Data Quality/ Information Quality Research: Research Literature Framework and Evolution

(Target Journal: MISQ – review section paper)

Wayne Huang et al. (Ohio University)

Abstract

Over the past three decades, data quality/ information quality (DQ/ IQ) is emerging into a possible distinct discipline. As the research overlaps with other disciplines or research fields such as IS, Marketing, Computing Science, etc., it is important to identify the core characteristics of DQ/IQ research and to study its development over time. Although scholars have make contribution to the identity of DQ/ IQ research through qualitative and quantity approaches, there is lacking of a more objective approach that comprehensively studies the identity and evolution of DQ/ IQ research.

In this study, Latent semantic analysis (LSA) approach was used to identify the core areas and evolution of DQ/IQ research field. Relevant keywords from selected 317 journal papers and conference proceeding papers during 1976 through 2012 were analyzed. We identified five core research areas of DQ/ IQ that have emerged from the research literature in the last 36 years: (1) assessment of DQ/ IQ; (2) computing and technological aspect of DQ/ IQ; (3) DQ/ IQ system application; (4) organizational level impact of DQ/IQ; (5) data process management of DQ/ IQ. By examining the evolution of DQ/ IQ research over the past 36 years, we found that the core areas have remained stable, but the topics within each core area appeared and disappeared over time. We conclude that DQ/IQ research has remained relatively stable by focusing on the DQ/IQ research cycle of data/ information management: technology → application → process management → assessment → impact. Insights and suggestions are discussed and presented finally for future research.

1. Introduction

1. INTRODUCTION: Identify the research gap and your research purposes. Justify why they are important to our community.

Over the past three decades, DQ/IQ has become an important research field that has made significant contribution to the IS academic research field. There are many research topics in this field, including technical oriented topics such as database related technical solutions for data quality, behavior oriented topics such as data quality impact, web and environment oriented topics such as data quality in the context of computer science and IT [1]. According to the senior editors of the *ACM Journal of Data and Information Quality*, after developing from different relevant disciplines, DQ/IQ research has been shifting to a new emerging distinct discipline of IS. However, as the research field overlaps with other disciplines or fields, it is very important to identify the core characteristics of DQ/IQ and to study its development over time. It is the strong and distinct identity that makes a continuous development and success of the discipline [2, 3].

Although scholars have make contribution to the identity of DQ/ IQ research through qualitative and quantity approaches, there is lacking of a more objective approach that comprehensively studies the identity and evolution of DQ/ IQ research. This research focuses on the identification of core topics and themes of DQ/IQ field. In addition, we aim to trace the evolution and development of DQ/ IQ research over the past 36 years. In this research, we applied latent semantic analysis (LSA) to identify the core areas and evolution of DQ/IQ field by analyzing the research keywords of 317 full-text DQ/ IQ research papers that were searched and selected through a structured approach as recommend by Webster and Watson (2002) [4].

The rest of this paper is organized as follows. In the next section, we review the current literature from three aspects: topics, frameworks and identities. Then, we describe the application of LSA and data collection process. Thirdly, results are analyzed. At last, conclusion of key findings, limitations were discussed and directions for further research were suggested.

2. Literature Review

The topic of identification of disciplines has been discussed for a long time. Sidorova et al. (2008) used LSA to determine the intellectual core of the IS discipline and examined the evolution of IS research over time, and posit IS

among the business disciplines. They identified five core research areas: (1) information technology and organizations; (2) IS development; (3) IT and individuals; (4) IT and markets; and (5) IT and groups. In addition, the five core areas remained quite stable, but the specific research themes in the core areas have changed greatly over time [5].

Scholars have made contribution to the identity of DQ/ IQ research through qualitative and quantity researches on categorizing/clustering DQ/ IQ's research dimensions, topics, and themes. Wang et al. (1995) did a qualitative research to analyze DQ research and developed a framework for DQ research by analogizing product manufacturing and data manufacturing adapted from ISO 9000 [6]. The framework includes seven elements: (1) management responsibilities; (2) operation and assurance costs; (3) research and development; (4) production; (5) distribution; (6) personnel management; and (7) legal function. However, the seven elements are not sufficient to characterize DQ/ IS research and are difficult to be used [1].

