FRAMING DATA QUALITY RESEARCH: A SEMANTIC ANALYSIS APPROACH

(Research-in-Progress)

Roger Blake
University of Massachusetts Boston
roger.blake@umb.edu

G. Shankaranarayanan
Babson College
gshankar@babson.edu

Abstract: Research in data and information quality has made significant strides in the last decade and has created an expansive body of knowledge. Given the multiple different research perspectives and research methodologies adopted, it is important for us to understand the research topics and themes that have evolved and currently define this body of research. Here, we present the results of a preliminary study that aims to provide a better understanding of this research area by identifying the core topics and themes. We analyze abstracts of 467 journal and conference articles published over the past ten years in data and information quality. Latent semantic analysis (LSA) is used to develop term-to-term semantic similarities and term-to-factor loadings. From the analysis, we identify five core topics and fourteen core themes of data quality research. The results from this research can significantly improve our understanding of the body of literature in data and information quality. Taken a step further, this research can offer insights into how themes and topics have shifted over time, what topics/themes have garnered the attention of researchers and when, and reveal the research trends in this area. Above all, it can motivate research by helping researchers associate research methods with research topics, identify themes/topics that have not been studied and data quality dimensions that have not been examined sufficiently.

Keywords: Data and information quality research, research taxonomies and frameworks, latent semantic analysis

INTRODUCTION

Research in data and information quality has shifted towards a unified body of knowledge from one spread through many reference disciplines [1]. It is therefore imperative that we formulate a unified understanding of the core research topics and themes within this research area. Many have attempted to summarize, classify, and develop frameworks to define the core concepts in this research area (e.g., [2] [3][4][5]). They have all examined the literature in data quality and defined what they believe is the area of data quality – their point-of-view of the research area. Although these offer invaluable lessons that help us understand the research area better, we posit that there is a more interesting point-of-view that comes, not from the researchers, but from the research itself. Can this body of literature tell us the core topics examined within the research area and the key themes within each topic? Can we understand what research themes have risen to the forefront and the ones that are ebbing? Can we understand the evolution of research themes? Can we associate research topics with data quality dimensions and determine the association (if any) between topics and dimensions? The summaries, classifications and frameworks already developed do not answer such questions. As data/information (as many others have done in work that precedes ours, we too use the terms *data* and *information* quality interchangeably in this paper) quality researchers it is important for us to understand the point-of-view that emanates from the body of literature. Our primary motivation for our research is to take a step towards determining this point-of-view by identifying the core topics and themes, as communicated by the body of literature in data quality. Another motivation is to develop a reproducible method to understand how this core changes over time.

The specific objectives of this paper are: (1) to identify a clear set of research topics and themes to define the body of literature in data quality. (2) To identify associations (if any) between dimensions and research topics to highlight dimensions that have been rigorously studied and those that not been studied. (3) To identify interesting additional analyses that can help us better understand research in data quality, based on our observations and experiences from conducting this research.

Seeking the identity of an area of research is not new. From almost forty years ago when Mason and Mitroff [6] published a program for information systems (IS) research, through more recent "crises of identity" described by Benbasat and Zmud [7], the identity of the IS discipline has been discussed. According to Benbasat and Zmud, finding the core topics and themes from IS research is critical to finding the identity of the IS discipline as a whole: "We argue that the primary way in which a scholarly discipline signals its boundaries – and in doing so, its intellectual core – is through the topics that populate discipline-specific research activities." Benbasat and Zmud were concerned that the topics in IS research were becoming amorphous, diffuse, and indistinct from reference disciplines. They expressed worry that an ambiguous identity for the IS discipline would ultimately undermine its very existence. This "search for identity" is a key motivation for us as we embark on this research. Lima, Maçada and Vargas [5] saw the ambiguity of an identity and the relation to reference disciplines as parallels between the IS discipline and data quality research. This further motivates our research.

We proceed by analyzing the content of the research documents in data quality. As a first step, we analyze the abstracts of journal articles and conference proceedings considered to have data quality as the primary focus. Abstracts are the primary source of data to develop research topics in a diverse range of fields including business strategy [8] and the sciences [9].

In the remainder of this paper, we first review the relevant literature by reviewing prior work that summarizes and/or classifies data quality research to define the scope of this paper. We also provide an overview of LSA, the tool we adopt in this research. We then describe the methodology and present the results of our analysis. We finally offer our conclusions together with directions for further research.

RELEVANT LITERATURE

The question of identity is central to our research. Albert and Whetten [10] defined an organization's identity by three claims that it can make: the claim to a central character, the claim to distinctiveness, and the claim of temporal continuity. Our research can help support two of these claims for data quality research: identifying distinct core topics and themes can support the claim to a central character. Evaluating how these topics and themes change over time can support the claim of temporal continuity.

