Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2010 Proceedings

Americas Conference on Information Systems (AMCIS)

8-2010

Identifying the core topics and themes of data and information quality research

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

Identifying the core topics and themes of data and information quality research

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

ABSTRACT

As data and information quality research has evolved towards becoming a unified body of knowledge, the importance of defining the identity of this research area has also grown. Our paper presents the results of a preliminary study with the aim of helping to define this identity from its core topics and themes. To do so we analyze the abstracts of 324 journal articles and conference proceedings published over the past ten years. Latent semantic analysis is used with these abstracts to develop term-to-term semantic similarities and term-to-factor loadings, from which six core topics and fifteen core themes of data quality research are identified.

This paper presents a quantitatively based and reproducible method for identifying topics and themes. In further research, this method will be used to analyze the frequency of papers published in each topic and theme to show their level of activity. Applying this method for publications within discrete periods of time can show how topics and themes change and evolve. This research has the potential to help define the identity of data and information quality research, find the topics and themes receiving the greatest attention, and reveal trends occurring in this growing area.

Keywords

Data quality research, information quality research, research frameworks, latent semantic analysis

INTRODUCTION

In the editorial of the inaugural issue of the ACM Journal of Data and Information Quality the senior editors noted that data and information quality research has shifted from research that is spread through several different reference disciplines towards a more unified body of knowledge. As this trend continues and possibly accelerates, it becomes important to establish the identity of data quality as a distinct area of research. (Please note that as others have done, we use "data quality" as a more succinct term meant to encompass both data quality and information quality.)

One primary motivation for our research is take a step towards defining that identity by identifying its core topics and themes. Another motivation is to develop a method that can be reproduced to evaluate how that core changes over time.

We analyze the abstracts of journal articles and conference proceedings considered to have data quality as the primary focus. Abstracts have been used as the primary source of data to develop research topics in a diverse range of fields including business strategy (Cummings et al. 2009) and the sciences (Stotesbury 2003).

For our analysis we use Latent semantic analysis (LSA), a statistical method for finding semantic relationships within a corpus of documents. LSA uses the context in which terms appear to measure term-to-term and document-to-document semantic similarities. This is quite different from an analysis of either keyword frequencies or citation counts; these can require literal, or near literal, matches between terms whereas LSA does not. LSA has been successfully used to predict the subjective ratings of essays made by human readers, to match human categorizations of terms (Laham 1997), and measure textual coherence (Landauer et al. 1998).

The literature review in the next section includes prior work, which defines what identities are for research areas and existing frameworks of data quality research. Following the literature review are sections for methodology, results, and finally a summary with directions for further research.

Literature review

Seeking the identity of an area of research is not new. From almost forty years ago when Mason and Mitroff (1973) published a program for information systems (IS) research, through more recent "crises of identity" (Benbasat et al.

2003), 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.

Lima, Maçada and Vargas (2006) saw the ambiguity of an identity and the relation to reference disciplines as parallels between the IS discipline and data quality research. This is an important motivation for our study.

Albert and Whetten (1985) 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, and evaluating how these core topics and themes change over time can support the claim of temporal continuity.

More recently Sidorova, Evangelopoulos, Valacich and Ramakrishnan (2008) sought to define the identity of the IS discipline within five core areas. These were information technology for organizations, individuals, markets, and groups for the first four, and information systems development for the fifth. Research themes were also identified within each core area. Sidorova et al. concluded that the core areas of IS research have remained stable, and the underlying themes within those core areas have continued to evolve.

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 in this regard. Our work can complement these but with a focus on data and information quality research.

One of the earliest frameworks of data quality was developed by Wang, Storey, and Firth (1995) from a comprehensive analysis of publications which appeared through 1994. 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.

Beyond profiling and categorizing more than 125 papers in their framework at an early point of the evolution of data quality research – the first International Conference on Information Quality was not held until the year after the framework was published – the authors pointed to specific research challenges. Calls were made to research the economics of data quality, standardization of data quality metrics, and effects of data quality policies. Because the method we develop can be used to analyze publications over discrete periods of time, our research can help measure how well the calls for research were answered.

Lima et al. (2006) developed conceptual maps of data quality research using articles published from 1995 through 2005. For this purpose they chose proceedings from conferences of central interest to the data quality research community as their primary data. Of the 171 proceedings reviewed by Lima et al., 86% were from the Internal Conference on Information Quality (86%), 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 a highly detailed conceptual map of these keywords was developed specifying relationships between keywords and their groupings.

