) The keyword was a part of another phrase irrelevant to data quality e.g. ``quality of data modelling'', ``quality of data transmission'', ``quality of data-flow diagram'', to name a few.

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

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

The application of the above criteria reduced the relevant data set to 764. This narrowing indicates that 2.4% of the above listed publication outlets between the years 1990-2009 were directly and explicitly discussing concepts relevant to data or information quality. However, it was evident that the data set may also contain articles in which the chosen broad keywords may not necessarily explicitly appear, but the articles could still be implicitly related to the area and contain valuable outcomes. For example, papers within the database/computer science community that focus on *record linkage* may not contain any of the aforementioned four keywords but are still relevant to data quality research.

Accordingly, as a next step following the identification of the reduced dataset of 764 relevant papers, we identified a set of ‘second level’ keywords to further review the literature. To obtain an objective and relevant list, two researchers independently reviewed a sample (5%) of the initial set of articles to obtain further relevant concepts/keywords. The researchers identified the high level main theme(s) of the papers and associated these with terms and/or phrases that are representative of the theme e.g. terms such as *entity resolution*, *record linkage*, *data profiling*, *provenance* and *lineage* etc. Through this resource intensive activity, a large number of second level keywords were identified. The results of the two independent researchers were then compared, followed by a discussion to resolve any keyword conflicts. The agreed set of keywords was then later reduced as several did not return search results that were meaningful for data quality research.

A review of the second level keywords identified that several had synonyms. For example, *record linkage* had several related techniques such as *approximate join*, *similarity join*, *fuzzy matching* etc. Thus our identification of the second level keywords resulted in the development of a keyword taxonomy (see Figure 1). Finally, the identified keywords were also compared with a number of existing studies that have contributed to developing concept maps and various taxonomies for data quality, see e.g. (Lima and Macada, 2006; Ge and Helfert, 1996; Madnick and Wang, 2009). A number of augmentations were made to the list, including some further categories of the so-called second (and sometimes further) level keywords in order to ensure wider and more complete coverage.

It is important to point out that the purpose of this exercise was *not* to produce a classification of data quality research, but to be able to identify a sufficiently wide number of topics, so that the subsequent searches would return relevant results and produce a comprehensive coverage of the data quality research landscape. Accordingly, these new keywords were then used to search the data set again. The same strategy was used to prune the returned results as for the general keywords. After this second phase of analysis, a total of 1364 relevant publications were identified. Where there was a large group of publications (>50 papers) within a given keyword, an attempt was made to find sub keywords if possible e.g *edit distance*, *q-gram* etc. for *approximate matching*.