However, identifying core topics and themes does not equate to defining an identity for an entire body of knowledge. Prior work that has analyzed keywords, citations, and developed frameworks of data quality research is all useful and important. Wang, Storey and Firth proposed one of the earliest frameworks of data quality from a comprehensive analysis of publications through 1994 [11]. Their framework used the analogy of data and data quality to a manufactured physical product and its quality, and consisted of seven elements and their subsections. The seven elements were Management Responsibilities, Operation and Assurance Costs, Research and Development, Production, Distribution, Personnel Management, and the Legal Function. The authors also pointed to specific research challenges and called attention to research on economics of data quality, standardization of data quality metrics, and effects of data quality policies. Our research can help measure how well the community has responded to these calls - the method we develop can analyze publications over discrete time-periods.

Neely and Cook [3] developed a novel framework by combining the factors of "fitness for use" as defined by Juran [18] and the elements in the framework by Wang et al. [11]. From codifying 74 articles on data quality published since 1995, they assigned each to the set of categories in their framework considered most applicable. The most frequent categorizations were for the four combinations of the "what" and "how" aspects of Juran's "fitness for use" [18], and the "Distribution" and "Dimensions" elements of Wang et al.'s framework [11].

From examining articles Neely and Cook were able to enumerate specific research questions. For those most frequent categorizations the research questions related largely to meta data, distribution processes and procedures, data integration, data flows and for data quality dimensions, defining, measuring, and application. The topics of these questions are consistent with the findings from this research, although from our analysis the research of dimensions has become more integrated with other research as will be discussed in the results.

Lima et al. [5] developed conceptual maps of data quality research using articles published from 1995 through 2005. They chose proceedings from conferences of central interest to the data quality community as their primary data. Of the 171 proceedings reviewed by Lima et al., 86% were from the Internal Conference on Information Quality, with the remainder chiefly from the proceedings of the International Workshop on Information Quality in Information Systems. From this body of work Lima et al. developed a list of 279 keywords and defined three high-level views of data quality research: the organizational, behavioral, and operational views. Within each view, they proposed a highly detailed conceptual map of these keywords specifying relationships between keywords and their groupings. They derived relationships based on the judgment and intuition of the researchers. The method we use in our research, LSA, is often used to reach the same ends. It is known that LSA makes judgments similar to those made by humans. Frameworks occasionally reference exemplar papers to illustrate their categories - this is another application where LSA performs well. Lima et al.'s framework of 279 keywords was much more detailed than the topics and themes developed in our study. Comparable to this framework, we can expand the research topics and themes from our study to a very substantial number of sub-themes.

A comprehensive review by Ge and Helfert [4] divided research into that focusing on the assessment, management, and contextual aspects of data quality. In particular, they examined data quality assessment in depth. They further divided DQ assessment into three sub-categories: problem identification, data quality dimensions, and assessment methodologies. From synthesizing prior research, the authors developed conceptual maps and models for those sub-categories and enumerated relevant papers in each. Ge and Helfert's review and conceptual models identify and categorize research topics, and relate to the goal of our study. The key difference is that our study builds topics and themes from a semantic analysis of text, and does not develop conceptual models or organize a body of literature. Further, while Ge and Helfert's framework offers a static picture of research built top-down, the framework developed in our study provides a more dynamic image that is built bottom-up.

A recent framework proposed by Madnick, Wang, Lee, and Zhu [1] used two dimensions, topics and methods, to categorize data quality research. They specified research methods at varying levels of granularity; treating some methods as subsets of others, such as statistical analysis as a type of quantitative method. Neither topics nor research methodologies are considered mutually exclusive in this framework, and papers from different disciplines could span multiple categories along both dimensions. Using the framework and keywords associated for topics and subtopics (shown in Table1), researchers can characterize their own research. Our framework can complement this and other frameworks proposed in literature, as it does not attempt to separate methodologies from themes. Instead, it attempts to determine associations between topics/themes and methodologies. We believe this helps us understand why certain methodologies are more popular with specific research topics/themes – an insight that allows researchers to choose more appropriate methodologies based on their research topic.

