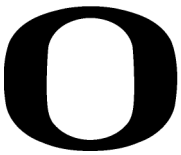
Presented to the Interdisciplinary Studies Program:
Applied Information Management and the Graduate School of the University of Oregon
in partial fulfillment of the requirement for the degree of
Master of Science



UNIVERSITY OF OREGON
APPLIED INFORMATION MANAGEMENT

Identifying and Prioritizing Information Quality Dimensions for Assurance in the Pre-Processing Stage of Data Storage for Business Intelligence

CAPSTONE REPORT

Hope Angel
Information Systems Manager
Pacific Star Corporation

University of Oregon
Applied Information Management
Program

February 2011

Continuing Education
1277 University of Oregon
Eugene, OR 97403-1277
(800) 824-2714

Approved by

_____

Dr. Linda F. Ettinger
Senior Academic Director, AIM Program

Identifying and Prioritizing Information Quality Dimensions for Assurance in the Pre-Processing Stage of Data Storage for Business Intelligence

Hope Angel

Pacific Star Corporation

Abstract

As business intelligence systems increase the amount of information stored in data warehouses, quality of content becomes more critical (Fisher, Lauria, Chengalur-Smith, & Wang, 2008). Selected literature published between 2001 and 2011 is analyzed to define key dimensions of information quality for consideration in the pre-processing stage, before data reach the warehouse, to ensure maximum quality assurance. The goal is to provide a framework to prioritize dimensions that align with business intelligence goals and objectives.

Keywords: data mining, business intelligence, information quality, information quality assurance, competitive advantages, knowledge discovery, data analytics, and data warehouse.

•

Table of Contents

List of Figures and Tables

Introduction to the Literature Review

Purpose

Business intelligence (BI) is defined as a decision support system that uses data mining techniques to extract information from data warehouses and predict future patterns for decision makers (Andersson, Fries, & Johansson, 2008; Negash, 2008). BI can be thought of as "extracting and analyzing relevant information and making it accessible for support in the decision-making process" (Andersson et al., 2008, p. 3). BI extrapolates and captures information from many other systems, such as online analytical processing tools (OLAP), data mining, decision support systems (DSS), and geographic information systems (GIS), among others (Negash, 2008). Figure 1 depicts some of the information systems that are used by BI.

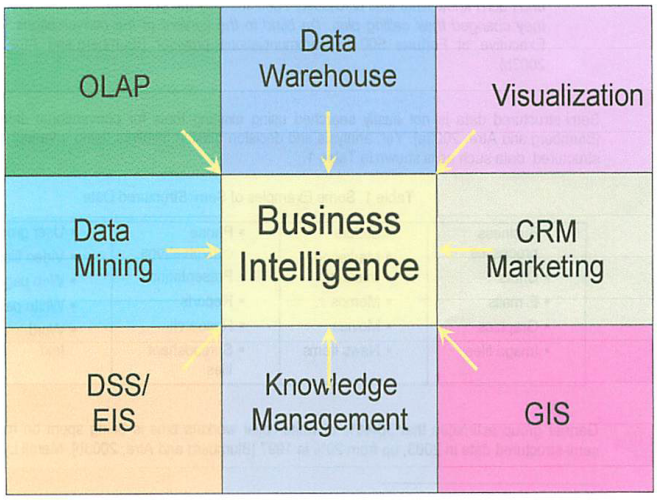


Figure 1. The relationship of business intelligence to other information systems (Negash, 2008)

According to Negash (2008) and McGilvray (2008), BI converts captured master data (i.e., key operational data) into useful information and, through analysis, into knowledge that is used to gain a competitive advantage. Lupu, Razvan, Sabau, and Muntean (2007) note that BI is the process of getting enough of the right information in a timely manner and in a usable form, and then analyzing the classification and metadata schemas to create a positive impact on the

integrity of the business and management information systems. In order to best ensure this outcome, Andersson et al. (2008) proposes devoting a significant portion of time in the pre-processing stage to identify and prioritize the dimensions that influence information quality assurance to ensure quality of content for storage in data warehouses.