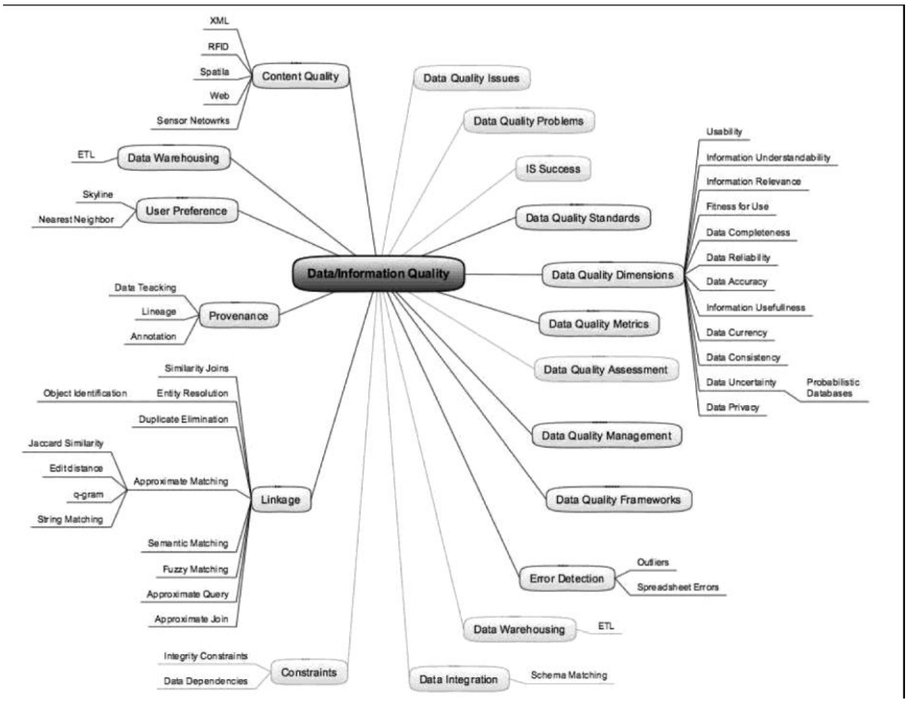

Figure 1. Taxonomy of Keywords

Finally the bibliographic data of the selected papers were recorded in a database together with the keyword(s) through which the paper was identified. Most publications were assigned more than one keyword. In general a publication with a lower level keyword (see Figure 1) was also associated with the higher level keywords. For example, a paper assigned the keyword *edit distance* will also have the keywords *approximate matching* and *linkage*. Further, some keywords were jointly assigned, e.g *data quality metrics* and *data quality assessment* co-exist due to the clear overlap in the topics. These steps were taken in order to ensure an effective search of the papers later through the database.

4 Key Findings

The taxonomy as presented in Figure 1, demonstrates the variability of themes in data quality research. In the sections below we present an analysis of the data based on above taxonomy from a number aspects. The main aim is to explore the connectivity (or conceptual distance) between IS and CS research contributions and thereby identify gaps and issues that may need to be addressed in order to work commonly towards a unified framework for data quality spanning across organizational, architectural and computational challenges.

4.1 Outlet Analysis

The outlet analysis has been conducted using the bibliographic database that was constructed as part of the study. Figure 2 presents a summary of the most widely published data quality topics, that is topics (or keywords) against which there was largest number of publications.

The top 3 topics of *data quality assessment*, *dimensions* and *metrics* depict a common set of papers as the three keywords were jointly assigned due to the high topical overlap and also to promote effective search in the database. Similarly *Constraints* and *Data Consistency* also represent a common set. Although the number of papers is too large to discuss individually, we studied further a selection of the papers in the first three sets, i.e. the combined *Assessment*, *Dimensions* and *Metrics* set (617), set of papers for *Linkage* (301) and for Content Quality (255), and present below a brief synopsis of the main themes as evident from the prominent (highly cited) papers in the area.

The first set, as expected is dominated by papers that focus on organizational aspects, particularly on measurement of *Information System Success* e.g. (Delone and McLean, 2003), (Rai and Lang, 2002), (Wixom and Watson, 2001). Classical papers such as (Wang and Storey, 1995), (Lee and Ling, 2002a), (Ballou and Wang, 1998), also appear here. However, some works from computational and architectural perspective also appear, predominantly on *Data Integration* related issues such as (Lenzerini, 2002), (Spaccapietra and Parent, 1992).

The second set on *Linkage* primarily consists of computationally focused papers and span both *Schema Matching* (Do and Rahm, 2002) as well as data matching (Hernandez and Stolfo, 1995) techniques. It also includes papers on data streams (Babcock and Babu, 2002), (Dobra and Garofalakis, 2002), semantics/ontologies (Rodriguez and Egenhofer, 2003), dominance and preference modelling (Kossmann. and Ramsak, 2002), (Papadias and Tao, 2003) rule based (Pirahesh and Hellerstein, 1992) and probabilistic (Dalvi and Suciu, 2007) approaches.

The third set on *Content Quality* contains papers from all three aspects, i.e organizational (mostly from IS outlets), architectural and computational. In addition publications specific to particular data types can be found such as XML (Lee and Ling, 2002), RFID (Jeffery and Garofalakis, 2006), sensor data (Sharaf and Beaver, 2004), spatial data (Koudas and Sevecick, 1997), web data (Katerattanakul and Siau, 1999), (Aladwani and Palvia, 2002), etc.