Through the qualitative research of 171 papers published form 1995 to 2005, Lima et al. (2006) came up with the conclusion that the relationship between IS and DQ/ IQ research is parallel, and then developed the conceptual maps of DQ research with three high-level perspectives of DQ research: the organizational, behavioral, and operational views, which have detailed conceptual map of relationships between keywords and clusters for each perspective [7].

Ge and Helfert (2007) did a comprehensive review of information quality research review, proposing three major areas of DQ/ IQ research: information quality assessment, information quality management and contextual information quality, which qualitatively built a framework of DQ/ IQ research and emphasized the characteristic of this research field [8].

Madnick et al. (2009) qualitatively introduced a framework for DQ/ IQ research along two dimensions: topics and methods, and provided several representative papers to illustrate relevant topics and methods [1]. The authors also provided intuitive and commonly used keywords to describe each topics, which helps people to characterize individual papers and determine which topics it will belong to.

Recently, Blake (2010) did a quantitative research to identify core topics and themes on DQ/ IQ. He analyzed the abstracts of 324 data quality and information quality papers, and used LSA to analyze all the abstracts of these articles. The author concluded that this research consists of six core topics and fifteen core themes of data quality and information quality research [9]. Blake (2010) also pointed out that future research should focus on the evolution of DQ/ IQ research over time [9].

Although the research on the identity of DQ/ IQ research has made contributions to DQ/IQ research, there are several limitations of the former research. First of all, the qualitative researches objectively lead, determined, or recommended the structure, core areas, topics, and methodology of DQ/ IQ research through manual statistical analysis, which may neglect the subject perspective that start from the mathematical analysis of published paper. However, quantitative research using LSA can provide a different perspective to look into the DQ/ IQ research [9]. Secondly, the recent quantity researches of DQ/ IQ are not sufficient; for example, the evolution of DQ/ IQ research over the past thirty years should be studied to find the trend of this field [9].

The next section introduces the methodology of LSA to identify core DQ/ IQ areas and corresponding themes through a quantitative way.

3. Methodology

3.1. Introduction of LSA

The main idea behind LSA is to collect all of the contexts within which words appear, and to establish common factors that represent underlying concepts [5, 10]. The main purpose of LSA is the reduction of dimensionality of original data through singular value decomposition (SVD). SVD is a form of factor analysis applied to a t by d term-document matrix.

In order to process the unstructured data efficiently in LSA, it has to be transformed into a structured form (i.e., vector/matrix). The normal method is the term frequency and inverse document frequency (TF-IDF) transformation. TF-IDF transformation or quantification is a term-frequency based mechanism by which a sequence of words (e.g. a sentence, an article) can be transformed into a structured quantified vector. The raw term frequencies are refined to the product $w_{ij} = tf_{ij} * idf_i$, where $idf_i = \log_2(N/n_i) + 1$, where N is the number of documents in the collection, tf_{ij} is the raw term frequency of term i in document j, n_i is the term frequency of term i in the entire collection of documents, and the inverse document frequency (IDF) idf_i serves as a metric of rarity of term i in the entire collection of documents. Thus, the occurrence of rare terms is promoted and the influence of more common non-stop words is discounted. Then,

the term frequencies are typically normalized so that the sum of squared transformed frequencies of all term occurrences within each document is equal to one. Hence, a document can be represented as a vector of term frequencies.

However, the original vectors of term frequencies are usually highly dimensional. Singular value decomposition (SVD) is utilized to reduce the dimension of the original document vectors. By SVD, a $t \times d$ matrix $\mathbf{X}$ can be expressed as $\mathbf{X} = \mathbf{TSD}^T$ where $\mathbf{T}$ and $\mathbf{D}$ are both orthogonal and S is diagonal. If X is a $t \times d$ matrix of terms by documents containing the raw or weighted term frequencies, $\mathbf{TS}$ is called the factor loadings for terms and $\mathbf{DS}$ is the factor loadings for documents. Interestingly, $\mathbf{T}$ is the $t \times r$ matrix of eigenvectors of $\mathbf{XX}^\mathbf{T}$ (term covariance), $\mathbf{D}$ is the $d \times r$ matrix of eigenvectors of the square symmetric matrix of $\mathbf{X}^\mathbf{T}\mathbf{X}$ (document covariance), and $\mathbf{S}$ is an $r \times r$ diagonal matrix containing the square roots of eigenvalues of both $\mathbf{XX}^\mathbf{T}$ and $\mathbf{X}^\mathbf{T}\mathbf{X}$, where r is the rank of $\mathbf{X}$. By retaining a proper number of significant factor k, X can be represented by its least squares approximation $\mathbf{X}' = \mathbf{T}_k\mathbf{S}_k\mathbf{D}_k$, which not only reduces the dimensions of some corresponding matrices but also filters some noises in the original representation.