The purpose of this study is to address dimensions identified in selected literature that most influence information quality assurance in the pre-processing stage of data stored in warehouses (Hakim, 2007a; Jafar, 2010). The goal is to identify and prioritize dimensions of information quality that align with BI goals and objectives for use in the pre-processing stage to improve and ensure integrity and consistency before the information reaches the warehouses. Su, Peng, and Jin (2009) describe key information as a vital business asset, and report that "information quality is a critical factor for the successful development of data warehouses and implementation of data mining" (p. 332). Information quality is not linear; identifying and prioritizing multiple dimensions such as accuracy, completeness, consistency, and timeliness, among others, is critical for effective information quality assurance strategies (Kahn, Strong, & Wang, 2002).

In order for a company to support BI, Popovic, Coelho and Jaklic (2009) state that tools such as data mining processes and information quality assurance assessments must align with and be embedded into every step of the pre-processing stage. These tools provide accurate and easily retrievable key information to improve decision making for increased performance management and competitive advantages in BI (Watson & Wixom, 2007). According to Lamont (2010) and Negash (2008), competitive advantages are a form of competitive intelligence in BI: constantly analyzing the existing market for any relevant changes, and adapting to those changes. Negash (2008) states that when combined with data mining tools, timeliness, consistency, and

quality of information improves the decision-making process and creates competitive advantages. Corporate goals and existing strategies provide a basis for analyzing BI resources (Davenport & Harris, 2007). This analysis is necessary to identify competitive advantages and disadvantages, which are the strengths and weaknesses of a corporation relative to its present and likely competitors (Lamont, 2010; Negash, 2008).

However, in order to maximize information quality assurance, it is crucial to understand the quantitative and qualitative value of the data available to decision makers (Seng & Chen, 2010). Computer systems, such as document management systems and enterprise content management, can only assist after the quality of information is assured by collecting, managing, storing, and retrieving content to better achieve the aims and goals of a comprehensive BI system (Negash, 2008; Olson, 2003). In one case example, a study conducted by Lupu et al. (2007) observes dimensions that influence information quality assurance of a real-world industry project. Analyses performed on levels of information quality of the data mining process for successful decision-making recommendations focus on the dimensions influencing and affecting project development, and solutions fulfilling dynamic BI requirements (Lupu et al., 2007). This study examines similar techniques to those presented by Lupu et al. (2007) but on a smaller scale, with specific emphasis on identifying and prioritizing dimensions of information quality that align with goals and objectives to assure information quality before the data reaches the warehouses (Cong, Fan, Geerts, Jia, & Shuai, 2007). So while the analysis performed by Lupu et al. (2007) of data mining standards for managing financial resources is indirectly related to BI, it brings clarity and focus to the research problem in this study.

The data analysis process in this study is designed to identify dimensions that most influence the quality of key information in the early stages of data preparation, so that data can

be correctly analyzed by data mining tools (Fayyad & Uthurusamy, 2002; Negash, 2008). Once information quality assurance is in place, Popovic et al. (2009) predict that pragmatic data mining techniques produce competitive advantage information and improve the decision-making process universally. According to English (2005) and McGilvray (2008), investing in information quality assurance is a means of showing benefits in returns on investment (ROI). Thus, the underlying assumption of this study is that establishing effective information quality assures capitalization of advantages and opportunities in the form of increased ROI for BI (Keeton, Mehra, & Wilkes, 2009).

Problem

Businesses recognize that change is constant, and adapting quickly to new demands is an opportunity to employ competitive business advantages and opportunities (Wixom & Watson, 2001). Data volumes have grown from megabytes to gigabytes to terabytes; some corporate databases are approaching one petabyte, a unit of information equal to one quadrillion bytes of memory (Davenport & Harris, 2007; Klein, 2002). However, while organizations have more data than ever at their disposal, sufficient capture, cleansing, and enhancement processes must be imposed to avoid data decay and de-duplication of the information (McGilvray, 2008; Panin, 2006). Two specific processes are noted in the literature: (a) knowledge discovery in databases (KDD), which is a process of extracting and capturing useful knowledge from increasing volumes of data (English, 2009; Lupu et al., 2007; Web4All, 2010), and (b) effective document management systems such as KDD, that combine data gathering, data mapping, data storage, and knowledge management with analytical tools to track, store, and efficiently extract information from data stored in warehouses (Andersson et al., 2008; Negash, 2008).