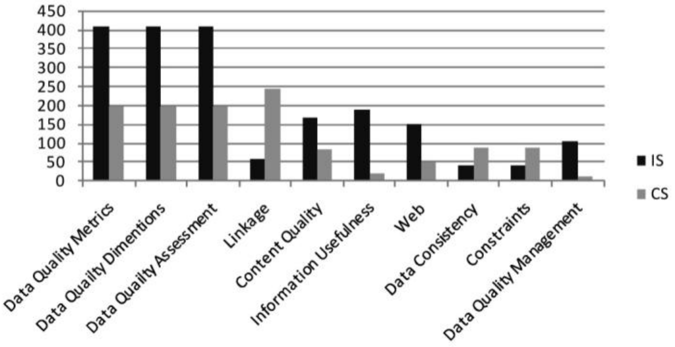

Figure 2. Keyword Frequency between IS and CS outlets

From the above, there is a clear indication that DQ themes are spread between IS and CS outlets. The overall distribution of papers between IS and CS outlets is summarized in figure 2. Clearly there are some topics where the overlap is greater (e.g *Data Quality Metrics*) than others (e.g *Linkage* and *Information Usefulness*). Although there is a generally increasing trend over the last several years in contributions to data quality research from both communities, the patterns of overlap have not shown substantial change (Sadiq, Yegenah and Indulska, 2011).

4.2 Topical Analysis

We conducted a further topical or thematic analysis of the papers through a text-mining tool called Leximancer (www.leximancer.com). Leximancer performs a full text analysis both systematically and graphically by creating a map of the concepts and themes re-appearing in the text. The tool uses a machine-learning technique based on a Bayesian approach to prediction. Once the optimal weighted set of evidence words is found for each concept, it is used to predict the concepts present in fragments of related text. In other words, each concept has other concepts that it attracts as well as concepts that it repels. The relationships are measured by the weighted sum of the number of times two concepts are found in the same block of text or between blocks of texts.

Leximancer uses concept maps to visualize the relationships. Each of the identified concepts is placed on the map in proximity of other concepts in the map through a derived combination of the direct and indirect relationships between those concepts (see Figure 3). Concepts are represented by labelled and color-coded dots. The size and the brightness of a concept dot on the map are indicative of the concept's strength within the body of analysed. The thickness and the brightness of connections between concepts is indicative of the frequency of co-occurrence of the two concepts. The relative distance of concepts on the map is indicative of similar conceptual contexts – i.e. the shorter the distance between the two concepts the closer in context they are. Thematic clusters of highly connected concepts are indicated through coloured circles, called themes.

To explore synergies and differences between data quality research in the CS and IS disciplines we conducted a series of Leximancer analyses for each of the top 10 keywords listed in Figure 2 (Sadiq, Yegenah and Indulska, 2011). For each of the keywords, data was analysed considering the CS publications in isolation, then considering the IS publications in isolation, followed by a joint analysis of both data sets to gain a better understanding of the common focus of the two disciplines.

Due to space limitations a detailed analysis is omitted in this paper. However as an example consider Figure 3, where we selected the set of 'Content Quality' related publications in the CS and IS publication outlets for a joint analysis.

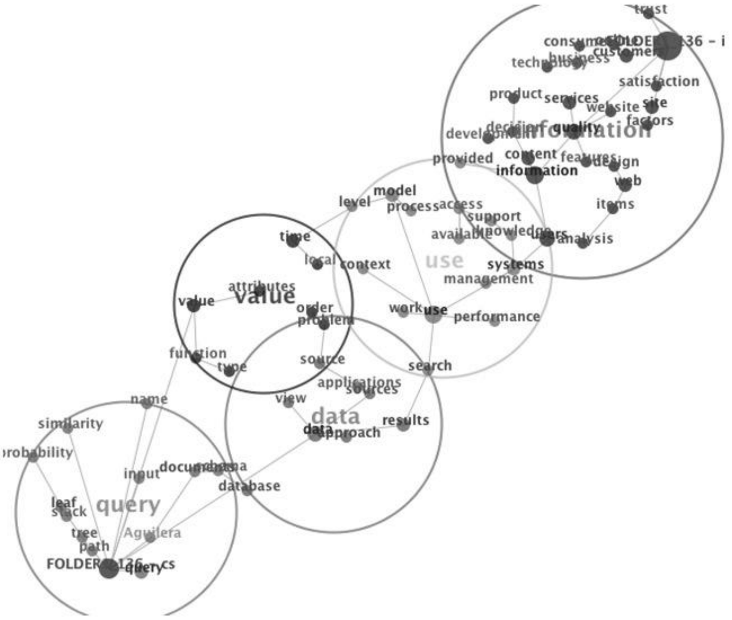

Figure 3. Content Quality - Information Systems vs. Computer Science publication focus.