Lima et al. established relationships between keywords using their own judgment. Latent semantic analysis is often used to reach the same ends; in many cases it is known to make judgments similar to those made by humans. Frameworks are sometimes accompanied by categorized papers; this is another use for which LSA performs well. Lima et al.'s framework of 279 keywords was much more detailed than the topics and themes developed in our study. However, the research topics and themes from our study can be readily increased to a very substantial number of sub-themes, affording a comparison to Lima et al.'s framework.

A comprehensive review by Ge and Helfert (2007) divided research into that focusing on the assessment, management, and contextual aspects of data quality. Data quality assessment was covered in particular depth, and was further divided 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.

Using studies with a focus on how data quality problems are identified, Ge and Helfert classified those problems using a two-by-two grid. Problems were classified as being either dependent on, or independent from, a specific context. Problems were additionally classified as having either a perspective related to data, or a perspective related to users. Examples of context independent problems from a data perspective would include missing or incorrect data; from a user perspective, these problems included inaccessible or insecure information. A context independent problem from a data perspective might be from a database in a state inconsistent with business rules; from a user perspective problems would include semantically ambiguous information.

Research of data quality dimensions was categorized by topic and methodology. Topics were classified by their primary focus as identifying, defining, or classifying data quality dimensions, or as analyzing their interdependencies. Methodologies were classified as being intuitive, theoretical, or empirical. Within the major category of data quality assessment, several well-known assessment methodologies were profiled by five attributes: how they defined data quality dimensions, how they classified those dimensions, their underlying assessment model (such as PSP/IQ), the tools they employed, and their case studies. A related framework was developed for classifying assessment methodologies by their specific features, such as their processes, tools, and standards, and how they benchmarked data quality whether measured objectively or subjectively.

Ge and Helfert's review and conceptual models identify and categorize research topics, and relate to the goal of our study. A 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. However, the method we develop can complement frameworks and conceptual models, and comparing both works could be helpful in defining the identity of data quality research.

A recent framework developed by Madnick, Wang, Lee, and Zhu (2009) used two dimensions, topics and methods, to categorize data quality research. Research methods in particular were specified at varying levels of granularity; some methods could be considered subsets of others, such as statistical analysis as a type of quantitative method.

Neither topics nor research methodologies were taken as 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, which are shown in Table 1, researchers can characterize their own research.

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 householding	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)

Madnick et al. noted that the quality of structured data which has received a considerable amount of attention in the past has been receding in favor of studying the quality semi-structured and unstructured data. They defined several research challenges, some still similar to the calls from Wang et al.'s earlier 1995 work. Among the challenges seen in both 1995 and 2009 were for research to find new techniques to manage and deliver quality data. In an active area of research it can be expected that some challenges will be met as others arise. This is certainly true of data quality research; the need to develop more rigorous data quality metrics and measures as seen in 1995 did not receive mention as a challenge by Madnick et al. in 2009. Conversely, the 2009 call to respond to rapidly changing expectations and views on the part of users, and the need for data quality research to go beyond the perspective of single individuals and organizations to scopes including cultures, groups, or societies; this was not a consideration in 1995.

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

METHODOLOGY

The abstracts chosen to build our corpus were from the proceedings of conferences central to the data quality research community, and from journal articles with data quality as their primary focus. The conferences selected included AMCIS and the International Conference on Information Quality; journal articles were selected from the Journal of Management Information Systems and the ACM Journal of Data and Information Quality. The counts of abstracts from outlets are shown in Table 2.

Outlet	**Abstract Count**
International Conference on Information Quality (ICIQ)	166
International Journal of Information Quality (IJIQ)	33
AMCIS	32
International Workshop on IQ in IS	18
Advances in Management Information Systems	14
ACMJDIQ	7
Journal of Mgt Information Systems (JMIS)	6
Others	48
Total	324

Table 2. Count of abstracts used to identify core topics and themes

Several routinely used pre-processing steps were applied to the text of the 324 abstracts in our corpus prior to latent semantic analysis. In the first, punctuation and numeric values were removed. The second step was to remove stop words. Stop words are short, commonly occurring words that are not useful for the analysis, such as "a", "I", and "the". Several words and phrases with high frequency in the corpus were removed for the same reason. To no surprise, these included the terms "data quality" and "information quality". In the third step all words in the corpus were stemmed. Stemming standardizes words having multiple variations with semantically equivalent meanings. Often these are words with the same root with multiple suffixes. For example, stemming might transform "decide", "deciding", and "decides" all to the same root term "decid". For our analysis we used the Snowball stemmer, which is an implementation of a popular stemming algorithm.

