Discovering Data and Information Quality Research: Insights Gained through Latent Semantic Analysis

Roger Blake, University of Massachusetts Boston, USA
G.Shankaranarayanan, Babson College, USA

ABSTRACT
In the recent decade, the field of data and information quality (DQ) has grown into a research area that spans multiple disciplines. The motivation here is to help understand the core topics and themes that constitute this area and to determine how those topics and themes from DQ relate to business intelligence (BI). To do so, the authors present the results of a study which mines the abstracts of articles in DQ published over the last decade. Using Latent Semantic Analysis (LSA) six core themes of DQ research are identified, as well as twelve dominant topics comprising them. Five of these topics—decision support, database design and data mining, data querying and cleansing, data integration, and DQ for analytics—all relate to BI, emphasizing the importance of research that combines DQ with BI. The DQ topics from these results are profiled with BI, and used to suggest several opportunities for researchers.

Keywords: Business Intelligence, Data and Information Quality (DQ), DQ Dimensions, Information Quality, Information Retrieval, Latent Semantic Analysis, Research Topics

INTRODUCTION
Research in data and information quality (DQ) crosses several research disciplines and is becoming a unified body of knowledge. Starting with the pioneering works of Wang and Strong (1996) and Redman (1998), DQ has borrowed and adapted theories and techniques from many other areas including information systems, operations management, cognitive psychology, and organizational behavior, to name a few. In doing so, DQ researchers have applied a variety of research methods. Quantitative methods have been used to propose and test measurements for DQ. Models and representations have been defined for managing quality in organizations. Qualitative methods have been used to identify dimensions of DQ from the perspective of users and their context of usage. In the recent past, DQ research has embraced the design science paradigm by validating the usefulness of models, artifacts, and techniques in real-life settings. In this paper we have used the term “DQ” interchangeably with “information qual-
ity”, consistent with earlier research (Ballou & Pazer, 1985; Pipino, Lee, & Wang, 2002).

Given the extensive growth of DQ research and its likely continued growth in the future, it is important for researchers to understand the key research themes in DQ research and the popular research topics within each theme. DQ clearly has an enormous impact on the effectiveness of business intelligence (BI) and concurrent with DQ research, research of business intelligence has also grown significantly. The capabilities BI offers organizations have sharply increased and fact-based decision-making has become not just the norm for many companies, but is considered critical for success. As BI assumes ever more significance, so too does the need for a conceptual understanding of this field, as has been investigated by researchers (Foley & Guillemette, 2010). Since DQ is fundamental to BI, it is important to understand how the topics and themes of DQ research have evolved over time. It is also important to target BI-areas, related to DQ, that have not been addressed in the literature, and to identify the BI topic(s) within DQ that can garner the attention of practitioners and academics, both in BI and in DQ.

Concepts analogous to those found in DQ research can be found in BI but, these are often constructed and represented differently. These concepts may overlap but are difficult to associate. For instance, accuracy and completeness are well-known to DQ researchers as two important DQ dimensions, each defined differently and distinctly from the other. Data mining research has investigated the same phenomena, but generally considered both as forms of “noise”. Some data mining studies have defined noise in a manner that is very similar to the definition of accuracy in DQ research. Other data mining studies examining noise have used definitions similar to those for completeness in DQ research (Blake & Mangiameli, 2011).

Finding the core concepts in DQ research and how they relate to BI in order to point to opportunities for researchers is an important motivation for this study. Although they have not been explicitly connected to BI, there have been many attempts to define the core concepts of DQ research which have proposed frameworks to summarize and/or classify this area (Ge & Helfert, 2007; Lima, Maçada, & Vargas, 2006; Madnick, Wang, Yang, & Zhu, 2009; Neely & Cook, 2008). By examining the literature, each defines the classification framework from the respective researchers’ point-of-view. Although they offer invaluable insights into DQ research, we posit that there is a more interesting point-of-view that comes not from the researchers but from the research itself. What if the body of literature can inform us about the core themes and in addition associate the dominant core topics within each theme? What if we could examine relationships between those topics and themes and how they relate to business intelligence? The existing literature does not answer these questions. We believe that our methodology can answer these questions and more. Further, our methodology can be replicated to define the status of this (any) research field at any time in the future.

Understanding the core topics and themes of an area of research is an important part of “establishing the identity” of that research area. Sidorova, Evangelopoulos, Valacich, and Ramakrishnan (2008) sought to do the same for the identity of the IS discipline, which they defined with five core areas. Our ambition in pursuing this research is similar to that of Sidrova et al. – to help define the identity of DQ research by understanding its core topics and themes. The objectives of this paper are to take an important step towards defining that identity through its topics and themes, to use the results of the analysis to guide researchers in this area, and to develop a reproducible method that can be used for similar purposes in the future.

This paper makes the following contributions: (1) it identifies, from the current body of literature, the key themes in DQ research. (2) It further defines the key core topics within each theme. We strongly believe this offers a superior representation of the current state of research in this area. (3) It derives a mapping between research topics and the dimensions of DQ, which are central to much DQ research. This sheds light on how DQ research themes
Improving Business Intelligence: The Six Sigma Way

www.igi-global.com/article/improving-business-intelligence/47195?camid=4v1a