In Figure 3 the collective Information Systems dataset related to the 'Content Quality'' topic is indicated by a 'FOLDER-136-IS'' concept. Likewise, the Computer Science dataset is represented by 'FOLDER-136-CS'' concept. Specifically, it shows the relationships of concepts related to Content

Quality across all considered publication years and how the data set relates to concepts that were identified to be the strongest common concepts across the two data sets. Our analysis indicates that, while there are concepts that are common to both data sets, the strength of the connection is weak (while this is not visible in Figure 3, due to resolution, the weakness is indicated in the Leximancer tool environment by the relative lack of thick, bright connections between both folder concepts and any one of the Content Quality concepts). Indeed, the analysis uncovers strong evidence that the Information Systems set of papers is strongly focused on information quality, issues relating to satisfaction and business value in general, yet it is not as strongly focused (as indicated by the relative distance of the themes from each other and the relative closeness of the themes to each of the two publication sets) on approaches for ensuring content quality. While this is not surprising in itself, given that Information Systems is a less technically oriented community of researchers, we see a weakness in a situation where the communities that should be collaborating together, appear to lack a strong collaboration and common focus. Strength of collaboration is further explored in subsequent sections and includes citation and network analysis to determine the degree of collaboration between the two areas.

4.3 Citation Analysis

In order to conduct the citation analysis we wrote a crawler script that searches all papers in the database within google scholar, and collects information regarding number of citations for the paper. Below we list the top cited authors and publication outlets. It is important to note that the citation counts are entirely based on the publications which are part of our collection and thus do not reflect the overall count for authors and/or publication outlets. The citation data has been further utilized to establish the strength of collaborations amongst the influential contributors in data quality research (see next section on network analysis).

Author	Citations	Author	Citations
Wang, R.Y.	4364	McLean, E.R.	1373
Widom, J.	2774	Halevy, A.	1308
Strong, D.	1986	Lenzerini, M.	1299
Ng, R.T.	1894	Lee, Y.W.	1183
Motwani, R.	1847	Gibbons, P.B.	1105
Datar, M.	1739	Knorr, E.M.	1071
Babcock, B.	1685	Koudas, N.	1061
Babu, S.	1607	Chaudhuri, S.	1056
Garofalakis, M.N.	1428	Shim, K.	1051
Rastogi, R.	1378	Hellerstein, J.M.	1014
DeLone, W.	1373		

Table 2. Authors with more than 1000 citations

Some of the earliest contributions came from *Wang, R.Y, Strong, D.* and associates on the identification of data quality dimensions and data quality assessment. These contributions have been heavily utilized by later researchers as is evident from the high citation count above. *Widom, J.* And co-authors have contributed substantially to the body of knowledge on data lineage and uncertainty especially through the Trio system (see infolab.stanford.edu/trio). Similarly works of *Ng, R.T.* on identification of outliers in large data sets, has profound applications in error detection, entity resolution and a number of data quality related problems. Although it is not possible to summarize the contributions of all highly cited authors, it is safe to conclude that the contributions of these influential contributors is indicative of the wide span of data quality research.

We further observe in Table 3 that in terms of top 10 cited outlets, the spread is roughly 60% between IS and CS contributions, in favour of CS. Even though there are publication outlets which have a larger number of papers in the database (presumably larger number of data quality relevant contributions), they do not seem to have a corresponding number of citations and have not made it in the top 10 venues listed below.

Publication Outlet	Area	Type	Abbreviation	Citations
International Conference on Very Large Databases	CS	Conference	VLDB	11388
International Conference on Management of Data	CS	Conference	SIGMOD	10693
Transactions on Knowledge and Data Engineering	CS	Journal	TDKE	5955
Communication of ACM	CS	Journal	CACM	5105
Principles of Database Systems	CS	Conference	PODS	4926
Journal of Management Information Systems	IS	Journal	JMIS	2938
Transactions on Information Systems	IS	Journal	TOIS	2703
Transons on Database Systems	CS	Journal	TODS	2523
Journal of Information Systems Research	IS	Journal	ISR	2337
Journal of Information and Management	IS	Journal	IM	1922

Table 3. Top 10 Cited Outlets (all keywords)

4.4 Network Analysis

In order to investigate patterns of co-authorship networks within the research community, we created a simple visualisation tool. The tool searches through our database of publications and constructs co-authorship graphs.