In order to promote the significance of more important terms and discount the significance of less important terms, factor rotation technique is often adopted to achieve this. Usually, the term loadings $\mathbf{L}_\mathrm{T} = \mathbf{T}_k\mathbf{S}_k$ is rotated into $\mathbf{L}_\mathrm{T}\mathbf{M}$ and the document loadings $\mathbf{L}_\mathrm{D} = \mathbf{D}_k\mathbf{S}_k$ is rotated into $\mathbf{L}_\mathrm{D}\mathbf{M}$ by multiplying them by a rotation matrix $\mathbf{M}$ according to some term structure simplification. A representative factor rotation method is called the varimax rotation [5].

3.2. Data Collection

To build our corpus, the keywords are chosen from relevant DQ/ IQ journal articles and proceedings of conferences central to the data quality research using a structured approach recommend by Webster and Watson (2002) [4].

(1) We use “Information Quality”, “Data Quality”, “Quality of Data”, and “Quality of Information” as keywords to search papers published on the leading journals in IS discipline before July 2012 in date. Journal databases ABI/Inform (ProQuest) was used to identify relevant articles in the top five leading journals from MIS Journal ranking (2011) by Association for Information Systems, including MISQ, ISR, CACM, MS, and JMIS. In addition, we reviewed all the available papers on proceedings of International Conference on Information Quality (ICIQ) from 1995 to 2007, which is the only international conference specifically targeting DQ/ IQ field with a reputation for quality.

(2) We go backward by reviewing the citations for the articles identified in step 1 including Information Quality framework and/ or instrument to determine prior articles that we should consider and were not included in the step 1.

(3) We go forward by using Web of Science to identify articles citing the key articles identified in the previous steps to determine which of these articles should be included in the review.

Through these three steps, we have collected 317 full-texts DQ/ IQ related papers.

3.3. Pre-processing

LSA begins with treating keywords of each article as a set of unstructured words. The text of the keywords of 317 articles were processed with several routinely used pre-processing steps such as stop words removed and stemming prior to latent semantic analysis [11].

Stop words. After numeric values and punctuation were removed, the following step is to remove stop words. Stop words are usually meaningless and not useful for the analysis.

Stemming. Stemming can transform these words that share a same root term to a related form in order to increase identification of similar words and avoid redundancy to reduce processing time at the same time. For example, stemming might transform “contribute”, “contribution”, “contributing”, “contributes” and “contributed” all to the same root term “contribute”. For our analysis we used the Snowball stemmer, which is an implementation of a popular stemming algorithm.

4. Result

4.1. An Overview of Different Factor Solutions

The result of LSA analysis was used to find semantically related terms and their loadings on each factor. Just like general factor analysis, each factor was related to its high loading terms and documents, which are both necessary to interpret the corresponding factor. For each factor, high loading terms and documents that are sorted by absolute loading value are selected to help interpret and label the factor. All factor solution can be proved meaningful by a reasonable

factor interpretation. Examination of different factor solution suggests that the body of DQ/ IQ can be aggregated at different level, in other words, different factor solution reveals different aspect of the DQ/ IQ discipline.

Combined with the reference of earlier empirical studies and the result of different factor solution, we finally chose five factors for topics and fifteen factors for themes in order to have the most meaningful representations.