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 matrix had 636 terms and 309 documents (several of the 324 abstracts were too brief to be included). 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 deemphasizes 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. In 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.

SVD is used 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. *S* contains the square roots of the eigenvectors from SVD in sorted order; if *T*, *S*, and *D* are multiplied they will approximate the original term-document matrix. A *t* by *d* matrix of similarities between terms can be found from the dot products of *TSD'*, an analogous matrix can be generated to find the similarities between documents. A much more detailed description of LSA, SVD, and similarity measures is provided by Deerwester, Dumais, Furnas, Landauer and Harshman (1990).

For the analysis procedure we used R version 2.10.1 and the LSA, Snowball, and RStem packages. This software is from the R Foundation and is available at http://www.r-project.org/. Our preliminary results are shown next.

RESULTS

LSA was used to find semantically related terms and their loadings onto individual factors. Several different combinations of factors were evaluated to find the most logical groupings of topics and themes. Terms with factor scores lower than .10 were ignored, as were words with no specific meaning to research topics and themes. For example, the terms "article" and "paper" were among those discarded. Terms such as "empir" and "statis" are related to methodology and might interesting in an extension of this study, but for our purposes here they too were discarded. Relatively few terms had significant loadings on multiple factors. In the fifteen-factor analysis there were 32 terms (5.0%) loading on three or more factors, and 161 (25.3%) loading on two factors, using a cutoff factor score of .20. This is consistent with other results of the analysis showing that 138 terms (21.7%) explained 80% of the total variance and 186 terms (29.2%) explained 90%.

Six factors for topics and fifteen factors for themes were chosen as having the most meaningful representations. Terms with high loadings most relevant to each topic and to each theme are shown in Table 3. Please note that the terms are words as transformed by stemming, and not necessarily the original word as it appeared in the abstracts.

Core topics and themes		**Relevant terms loading onto themes**
1 Data quality assessment	1 Fitness for use	Suitabl, fit, satisfi, expect, requir, measur, model
	2 Metrics	Featur, express, aggreg, define, paramet, metric, specifi
2 Management of data quality	3 IT Management	Prevent, profession, chang, hierarch, network, employ, domain
	4 Data manufacturing	Product, manufactur, investig, infrastructure, recogn, secur
	5 Operations	Output, input, report, process, uncertainti
	6 Data quality improvement	Architec, institut, implement, function, accuri, negoti, project, product
3 The impact of data quality at organizational levels	7 Enterprise level	Encompass, corpor, strateg, compani, success, demand, organiz, market, enterpris, build, wide, interdepend
	8 Stakeholder level	Environ, busi, comprehens, resourc, stakehold
4 Data quality and databases	9 Database design and data mining	Algorithm, cluster, dataset, datamin, entiti, object, rule, attrib, pattern, linkag, score
	10 Querying and cleansing	Extract, manipul, storag, dirtydata, system, queri, transform, field, clean

	11 Data integration	Constraint, schema, sourc, solut, integr, specifi, monitor, view, semantic
5 The impact of data quality on decision making	12 Decision making	Understand, decisionmak, learn, effort, reflect, impact, issu, accuri
	13 Economic impact	Cost, benefit, tradeoff, optim, util, feasibl, polici, valu, estim, economic, dss
6 Data quality application areas	14 Internet-related	Web, webpage, publish, portal, transact, user, design, internet, site
	15 Business applications	Practic, community, interdepend, datawarehous, crm, applic, network, develop

Table 3. Relevant terms in the loadings onto core topics and themes

The results shown here do not include a planned analysis to find the specific papers, and then the number of papers, which load onto each factor. When achieved we will be able to determine the activity level for each core topic and theme. When repeated for discrete intervals of time, the analysis will reveal trends in data quality research. This is an importance continuation of the research presented in this paper.

There are clearly similarities and differences between these core topics and themes and existing frameworks. While organized differently from our topics and themes, frameworks often have categorizations of data quality assessment, data quality management, and data quality in databases. However, several differences can be noted.

The importance of data curation is evident by its placement as one of the four major topics in Madnick et al.'s (2009) framework. In our study data curation did not emerge either as a topic or as a theme. Even the term "curation" did not have enough global occurrences in our base of abstracts to reach the threshold for inclusion in the latent semantic analysis. We speculate that the increasing interest in data curation may be a more recent development, accounting for its absence in relatively recent abstracts and therefore in our results, but this is a subject for further investigation.