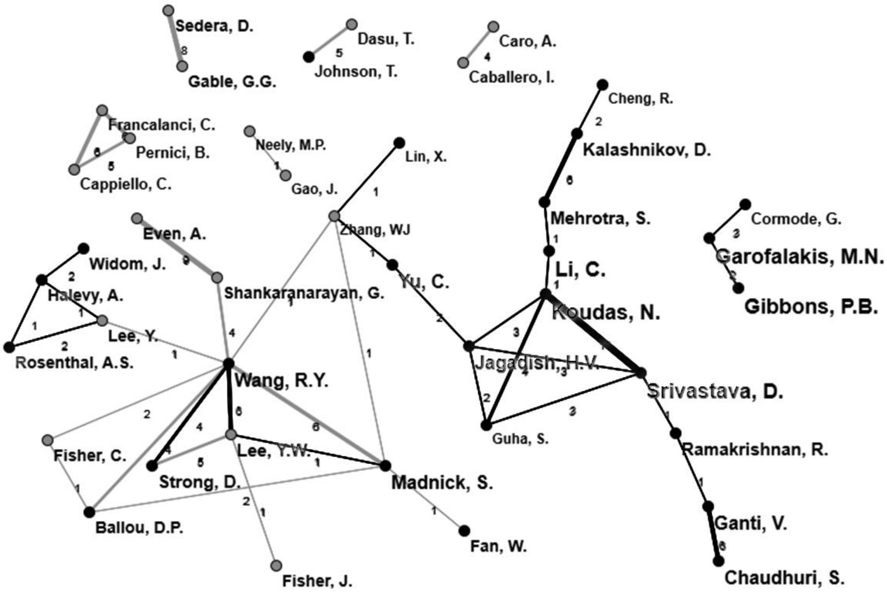

Figure 4. Co-authorship network for authors with 5 or more papers

These graphs are colour coded based on the area (IS/CS) of the author or relationship. Grey colour represents IS whilst Black colour represents CS. An author is classified as CS (IS) if at least half of the papers in which the author is a first author appear within CS (IS) outlets. Likewise, if half or more of the co-authorship papers of two authors are published in CS (IS) venues, the co-authorship link is colored Black (Grey). The size and boldness of the author name is a reflection of number of publications within the database (presumably DQ related). Lastly, co-authorship link is labelled by number of co-authored publications.

The graph in Figure 4 depicts the co-authorship network of authors who are first author of 5 or more papers in our database. The graph shows a number of fragmented smaller groups of authors, some larger groups of co-authors e.g those centred around *Wang, R.Y* and some strong co-authorship links e.g 11 co-authored papers between *Koudas, N* and *Srivastava, D.*

The graph in Figure 5 shows the co-authorship network for top 50 highly cited authors in our database. However only a little more than half of the top 50 authors appear in a co-author relationship and hence appear in the network visualization. The relationship strength is mostly limited to 1-2 papers with some exceptions e.g *Wang, R.Y* and *Lee, Y.W* with 6 co-authored papers. An interesting pattern in the figure is that many influential contributors seem to have links that span IS/CS domains, e.g. *Spaccapietra, S.* with more CS publications, is co-author with *Parent, C.* with more IS publications. Similarly *Lee, Y.W.* with more IS publications is co-author with *Wang, R.Y.* and *Strong, D.* with more CS publications.