Research Topics	Research Methods
1. Data quality impact	1. Action research
1.1 Application area (e.g., CRM, KM, SCM, ERP)	2. Artificial Intelligence
1.2 Performance, cost / benefit, operations	3. Case study
1.3 IT management	4. Data mining
1.4 Organizational change, processes	5. Design science
1.5 Strategy, policy	6. Econometrics
2. Database related technical solutions for data quality	7. Empirical
2.1 Data integration, data warehouse	8. Experimental
2.2 Enterprise architecture, conceptual modeling	9. Mathematical modeling
2.3 Entity resolution, record linkage, corporate	10. Qualitative
2.4 Monitoring, cleansing	11. Quantitative
2.5 Lineage, provenance, source tagging	12. Statistical analysis
2.6 Uncertainty (e.g. imprecise, fuzzy data)	13. System design, implementation
3. Data quality in the context of computer science and IT	14. Theory and formal proofs
3.1 Measurement, assessment	
3.2 Information systems	
3.3 Networks	
3.4 Privacy	
3.5 Protocols, standards	
3.6 Security	
4. Data quality in curation	
4.1 Curation-Standards and policies	
4.2 Curation-Technical solutions	

Table 1: Framework of data quality research by Madnick, Wang, Lee, and Zhu (2009)

Our methodology makes use of latent semantic analysis (LSA), a statistical method for finding semantic relationships within a corpus of documents. There is a wide range of uses for this method; LSA has been successfully used to predict the subjective ratings of essays made by human readers, to match human categorizations of terms [12], and measure textual coherence [13]. LSA is a "bag-of-words" approach that analyzes the frequency and co-occurrences of terms within a corpus to infer semantic similarity (or lack thereof) between terms and between documents. This is quite different from other approaches such as analyses of keyword frequencies or counts of citations; these approaches can require terms to be exact, or near exact, matches. LSA can relate terms that are different, yet used in similar contexts.

Sidorova et al. [2] used LSA to derive the core research topics within the discipline of Information Systems (IS). These authors analyzed a corpus of 1,615 abstracts from papers published in *MIS Quarterly, Information Systems Research*, and the *Journal of Management Information Systems* through 2006. With LSA they were able to define five major core topics: information technology (IT) for organizations, IT and individuals, IT and markets, IT and groups, and IS development for the fifth. They also developed thirteen themes, such as for Decision Support Systems, Virtual Collaboration, and Research Methodology. As we do, these authors used the abstracts of papers in their analysis.

Examining both topics and themes, Sidorova et al. determined that although the core topics of IS research have remained stable, the underlying themes have continued to evolve from the 1980's through the present. We model our work based on this research and have similar objectives.

The next section presents our methodology for producing factor loadings for terms based on their semantic similarities.

METHODOLOGY

We first collected abstracts of articles that deal with data quality from both journals and conference proceedings. We then prepared the data for analysis by performing a sequence of steps to remove terms that could potentially obfuscate our analysis. We also combined some terms to create a consistent representation for the analysis. We finally applied latent semantic analysis to extract semantically related terms and to identify the factors based on the loadings of these terms. We provide a more detailed explanation of our methodology in the following paragraphs.

Abstract corpus

The abstracts used to build our corpus were from the proceedings of conferences and journals of central interest to the data-quality research community and from journal articles with a focus on data quality. These conferences included the Americas Conference on Information Systems (AMCIS), International Workshop on Information Quality in Information Systems, and the International Conference on Information Quality (ICIQ - from 1999 until 2008). The journals included the Journal of Management Information Systems, ACM Journal of Data and Information Quality, Communications of the ACM, Decision Support Systems, Advances in Management and the International Journal of Information Quality. Abstracts were included from journal articles having either the keywords "data quality" or "information quality" from a search of the EBSCO, ACM, IEEE, and ISI databases, and from special issues of journals with that focus. We included 467 abstracts of which 314 were from conference proceedings and 153 were from journal articles. Table 2 has the number of abstracts from the most popular outlets that formed the basis of our research.

Publication Outlet	Abstract count
ICIQ - International Conference on Information Quality	235
IJIQ - International Journal of Information Quality	38
AMCIS	32
International Workshop on Information Quality in Information Systems	18
Advances in Management Information Systems	14
ACMJDIQ	12
CACM - Communications of the ACM	10
JMIS - Journal of Mgt Information Systems	8
Decision Support Systems	7
Others	93
Total	467

Table 2: Abstract Counts – By Publication Outlet

The majority of abstracts (95%) were from journal articles and from conference publications from 2000 through the beginning of 2010 with the count of abstracts as shown in Table 3.