A meaningful view of the intellectual core of the DQ/ IQ discipline are provided by five research topics, and each topic can be represented through several research themes. Four of the five factors containing the solution can be interpreted as the business or organizational problem-oriented research, examining the application, assessment or impact of one or some specific research issues; they include (1) assessment of DQ/ IQ, (3) DQ/ IQ system application, (4) organizational level impact of DQ/IQ, and (5) data process management of DQ/ IQ. DQ/ IQ system application examines the specific applications in the practice of DQ/ IQ research; from its corresponding research themes (see Table 1), this research topic focuses primarily on the improvement of data quality in systems, entity resolution application, corporate householding and supply chain, entity resolution application. Data process management of DQ/ IQ examines how to process information and data effectively in the management of DQ/ IQ. Here, the research themes focus on DQ/ IQ standardization, measure tool and processes management. At the next level, the assessment of DQ/ IQ examines how to assess the quality of information and data in practice. Here, research focuses on information quality measurement model, metrics and assessment, user satisfaction and IS success. Last, the research on the organizational level impact of DQ/IQ focuses on the implications of DQ/ IQ research for organizations, such as the effect of DQ/ IQ on business processes, and the impact of DQ/ IQ investments on organizational performance. Research at this level has also focused on business practices related to the impact of DQ/ IQ, such as security and trust, as well as the cost-benefit analysis of DQ/ IQ.

While these four research areas are business-oriented, the fifth distinct research area, (2) computing technology of DQ/ IQ is technology-oriented. It examines the computing technology itself, and how it processes to improve DQ/ IQ. Here, the research tends to be more technical in nature, largely focusing on the following themes, including entity identification, database and programming method, DQ/ IQ in network analysis technology, as well as data integration and cleansing.

According to the interpretations of the five research topic, the intellectual core of the DQ/ IQ discipline can be summarized into business-oriented research and technology-oriented research, which specifically are the application, process management, assessment and impact of specific DQ/ IQ issues in practice, as well as how computing technology is processed and applied on DQ/ IQ.

In the following section, we will discuss how the topics and themes of DQ/ IQ research have varied over time.

4.2. Dynamics of IQ and DQ research

To examine the dynamics of DQ/ IQ research, we account the total number of publications in different period of time in 15 research topics identified above and calculate the proportion of each topic in corresponding period. The time periods are chosen as 1990- 1999, 2000- 2009, and 2010- 2012 (as our data collection was ended in 2012). This analysis indicates the focus and popularity of changing research topics in different time periods, which provides an evidence of the evolution of DQ/ IQ research. The result is in Table 1.

Figure 1 examines the proportion of total number of five areas' publications and Figure 2 examines the proportion of each theme. (1) Assessment of DQ/ IQ accounts for the largest attention of researchers, and (3) DQ/ IQ system application ranks the second. Figure 3 shows the proportion of each area in different time period, which indicates the popularity of each area in different times. We can see from Figure 3 that most of the five areas remain constant over 30 years. Specifically, the proportion of (3) DQ/ IQ system application and (4) organizational level impact of DQ/IQ grew steadily over time; (1) assessment of DQ/ IQ and (5) data process management of DQ/ IQ first grew from 1990s to 2000s, but fell down slightly in 2010s. The last research area, (2) computing technology of DQ/ IQ experienced a dramatic fluctuation over time, however. It fell from 1990s to 2000s sharply and grew significantly when time came to 2010s.

Table 1. Areas and Topics in Different Times

Area	Topic	1990s	2000s	2010s (part of)
1.assessment of DQ/ IQ	user satisfaction and IS success	0.12	0.08	0.10
	information quality measurement model	0.12	0.18	0.10
	metrics and assessment	0.04	0.06	0.00
2.computing technology of DQ/ IQ	entity identification technology	0.04	0.00	0.04

	database and programming method	0.04	0.05	0.10
	DQ/ IQ in network analysis technology	0.12	0.06	0.08
	data integration and clean	0.12	0.00	0.05
3.DQ/ IQ system application	data quality improvement in systems	0.04	0.00	0.05
	corporate householding and supply chain	0.04	0.14	0.09
	entity resolution application	0.12	0.05	0.10
4.organizational level impact of DQ/IQ	security and trust	0.00	0.04	0.00
	cost-benefit analysis	0.08	0.07	0.05
	operation and decision system	0.04	0.08	0.14
5.data process management of DQ/ IQ	measure tool and processes management	0.12	0.15	0.10
	DQ/ IQ standardization	0.00	0.05	0.00