Also mentioned by Madnick et al. was the importance of data quality research with a scope beyond that of single individuals and organizations. Terms such as "market", "interdepend", and "encompass" in the Enterprise Level theme may reflect recognition of this importance, as could the "Stakeholder" theme on which high-loading terms included "environ" and "resource".

One might reasonably expect that the dimensions of data quality would emerge as separate topics or themes, but they do not. However, rather than having disappeared, 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 fifteen themes. This may mean that a focus on data quality dimensions has shifted towards a focus on other topics but with data quality dimensions remaining an integral and important aspect. This is consistent with Madnick et al.'s framework in which data quality dimensions do not appear either as explicit topics or as subtopics. Table 4 shows the high loading (factor scores .20 or above) themes for terms related to data quality dimensions.

Dimension	**High loading research themes**
Accuracy	Database design and data mining, Operations, Decision-making, Metrics
Completeness	Fitness for use, Economic impact, Metrics
Consistency	Operations, Enterprise level, Metrics
Currency	Fitness for use, Economic impact, Data quality improvement
Timeliness	Fitness for use, Decision-making, Economic impact, Data quality improvement
Security	Fitness for use, Data Manufacturing
Representation	Database design and data mining, Decision-making, Internet, Metrics
Accessibility	Fitness for use, Data Manufacturing, Internet, Business applications, Metrics

Believability	Data manufacturing, Fitness for use, Decision-making, Metrics
Relevancy	Fitness for use, IT Management, Data Manufacturing
Reliability	Fitness for use, IT Management

Table 4. Themes with high loadings on data quality research themes

As might be expected, most dimensions had high loadings on the "Fitness for Use" theme. It also might be expected those dimensions considered as intrinsic dimensions, such as accuracy and completeness which can be more objectively measured, generally had high loadings on the "Metrics" theme, but extrinsic dimensions such as believability and relevancy did not. Perhaps less expected was that the themes of "Decision-making" and "Economic impact" were more closely associated with intrinsic rather than extrinsic dimensions. When the document-term loadings are completed a more detailed analysis of the relation between dimensions and research themes can be conducted.

The results of this preliminary study show that logically consistent core topics and themes can be identified for the area of 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 quantitative and method of analyzing data and information quality research that can be replicated. This study and the steps outlined in the next section can help to define the identity of data quality research.

SUMMARY AND NEXT STEPS

This paper presents a preliminary study to identify core topics and themes of data quality research. The resulting topics and themes are based on analyzing the texts of abstracts from 324 journal articles and conference proceedings published over the past ten years. Latent semantic analysis was used to measure term-to-term semantic similarity, and by these measures terms were loaded onto factors. High-loading terms were used to identify six core research topics and fifteen core themes comprising those themes. These topics and themes were compared for alignment with an existing framework.

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 therefore 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 find if any additional journal articles or conference proceedings should be added.

Another important step will be to compare the core topics, core themes, and paper counts over discrete periods of time. Combined with document-factors loadings, this could be a viable way to measure changes and trends, both now and at points in the future. In a related direction, from preliminary work it appears a similar analysis of core research methodologies is feasible; this too could produce meaningful results.

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 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 those challenges are being met, and in so doing contribute to defining the identity of data quality research as a distinct body of knowledge.

BIBLIOGRAPHY

Albert, S., and Whetten, D. "Organization identity," *Research on Organizational Behavior* (7) 1985, pp 263-295.

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

Cummings, S., and Daellenbach, U. "A Guide to the Future of Strategy?:: The History of Long Range Planning," *Long Range Planning* (42:2) 2009, pp 234-263.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. "Indexing by latent semantic analysis," *Journal of the American society for information science* (41:6) 1990, pp 391-407.

Ge, M., and Helfert, M. "A Review of Information Quality Research," International Conference of Information Quality, Cambridge, MA, 2007.

Laham, D. "Latent Semantic Analysis approaches to categorization," The 19th annual conference of the Cognitive Science Society, 1997.

Landauer, T.K., Foltz, P.W., and Laham, D. "An introduction to latent semantic analysis," *Discourse processes* (25) 1998, pp 259-284.

Lima, L., Maçada, A., and Vargas, L. "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.

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

Mason, R., and Mitroff, I. "A program for research on management information systems," *Management Science* (19:5) 1973, pp 475–487.

Sidorova, A., Evangelopoulos, N., Valacich, J.S., and Ramakrishnan, T. "Uncovering the intellectual core of the information systems discipline," *MIS Quarterly* (32:3) 2008, pp 467-482.

Stotesbury, H. "Evaluation in research article abstracts in the narrative and hard sciences," *Journal of English for Academic Purposes* (2:4) 2003, pp 327-341.

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