Prior to 2000	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
23	27	33	22	29	45	61	39	67	51	61	9

Table 3: Distribution of abstracts by year of publication

We applied several routine pre-processing steps to the texts of the 467 abstracts in our corpus prior to applying LSA. First, we removed punctuation marks, special characters, and numeric values. In the second step, we removed stop words. Stop words are short, commonly occurring words such as "a", "I", and "the", that add little to the ability to distinguish among abstracts. For the same reason we removed several other words and phrases with high frequency. Among these were the phrases "data quality" and "information quality". After examining the remaining words, we removed additional words with little relevance to data quality such as "during", "largely", and "itself". Finally, we excluded words consisting of less than three characters and words appearing less than five times throughout all abstracts.

In the third step, we standardized frequently occurring phrases. For example, the phrase "total data quality management" was standardized to TDQM so that occurrences of the former would match the later.

The fourth step was to stem all words in the corpus. Stemming standardizes words having multiple variations with semantically equivalent meanings. Often these are words with the same root but with multiple suffixes. For example, stemming might transform "decide", "deciding", and "decides" all to the root term "decid". For our analysis, we used the Snowball stemmer, an implementation of the popular Porter stemming algorithm. At the end of the stemming process there were 453 unique terms in our corpus.

Latent Semantic Analysis

Latent semantic analysis (LSA) is a dimension reduction technique that uses singular value decomposition (SVD). SVD is a form of factor analysis applied to a t by d term-document matrix. In our study, this was a matrix of 453 terms and 467 documents. Term-document matrices represent the frequencies of terms as they appear in each document. Raw frequencies counts are usually transformed using a weighted value proportional to a term's frequency in a document and inversely proportional to the number of documents in which the term appears. This weighted value de-emphasizes the significance of terms that appear in many documents; a term appearing in all documents contributes little value to being able to differentiate among documents. For our analysis, the weights for each term were proportional to a binary transformation of term frequency and the logarithm of the inverse frequency of the term across all documents.

We used SVD to reduce the weighted term-document to r dimensions and produces three component matrices: a matrix T with t rows and r columns, a matrix D with r rows and d columns, and a diagonal scaling matrix S with r rows and r columns. Matrix S contains the square roots of the eigenvectors from SVD in sorted order; if T, S, and D are multiplied together, they will approximate the original term-document matrix. A t by d matrix of similarities between terms can be found from the dot-product of TSD', an analogous matrix that can be generated to find the similarities between documents. For a more detailed description of LSA, SVD, and similarity measures, readers are referred to Deerwester, Dumais, Furnas, Landauer, and Harshman [14].

For analyzing n-grams and word frequencies we used AntConc 3.2.1 and for LSA we used R version 2.11.0 and the LSA, Snowball, and RStem packages. This software is from the R Foundation and is available at http://www.r-project.org/. We present the results of our analysis next.

RESULTS

We used LSA to find semantically related terms and their loadings onto individual factors; terms with factor scores lower than 0.15 were set aside. Eighty percent of the total variance was explained by 156 of

the terms (34.4%), and 214 terms (47.2%) explained 90% of the total variance. Fifty-four terms (11.9%) loaded on three or more factors, and 161 (25.3%) loaded on two factors.

We evaluated several different combinations of factors to find the most logical groupings of terms into topics and themes. From this process, we determined that the most meaningful topics may be constructed with five factors and the most meaningful themes with fourteen factors.

Table 4 shows the terms with high loadings most relevant to each of these five topics and the fourteen themes. It is important to note the following: (a) Terms are words transformed by stemming, and not necessarily the original word as it appeared in abstracts. (b) Several terms loaded across multiple themes – we selected the theme for a specific term based on where it loaded the highest. (c) A theme is a name that we assigned to a "factor" based on our knowledge of what the collection of terms under this factor represent. (d) The grouping of themes into core topics is reducing from the number of factors (or themes) from LSA and naming the group in accordance with how the terms reloaded. We named these groups based on our knowledge of the research area. (e) The tool used offered us the ability to "look-up" a term and connect it to the abstract(s) from which the analysis extracted that term. We used this "check-back" capability to confirm our themes and topics.

We identified the first theme, "Methods to Analyze Data Quality", from the factor that included terms such as *check, defect, failur, inaccur, incomplet, inconsist, monitor, period, root, solv, caus, problem, improv, flow,* and *resource.* From our understanding and by checking-back, these terms came from abstracts of articles that proposed methods to examine data quality and addressed more than one data quality dimension. After examining the factor with terms such as *applic, architectur, benefit, framework, implement, informationproduct, ipmap, link, mean, measur, meta, metadata, metric, model, procedur, process, prototyp, repositori, softwar, system,* and *tool*, we named it "Systems for Measuring Data Quality". Several of the abstracts (articles) from which these terms were extracted proposed frameworks and/or embedded frameworks within prototype systems for measuring data quality. Similarly, we named the factor that included terms such as *cost, optim, tradeoff, util* as "Economic Aspects of Data Quality". These are exemplar factors that were relatively easy to identify and name as themes.