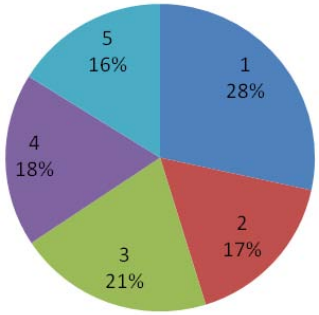

Figure 1. Ratio of Each Area

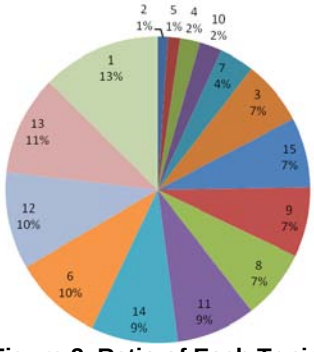

Figure 2. Ratio of Each Topic

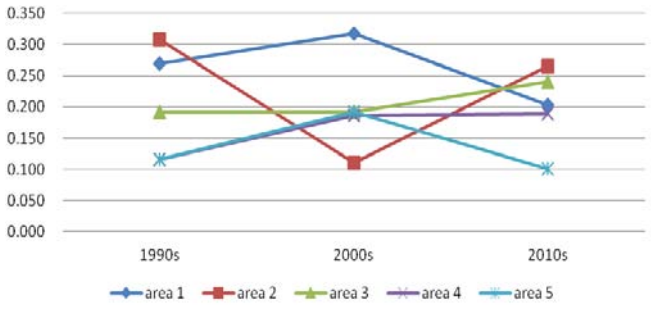

Figure 3. Dynamics of Research Areas

Examination of cross-loadings of 5 research areas and 15 research topics shows that the focus topics in some research area in different time period changes (see Table 1). For example, in (2) computing technology of DQ/ IQ, database and

programming method was the focus topic in 1990s and 200s, and was replaced by database and programming method in 2010s. In (3) DQ/ IQ system application, the topic of corporate householding and supply chain was not popular in 1990s, but gained the largest focus in 2000s. Similarly, operation and decision system in (4) organizational level impact of DQ/IQ gained continuous increasing attention over time and has grown into the focus topic of this area. Implications of significant changes in the five research areas are examined in the next part.

5. Discussion

The focus areas in DQ and IQ research change over time as the landscape of data quality in research and data quality in practice changes quickly [9]. To identify the evolution of DQ/ IQ research, we need to examine the dynamic changes and focus shifts in the five core areas identified above. In this part, we discuss about the changes in DQ/ IQ research over the last 30 years, based on the results of our former part.

Firstly, the examination of dynamics of five research areas shows that although some of the areas fluctuated in the past 30 years, the areas remained stable over time. DQ/ IQ research is tightly combined with technology and management [1]. The raise and falls in one research area provides research space for other areas, for example, the development of DQ/ IQ research technology promotes the application, assessment, and data process management, and change the impact of DQ/ IQ, while the application encourage the improvement of technology and data process management. The five research areas form a cycle of DQ/ IQ research: technology → application → process management → assessment → impact → new demand of technology. In addition, the uptrend of (3) DQ/ IQ system application and (4) organizational level impact of DQ/IQ shows that DQ/ IQ research has changed from monotonous areas of computing technology and assessment to more balanced areas that also focus on DQ system application and impact. The combination of technology and management, and the movement from technology toward process-related and managerial issues makes DQ/ IQ research more identifiable from IT, computing science, and other management research [1, 3].

Secondly, we can explore the results of each research area, the significantly changing of the topics, to find the dynamics characteristics. The result indicates that although many topics remain stable over time, such as user satisfaction and IS success, DQ/ IQ in network analysis technology, and measure tool and processes management, there are some interesting and important changes. We can see from Table 1 that the topic of information quality measurement model accounts for the largest amount of publications in DQ/ IQ research, which is one of the core concerns of DQ/ IQ field. However, examining the dynamic development of this topic, after going up from 1990s to 2000s, this topic fell down sharply through 2010s (from 18% to 10%). This suggests that after ample study of measurement of DQ/ IQ model in 2000s, researchers have gradually reached a consensus of how to measure data quality or information quality. The focus of (1) assessment of DQ/ IQ is shifting to user satisfaction and IS success, which take more factors into consideration other than measurement to assess DQ/ IQ.