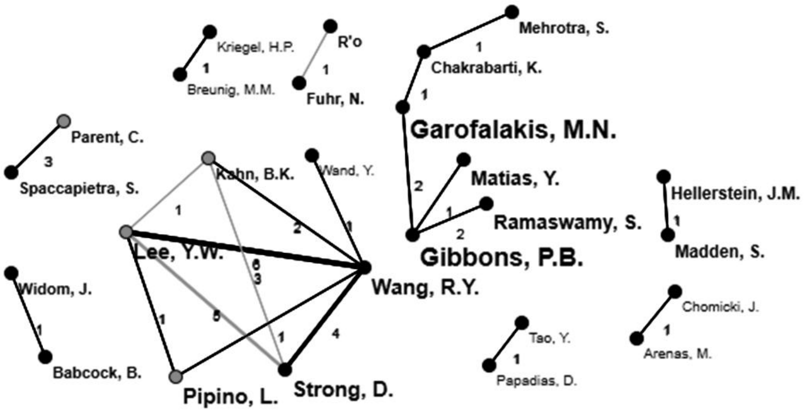

Figure 5. Co-authorship network for top 50 highly cited authors

We also conducted an analysis of authors with more than 3 papers together. Due to space limitations the full analysis cannot be presented, but it was noted that many authors in CS individually had co-authorship relationships amongst them that were of IS nature. For example *Ballou, D.P., Wang, R.Y.,* and *Madnick, S.*, have more CS publications individually, but their co-authorship is more of IS nature.

5 Conclusions

The dataset used to present the analysis in this paper has been recorded in a bibliographic database. The database is publically available (excluding full text of publications) from dqm.itee.uq.edu.au. Access to the databases is restricted to members of the portal, but membership is free. The bibliographic database has advanced search functionality i.e. it can be searched on various fields within the bibliographic entry including the keywords as presented in the taxonomy in Figure 1.

The results of the analysis as presented in this paper provide key insights into the nature of cross-disciplinary research on data quality with respect to outlets, topics, citations and collaboration networks. The study is by no means complete or without limitations. One of the key criteria for assessing degree of collaboration is co-citation. Unfortunately this data is not very easy to obtain and hence not included in the analysis so far. Furthermore, the body of knowledge considered so far has been restricted to academic contributions. A number of experienced practitioners have made significant contributions to data quality solutions (see e.g. www.infoimpact.com, www.gfalls.com, knowledge-integrity.com and www.dataqualitysolutions.com). Similarly commercial products provide a number of sophisticated solutions, and address data quality topics conspicuously missing from academic outlets e.g. data profiling and standardization. Further the perspective of data users is essential to understanding the key issues and challenges, particularly given that data quality is a largely domain specific problem due to the importance of fitness for use in the definition of data quality metrics. We hope as a first step in our future work to complement the research analysis by industry input relating to key issues and challenges. We hope that the extended work will provide a comparison platform against the research (solution) contributions and help identify true gaps and an industry relevant research agenda for this area.

Overall, based on the above analysis, we see a need for a stronger connection of technical solutions published by the Computer Science publication outlets and the application (and evaluation) of those technical solutions in the Information Systems discipline. Organizations rely strongly on information to support their operational and strategic decision making, thus any quality problems related to information that resides in organizational systems will have a negative impact on the overall operations of the organization in the long term. Therefore, we see the practice of developing data quality solutions in isolation from the real-world application of those solutions, and the practice of developing solutions without a strong business motivation for them, as counter-productive to reducing data quality problems in organizations. Accordingly, we call for the two communities to increase their awareness of the relevant leading research that is often published in publication outlets that are not on the immediate radar of the research community. This paper is the initial step to increase the connection and highlight synergies and differences between the foci of the two communities in data quality related research.

References

Aladwani, A.M. and Palvia, P.C. (2002). Developing and validating an instrument for measuring user-perceived web quality. Information & Management, 39(6): pages 467-476

Babcock, B. Babu, S. Datar, M. Motwani, R. and Widom, J. (2002). Models and issues in data stream systems. In Proceedings of the twenty first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 1-16.

Ballou, D. Wang, R. Pazer, H. and Kumar, G. (1998). Modelling information manufacturing systems to determine information product quality. Management Science, 44(4): pages 462-484.

Benbasat, I. and Zmud, R.W. (2003) The identity crisis within the IS discipline: Designing and communicating the discipline's core properties. MIS Quarterly, 27(2):pages 183-194.

Carr, N. (2004). Does IT Matter? Information Technology and the Corrosion of Competitive Advantage, Harvard Business School Press.

Chen, C. Song, I.Y. and Zhu, W. (2007). Trends in conceptual modeling: Citation analysis of the ER conference papers (1979-2005). In Proceedings of the 11th International Conference on the International Society for Scientometrics and Informatrics, pages 189-200.

Dalvi, N. and Suciu, D. (2007). Efficient query evaluation on probabilistic databases. The VLDB Journal, The International Journal on Very Large Data Bases, 16(4): pages 523 - 544.