Amongst the challenging factors was one with term loadings such as *actual, error, estim, linear, manageri, paramet, prioriti, rate, simul,* and *valu.* Initially it appeared that this factor should have been part of the first factor "Methods to Analyze Data Quality", but these terms either did not load at all on the first factor or had loadings that were extremely small. Looking back, we identified that the terms came from articles that measured a single dimension of data quality, but with the intent of suggesting methods to prioritize data quality improvements (along that dimension). We hence named this factor as "Prioritizing Data Quality Improvements".

There were two other factors, whose identification was challenging. The first factor included terms such as *consum, contextu, firm, govern, industri, profession, social, task, manag,* and *busi.* The second factor included the terms *context, implic, interpret, percept, polici, practition, report, usag,* and *perceiv.* At the outset, both factors appeared to deal with contextual data quality. However, some of these terms, such as *context, usag,* and *perceive,* loaded only on the second of the two factors and not on the first. Also, terms such as *task, manag, consum* and *contextu* loaded only on the first of the two factors and not on the second. In the process of checking-back, we identified that research in contextual data quality has two distinct themes – one that deals with using data quality for decision support in the context of the task and managerial decision-making, and the other that deals with the contextual assessment of data quality. This helped us identify these two factors as two separate themes: the first as the "Role of Data Quality in Decision Making" and the second as "Contextual Assessment of Data Quality".

Core topics and themes		Sample terms
1 Data quality assessment	1 Methods to analyze data quality	Check, defect, failur, inaccur, incomplet, inconsist, monitor, period, root, solv, caus, problem, improv, flow, resourc
	2 Information systems for measuring data quality	Applic, architectur, benefit, framework, implement, informationproduct, ipmap, link, mean, measur, meta, metadata, metric, model, procedur, process, prototyp, repositori, softwar, system, tool
	3 Economic aspects of data quality	Cost, document, optim, plan, scale, size, tradeoff, transact, util, work
	4 Ontology and Knowledge Management	Formal, identif, knowledg, ontolog, semantic, rule, expert, automate, meaning
2 Management of data quality	5 Prioritizing data quality improvements	Actual, error, estim, linear, manageri, paramet, prioriti, rate, simul, valu
	6 Methods to improve data quality for applications	Appli, attribut, classif, cluster, dataclean, datacleans, datamin, dataset, detect, duplic, entiti, linkag, object, pattern, rule, schema, semant, transform
	7 Data quality at the enterprise level	Chain, competit, custom, deliveri, economi, infrastructur, invest, manufactur, product, project, stakehold, supplier, communiti, corpor, healthcar, market, organis, organiz, personnel, regul, structur
	8 The role of data quality in decision making	Consum, contextu, firm, govern, industri, profession, social, task, manag, busi
3 Quality of data in repositories	9 Data quality in data warehouses	Entiti, engin, merg, store, subject, datawarehous, metadata, captur, maintain, databas, entiti, collect, structur, subject
	10 Data quality in relational databases	Constraint, criteria, databas, domain, extract, manipul, queri, standard, updat, user
	11 Contextual data quality	Context, implic, interpret, percept, polici, practition, report, usag, perceive
4 Data quality in networked data	12 Data quality on the world wide web	Group, internet, onlin, satisfact, site, commerce, world, wepag, communit, user
	13 Data quality in sensor networks	Devic, dynam, intellig, locat, mobil, network, sensor
5 Research design and methodologies	14 Research methods in data quality	Analysi, factor, indirect, instrument, interdepend, principl, taxonomi, case, effect, control, hypothes, impli, mediat, relationship, signific, theori, treat, design, survey, study, simul, empir

Table 4: Sample of relevant term loadings

Terms such as *analysi, factor, indirect, instrument, interdepend, effect, control, hypothes, impli, mediat, relationship, signific, theori, treat, design, survey,* and *empir* loaded on more than one factor. However, all of them loaded more heavily on one specific factor indicating that they had an identity of their own. We hence named this theme as "Research Methods in Data Quality".

We were able to identify four clear groupings of themes as topics: (1) Data Quality Assessment with four themes; (2) Management of Data Quality with four themes; (3) Data Quality in Data Repositories with three themes; (4) Data Quality in Networked Data with two themes. The fifth grouping, "Research Methods", has only one theme. This last topic and its theme is consistent with the framework proposed by Madnick et al. in 2009 in which research methodologies were separated from the rest.