McGilvray (2008) and Klein (2002) suggest that information quality problems are not restricted to any particular entity, and that IT teams are responsible for the quality of the systems that store and move the data, but not the content. Thus, a well-designed management information system is a pivotal performance indicator and starting point to provide timely, effective, and intuitive knowledge for decision makers in BI systems (Gallo, 2010; McGilvray, 2008; Popovic et al., 2009; Su et al., 2009).

Seng and Chen (2010) suggest that data mining for business decisions requires an analytical approach to reducing data in order to manage, analyze, and apply it. Extract, transform, and load (ETL) is a common three step approach designed for data transformation and integration; it is used in data mining to extract information, index it, and load it into a target database (Keeton et al., 2009; McGilvray, 2008). In the case of structured data, analysts use Enterprise Resource Planning (ERP) to create BI information by searching, analyzing, and delivering information to the decision maker (English, 2005). The data mining process starts with analysis, and understanding the characteristics of the attributes of the data is critical so the analyst can accurately process and present the results (Jafar, 2010). Accurate information can lead to improved business performance; however, the data mining process can only generate useable patterns from data when information quality assurance is in place in the pre-processing stage (Andersson et al., 2008; English, 2009).

Significance

Improvements in technology have significantly increased the amount of data that can be stored; however, many organizations struggle with the ability to manage, analyze, and apply it successfully to BI (Davenport & Harris, 2007). Information must be managed as a resource and as an asset; it must be recent, relevant, and an accurate reflection of real-world environments to

help the business meet its goals (McGilvray, 2008). This study is significant for several reasons: (a) improving information quality assurance and data mining processes are burning issues for businesses, (b) focusing on dimensions that most influence information quality assurance at the beginning allows data mining tools to successfully search through structured information, and (c) enabling smarter decision making techniques ensures BI success, which is the goal of all businesses and organizations (English, 2005; Lee et al., 2006; Popovic et al., 2009; Wang & Wang, 2007).

According to Web4All (2010), "the most successful companies are those that can respond quickly and flexibly to market changes and opportunities; the key to this response is the effective, efficient, ease of use of data and information" (p. 1). English (2009) notes that inaccurately defining data in the early stages, mismatching definitions and the simple reality of real-world object changes can produce obscured, misidentified, or incorrectly interpreted trend findings. McGilvray (2008) finds that poor data quality impacts project timelines, hampers data mining processes, and reduces confidence in data analysis results. As a result, BI processes fail when vital information is captured inaccurately (English, 2009; Forcada, Casals, Fuertes, Gangolells, & Roca, 2010).

Andersson et al. (2008) find that timely, accurate knowledge contributes to improved business performance. Indeed, for organizations that depend on data for decision-making processes, information quality assurance is one of the key determinants of the quality of their decisions and actions (Stvilia, Gasser, Twidale, & Smith, 2007). According to Halonen and Thomander (2008), the ability to assure quality information prior to data reaching the data warehouse is significantly important to enhancing the decision-making process and identifying competitive advantages for BI.

Past methods for delivering BI solutions focus on quick turnarounds, rather than embedding the implementation of guidelines into each and every step along the way (Gallo, 2010; IBM, 2010). Lupu et al. (2007) and Gallo (2010) find that simply extracting unstructured or undefined data stored in a warehouse does not provide a viable response to changing business needs. Moreover, failing to address issues of unstructured data reduces information quality both in data collection and definitions (Forcada et al., 2010). English (2005) describes information quality assurance as critical to BI success; in fact, he states that "problems in information definition, data content, data preparation, and misinformation can cause BI processes to fail" (p.1). Lack of information quality assurance compromises data integrity, and is prevalent in companies experiencing inefficient, non-integrated reports and analyses (Fisher, Lauria, Chengalur-Smith, & Wang, 2008; Lupu et al., 2007; Stvilia et al., 2007).

In order for BI decision-making processes to be successful, information quality assurance must be in place early on (Gallo, 2010; Popovic et al., 2009; Su et al., 2009). Lupu et al. (2007) find that empowering dynamic analysis and making the right decisions towards a competitive advantage can only be obtained by focusing on dimensions that most influence information quality assurance in the pre-processing stage of data storage. Thus the key to maximizing information quality becomes getting the right set of structured information to the right people at the right time, for their use in decision making to achieve company goals (IBM, 2009; McGilvray, 2008).