Delone, W.H. and McLean, E.R. (2003). The DeLone and McLean model of information systems success: A ten-year update. Journal of management information systems, 19(4): pages 9-30.

Do, H.H. and Rahm, E. (2002). COMA: a system for flexible combination of schema matching approaches. In Proceedings of the 28th international conference on Very Large Data Bases, pages 610-621.

Dobra, A. Garofalakis, M. Gehrke, J. and Rastogi, R. (2002). Processing complex aggregate queries over data streams. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data, pages 61-72.

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

Ge, M. and Helfert, M. (1996). A Review of Information Quality Research. In The 12th International Conference on Information Quality, MIT, Cambridge, Massachusetts, USA, pages 1-9.

Hernandez, M.A. and Stolfo, S.J. (1995). The merge/purge problem for large databases. In Proceedings of the 1995 ACM SIGMOD international conference on Management of data.

Jeffery, S.R. Garofalakis, M. and Franklin, M.J. (2006). Adaptive cleaning for RFID data streams. In Proceedings of the 32nd international conference on Very large data bases, pages 163-174.

Juran, J.M. (1962) Quality control handbook.

Katerattanakul, P. and Siau, K. (1999). Measuring IQ of websites: development of instrument. In Proceedings of 20th international conference on Information Systems, pages 279-285.

Kossmann, D. Ramsak, F. and Rost, S. (2002). Shooting stars in the sky: An online algorithm for skyline queries. In Proceedings of the 28th international conference on Very Large Data Bases, pages 275-286.

Koudas, N. and Sevcik, K.C. (1997). Size separation spatial join. ACM SIGMOD Record, 26(2): pages 324-335.

Lee, M. Ling, T. and Low, W. (2002). Designing functional dependencies for XML. Advances in Database Technology EDBT 2002, pages 145-158.

Lee, Y.W. Strong, D.M. Kahn, B.K. and Wang, R.Y. (2002). AIMQ: a methodology for information quality assessment. Information & Management, 40(2): pages 133-146.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In Proceedings of the twenty first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, page 233-246.

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

Madnick, S.E. Wang, R.Y. Lee, Y.W. and Zhu, H. (2009). Overview and Framework for Data and Information Quality Research. Journal of Data and Information Quality (JDIQ), 1(1): pages 1-22.

Neely, M.P. and Cook, J. (2008). A Framework for Classification of the Data and Information Quality Literature and Preliminary Results (1996-2007). AMCIS Proceedings.

Papadias, D. Tao, Y. Fu, G. and Seeger, B. (2003). An optimal and progressive algorithm for skyline queries. In Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pages 467-478.

Pirahesh, H. Hellerstein, J.M. and Hasan, W. (1992). Extensible/rule based query rewrite optimization in Starburst. In Proceedings of the 1992 ACM SIGMOD international conference on Management of data, pages 39-48.

Rai, A. Lang, S.S. and Welker, R.B. (2002). Assessing the validity of IS success models: An empirical test and theoretical analysis. Information Systems Research, 13(1): pages 50-69.

Rodrguez, M.A. and Egenhofer, M.J. (2003). Determining semantic similarity among entity classes from different ontologies. IEEE transactions on knowledge and data engineering, pages 442-456.

Sadiq, S., Yeganeh, N. K. and Indulska, M. (2011). 20 years of data quality research: Themes, trends and synergies. In Proceedings of the 22nd Australasian Database Conference (ADC 2011). Perth, WA, Australia, (1-10). 17-20 January 2011: pages 1-10.

Sharaf, A. Beaver, J. Labrinidis, A. and Chrysan, K. (2004). Balancing energy efficiency and quality of aggregate data in sensor networks. The VLDB Journal, The International Journal on Very Large Data Bases, 13(4): pages 384-403.

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

Spaccapietra, S. Parent, X. and Dupont, Y. (1992). Model independent assertions for integration of heterogeneous schemas. The VLDB Journal, 1(1): pages 81-126.

Wang, R.Y. Storey, V.C. and Firth, C.P. (1995). A framework for analysis of data quality research. IEEE Transactions on Knowledge and Data Engineering, 7(4): pages 623-640.

Wixom, B.H. and Watson, H.J. (2001). An empirical investigation of the factors affecting data warehousing success. MIS quarterly, pages 17-41