Key terms that did not appear in our analysis are *data lineage, entity resolution, provenance, tagging,* and uncertainty. There are two possible reasons for this. First, we looked for terms that appeared at least five times in our corpus of abstracts. Second, we did not collect articles from journals or proceedings in computer science. Computer science research often addresses data lineage and provenance in association with data warehouses. We believe that entity resolution does not appear because we did not have sufficient article abstracts in this area.

To establish a sense of how well our research themes are semantically distinct we conducted a follow-up analysis. The cosine-similarity measure normalizes for the effects of varying frequencies of term appearances and hence similarity is determined using this measure. The measures are calculated from the dot products of vectors in the reconstructed term-document matrix *W* produced by latent semantic analysis. For any two terms *i* and *j*, the cosine measure is calculated from the row vectors *i* and *j* in *W*. The cosine-similarity measure is defined as:

$\frac{t_i \cdot t_j}{\|t_i\| \|t_j\|}$, equivalent to: $\frac{\sum_k t_{ik} t_{jk}}{\sqrt{\sum_k t_{ik}^2 \sum_k t_{jk}^2}}$

We calculated the cosine similarity between each term and the average cosines between the terms constituting each of the fourteen research themes. When the cosine between a term and the terms comprising a research theme is low, there should be little semantic relationship between the term and that theme. Correspondingly, a higher cosine should indicate a high semantic relationship.

The cosines of terms within two themes with the average cosine of terms in two of the research themes, "Data quality at the enterprise level" and "Data quality in relational databases", is shown in Figure 1. For example, the term "queri" might logically be more associated with databases than enterprises. The cosine between this term and the average cosine of terms comprising the data warehousing theme (excluding the term "queri") is approximately 0.12, whereas the cosine of this term with the enterprise level theme is approximately 0.05. Conversely, one might expect the term "market" to be closer to enterprises than databases. This term has a cosine of approximately 0.11 with the other terms in the enterprise level theme and 0.04 with the terms in the theme of relational databases.

What is most significant is that Figure 1 shows the clear separation between the terms loading onto each of these two research themes, indicating that the analysis is distinguishing between the semantics of these two themes. Examination of analogous plots for other combinations of research themes showed this general pattern, leading us to conclude our research is finding themes that are semantically distinct.

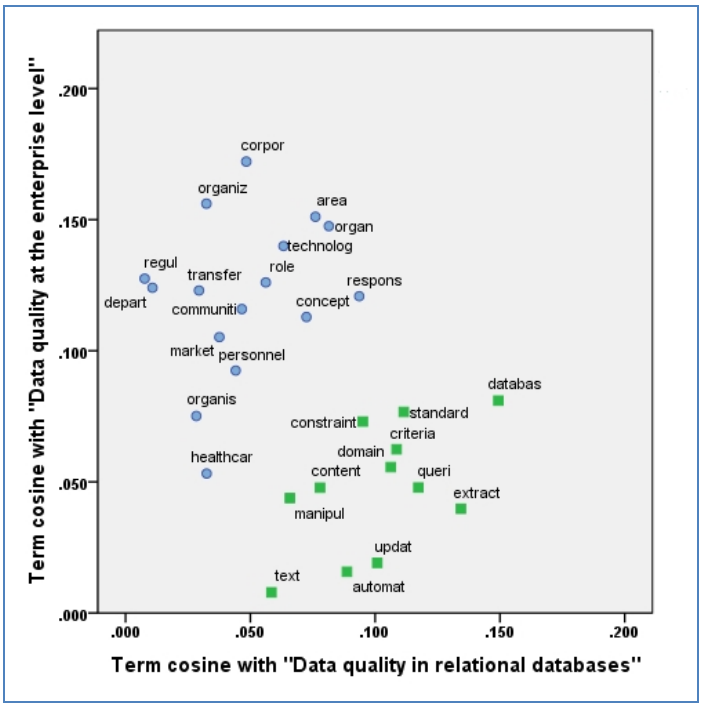

Figure 1: Plot of cosine similarity measures between terms and two research themes

One might reasonably expect that the dimensions of data quality would emerge as a separate topic or themes, but they do not. However, rather than disappearing, our analysis shows that of all the terms, terms related to dimensions were among those with significant loadings on the greatest number of factors: several loaded onto four or five of the fourteen themes. This may mean that the research focus has shifted from data quality dimensions *per se* to other topics. However, data quality dimensions continue to be an integral aspect of a majority of research. This is consistent with Madnick et al.'s [1] framework in which data quality dimensions do not appear either as explicit topics or as subtopics. Table 5 shows the high loading (loadings of 0.15 or above) themes for terms related to data quality dimensions.