Main computing technologies of DQ/ IQ evolved over time, too. Database and programming method gains more and more concern over three period of time and finally ranks the top concern in 2010s among these four technologies, while data integration and clean fell sharply. The result suggests that as computing technology evolves, more and more researches focus on improving database and programming method to study DQ algorithm and solve DQ problems.

In the area of organizational level impact of DQ/ IQ, topic of cost-benefit analysis declined while topic of operation and decision system increased over time, which indicates that more and more organizations accept the necessity and significance of DQ & IQ, and impact of operation and decision system for DQ solution are becoming more and more popular while the cost-benefit issue has reached an agreement.

6. Conclusions, Limitations, and Directions for Future Research

In this paper, we attempted to reveal the intellectual core and corresponding research dynamics of DQ/ IQ research. The key contribution of the paper is that we applied latent semantic analysis to identify and interpret the five core topics of DQ/ IQ and fifteen research themes to identify DQ/ IQ research. In order to better understand the trend of DQ/ IQ research, the evolution of this field in the past three decades was also discussed.

It is also important to point out several limitations of our study. First, we used keywords instead of full texts or abstracts, which may loss of information. Second, since factor labeling was done by researchers, it is hard to avoid the limitation of subjectivity biases. In addition, our sample was consisted by certain mainstream journal articles and conference proceedings that may limit the scope of our study.

Our research is a quantitative methodology based on latent semantics analysis. Our future work may consider the combination of earlier qualitative empirical studies and the method in our paper, eliminating the shortcomings of the

empirical studies and adding their advantages. By comparing the two methods, we can try to propose a comprehensive and complementary approach to address the DQ/IQ topic identification. Besides, our sample covers most of mainstream journal articles and conference proceedings such as ICIQ, CACM, MISQ, JMIS, etc. Some of them are oriented for North American, while some focus on European research. In this paper, we do not take this locality distinction into consideration. Maybe, a comparative study of North American versus European DQ/IQ research identities could be an interesting direction for future research.

For future research, researchers can explore new research topics, such as new technology and management to ensure semi-structured and unstructured data quality, cross impact of DQ/ IQ on individual and organizational, and so on. In addition, the cycle of these five core topics need more theoretical and empirical study to confirm.

In conclusion, the core view of DQ/ IQ in research as well as in practice is changing quickly overtime. As the development of discipline research and information technology, organic combination of DQ/ IQ and other disciplines will become a new main trend, and also bring new research challenges at the same time. We hope our study of DQ/ IQ discipline can enable new scholars to efficiently gain an understanding of what DQ/ IQ research is and identify potential areas of interest in a clearer and easier way.

7. References

[1] 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.

[2] Benbasat, I., and Zmud, R. W. 2003. “The Identity Crisis Within the IS Discipline: Defining and Communicating the Discipline’s Core Properties,” MIS Quarterly (27:2), June, pp. 183-194.

[3] Robey, D. 2003. “Identity, Legitimacy and the Dominant Research Paradigm: An Alternative Prescription for the IS Discipline. A Response to Benbasat and Zmud’s Call for Returning to the IT Artifact,” Journal of the AIS (4:7), December, pp. 352-359.

[4] Webster, J., and Watson, R.T. “Analyzing the past to prepare for the future: writing a literature review”, *MIS Quarterly*, 26 (2), 2002, pp. xiii-xxiii.

[5] 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.

[6] 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.

[7] 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.

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

[9] Blake, R., "Identifying the core topics and themes of data and information quality research" (2010). AMCIS 2010 Proceedings. Paper 221.

[10] Larsen, K.R., Monarchi, D.E. (2004). A Mathematical Approach to Categorization and Labeling of Qualitative Data: the Latent Categorization Method. Sociological Methodology, 34 (1) 349-392.

[11] Hovorka, D, Larsen, K and Monarchi, D (2009) Conceptual convergences: Positioning information systems among the business disciplines. In Proceedings of the 17th European Conference on Information Systems (ECIS) (NEWELL S, WHITLEY E, POULOUDI N, WAREHAM J and MATHIASSEN LEds), manuscript 0217.R1.