	More instrinsic dimensions ⟷ More extrinsic dimensions							
	Accuracy	**Completeness**	**Consistency**	**Timeliness**	**Security**	**Accessibility**	**Believability**	**Interpretability**
1 Methods to Analyze DQ	√			√				
2 Systems for measuring DQ			√					
3 Economic Aspects of DQ	√	√						
4 Ontology and KM for DQ								
5 Prioritizing DQ Improvements	√		√					
6 Improving DQ in Applications	√	√						
7 Enterprise Data Quality		√		√	√		√	
8 Data Quality in Decision-Making	√		√					
9 Data Quality in Data Warehouses			√	√		√	√	
10 Data Quality in Relational DB	√		√	√	√	√		
11 Assessment of Contextual DQ	√		√	√				√
12 Data Quality for the WWW			√					√
13 Data Quality in Sensory Networks		√		√				

Table 5: Association between Dimensions and Themes

As might be expected, accuracy (8 of 14), consistency (8 of 14) and timeliness (6 of 14) were associated with the largest number of themes. Completeness was next (4 of 14). By examining the loadings presented in Table 5, we make the following observations. Accuracy and Timeliness were the two dimensions examined most in the first theme, "Methods for Analyzing Data Quality". Surprisingly, the only dimension to be examined under "Systems for Measuring Data Quality" is consistency. Although, a seminal work in data quality by Ballou et al. [15] did address accuracy and timeliness besides consistency, it appears that most other systems proposed for measuring quality do not address these dimensions. A possible explanation may be that measuring accuracy is challenging and both dimensions do have a significant contextual component. If so, most of these systems support the user gauging quality along these dimensions in context, but do not directly measure those dimensions of quality. Further, our identification of themes separates "systems" from "data repositories" – accuracy is a dimension examined by "Data Quality in Relational Databases" and timeliness is a key focus for data quality research dealing with databases and data warehouses.

Accuracy and Completeness are the two dimensions addressed by research dealing with "Economic Aspects of Data Quality". Timeliness and consistency are very similar to accuracy and completeness in terms of economic implications – it is surprising to see that this research theme has not treated timeliness and consistency the same way. While accuracy and consistency appear to be the focus of "Prioritizing Data Quality Improvements", accuracy and completeness appear to be the focus of "Improving Data Quality in Applications". "Enterprise Data Quality" appears to deal with completeness, timeliness, security and believability. "Data Quality for Decision-Making" deals only with accuracy and consistency – one would have expected this research theme to examine believability and interpretability as well. Not surprisingly, both themes, "Data Quality in Data Warehouses" and "Data Quality in Relational Databases" deal with consistency, timeliness, and accessibility. However, data warehouses also focus on believability (consistent with the fact that data warehouses are decision support environments) while databases focus on security and accuracy – dimensions that are important for transactional databases and less so for analytical repositories.

"Assessing Contextual Data Quality" examines accuracy, consistency, timeliness and interpretability. It is interesting that the first dimension has been regarded as very intrinsic while the last as very extrinsic. The fact that contextual assessment deals with both extremes informs us that all data quality dimensions have some dependency on context, some more dependent than, others. On a related note, "Data Quality for the Web" deals with consistency and interpretability reflecting the contextual importance of data quality for the consumers of web data.

It is also interesting that this research has identified a very small subset (shown in table 5) of the quality dimensions proposed in literature. In particular, the dimensions of objectivity, reputation, value-added, appropriate amount of data, relevancy, representational consistency, ease of understanding, and concise reputation are all dimensions proposed in seminal work [16][17] in data quality, but do not appear here. Does this mean that both practitioners and academics are interested in only the dimensions we see in Table 5? Does it imply that it is more difficult to measure/examine the other dimensions? Our next step is to seek answers to these questions. We also note that our current analysis did not associate dimensions to research methods or research methods to topics/themes. We are currently in the process of investigating these associations.

The results of this study show that we can identify logically consistent core topics and themes within data quality research, and that these topics and themes not only align with existing work, but also can offer a different perspective. This study also demonstrates a replicable, quantitative method for analyzing the core research topics and themes in this area. We believe that this is a powerful first step towards defining the identity of data and information quality research.

CONCLUSIONS AND FURTHER RESEARCH

In this paper, we have presented a preliminary study to identify core topics and themes of data quality research. We identified the topics and themes by analyzing the texts of abstracts from 467 journal and conference articles published primarily over the past ten years. We used latent semantic analysis to measure term-to-term semantic similarity, and we used those similarity measures to load terms onto factors. We identified five core topics and fourteen themes based on the terms that loaded heavily on factors. We compared the framework derived in this research with the framework proposed by Madnick et al. [1] for consistency of alignment. We further used cosine-similarity to gauge the extent to which our themes are distinct.

We further derived the mapping of data quality dimensions to the themes identified in our framework. The results offer very interesting insights into well-researched dimensions in data quality and the dimensions addressed within each theme. From this analysis, we are also able to draw attention to minimally addressed dimensions.

Our next step is to employ the method developed in this paper to produce document-to-document factor loadings. Doing so will enable us to evaluate the number of papers associated with each topic and theme, and to determine which topics and themes are being emphasized, and which are not. Prior to this step we are revisiting the abstracts in our corpus to identify any additional journal articles or conference proceedings that should be included.

Another step will be to compare the core topics, core themes, and paper counts over discrete time-periods. Combined with document-to-document factors loadings, this could be a viable way to measure changes and trends now and in the future, as researchers begin to focus on new directions.

We know that the landscape of data quality in research and data quality in practice is changing quickly. Fifteen years ago the term “data provenance” was not in use, state regulations regarding data security were unheard of, Federal Enterprise Architecture was not yet on the horizon, and the volume of unstructured non-transactional data was a miniscule portion of what it is today. All of these changes pose new research challenges. The method shown in this paper can assess how well researchers are meeting those challenges. In so doing, the method can continually help define the identity of data quality research as a distinct body of knowledge on an on-going basis, in step with the rapid changes occurring in this discipline.

REFERENCES

[1] Madnick, S., Wang, R. Y., and Lee, Y. W., "Overview and Framework for Data and Information Quality Research," *ACM Journal of Information and Data Quality*, vol. 1, 2009, pp. 1-22.

[2] Sidorova, A., Evangelopoulos, N., Valacich, J. S., and Ramakrishnan, T., "Uncovering the intellectual core of the information systems discipline," *MIS Quarterly*, vol. 32, 2008, pp. 467-482

[3] Neely, M. P. and Cook, J., "A Framework for Classification of the Data and Information Quality Literature and Preliminary Results (1996-2007)," *Americas Conference on Information Systems (AMCIS)*, Toronto, Canada: 2008.

[4] Ge, M. and Helfert, M., "A Review of Information Quality Research," *International Conference on Information Quality*, Cambridge, MA: 2007.

[5] Lima, L., Maçada, G., and Vargas, L.M., "Research into information quality: A study of the state-of-the-art in IQ and its consolidation," *International Conference on Information Quality*, Cambridge, MA: 2006.

[6] Mason, R. and Mitroff, I., "A program for research on management information systems," *Management Science*, vol. 19, 1973, pp. 475-487

[7] Benbaset, I. and Zmud, R., "The identity crisis within the IS discipline: Defining and communicating the discipline's core properties," *MIS Quarterly*, vol. 27, 2003, pp. 183-194

[8] Cummings, S. and Daellenbach U., "A Guide to the Future of Strategy: The History of Long Range Planning," *Long Range Planning*, vol. 42, 2009, pp. 234--263.

[9] Stotesbury, H., "Evaluation in research article abstracts in the narrative and hard sciences," *Journal of English for Academic Purposes*, vol. 2, 2003, pp. 327-341

[10] Albert, S. and Whetten, D., "Organization identity," *Research on Organizational Behavior*, vol. 7, 1985, pp. 263-295

[11] Wang, R. Y., Storey, V. C., and Firth, P., "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, 1995, pp. 623-640.

[12] Laham, D., "Latent Semantic Analysis Approaches to Categorization," *Components*, 1997, pp. 80309-80309.

[13] Landauer, T. K., Foltz, P. W., and Laham, D., "Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, 1998, pp. 259-284.

[14] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A., "Indexing by latent semantic analysis," *Journal of the Society for Information Science*, vol. 41, 1990, pp. 391-407.

[15] Ballou, D. P. and Pazer, H. L., "Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff," *Information Systems Research*, vol. 6, 1985

[16] Lee, Y. W., Pipino, L. L., Funk, J. D., and Wang, R. Y., *Journey to Data Quality*, The MIT Press, Cambridge, MA, 2006.

[17] Wang, R. Y. and Strong, D. M., "Beyond Accuracy: What Data Quality Means to Consumers," *Journal of Management Information Systems*, vol. 12, 1996, pp. 5-34.

[18] Juran, J. M. and Godfrey, A. B., *Juran's Quality Handbook,* McGraw Hill International Editions: Industrial Engineering Series, 5th Edition, September 2000.