Audience/Outcome of Study

Everyone makes decisions; enabling smarter decision-making techniques from every level of a business is what makes businesses intelligent (IBM, 2010). BI is widely used and has become a strategic initiative recognized by CIOs and business leaders as instrumental in driving

business effectiveness and innovation (Rodriguez, Daniel, Casati, & Cappiello, 2010; Watson & Wixom, 2007). Cong et al. (2007) suggest that demonstrating the ability to identify dimensions that influence information quality assurance combines knowledge with information, thus producing successful BI. According to IBM (2009), businesses are most likely to reach desired outcomes when they have access to "complete, consistent, and trustworthy information" (p.1). Thus it is critical that analysts have the right information before decisions are sorted out and weighed in order to make full use of BI capabilities (Hakim, 2007b; Keeton et al., 2009). Negash (2008) finds that the demand for BI applications continues to grow even if the demand for IT products does not. It is implied, then, that BI provides actionable information "delivered at the right time, at the right location, and in the right form to assist decision makers" (Negash, 2008, p. 178). However, all businesses and organizations, at one point or another, confront information quality problems (Lee, Pipino, Funk, & Wang, 2006).

Knowing the business, its market, its customers, and its competition is a precursor to understanding how information quality is defined for any organization (IBM, 2010). The audience for this study is executives and professionals within organizations who require data analysis as key performance indicators (KPI) to generate analytical solutions for increasing revenue and moving the company to the forefront of the competition (McGilvray, 2008). This group includes knowledge workers, technologists, and professionals (Gallo, 2010; Kriegel, Borgwardt, Kroger, Pryakhin, Schubert, & Zimek, 2007; McGilvray, 2008). They must have access to a reliable system for creating, processing, and enhancing their own knowledge (McGilvray, 2008). In general, this study is designed to be beneficial to key decision makers faced with dynamic corporate goals and demands (Lee et al., 2006). Due to the rapid global expansion of information-based transactions and interactions being conducted via the Internet,

there is an increased demand for a workforce that is capable of performing these activities (Haag, Cummings, McCubbrey, Pinsonneault, & Donovan, 2006). In fact, Haag et al. (2006) note that knowledge workers are now estimated to outnumber all other workers in North America by at least a four to one margin. Table 1 illustrates the audience profile, characteristics, and professions of those who are most likely to benefit from information quality for decision making in BI.

Table 1
Audience Profile indicating Categories, Profession characteristics, and Titles

Broad Category of Audience	Characteristics	Job Title
Knowledge Workers	Oriented towards research and analysis of data; thus quality is essential to outcome.	Chief Information Officer (CIO) Chief Knowledge Officer (CKO) Knowledge Manager (KM) Content Manager Knowledge Steward Program Managers Project Managers Project Team Members Executives, Sales, Marketing Finance, Legal, Human Resources
Knowledge Technologists	Focus is on developing an increasing value of intellectual capital, gaining insight into customer preferences, and a variety of other important gains in knowledge that aid the business.	Computer Analysts Software Designers Software Analysts IT Professionals Administrative Assistants
Knowledge Professionals	Professionals who are valued for their ability to act and communicate with knowledge within a specific subject area.	Teachers, Librarians Lawyers, Architects Practitioners, Physicians, Nurses Engineers, Scientists

The intended outcome of this study is a framework that identifies and prioritizes key dimensions of information quality that align with BI goals and objectives to ensure the effectiveness of information quality in the pre-processing stage of data storage. Furthermore,

dimensions that the selected literature indicates are the most influential for the assurance of information quality in the pre-processing stage of data storage within the context of BI are addressed. Decision makers, such as knowledge workers, executives, and professionals faced with dynamic corporate goals and demands, must consider the dimensions of information quality from the perspective of the users of data (Lee et al., 2006). Thus dimensions are addressed in relation to the goal of successful decision-making to gain competitive advantages for BI, which is identified in the selected literature (Lamont, 2010; Negash, 2008). In this context, competitive advantage refers to the strengths of a corporation relative to its present and likely competitors by constantly analyzing the existing market for any relevant changes, and adapting to the changes quickly (Lamont, 2010; Negash, 2008). The framework for identifying and prioritizing key dimensions is organized around two themes: (a) discussing the role of information quality assurance, and (b) examining information quality dimensions and the effect that they have on information quality assurance in the pre-processing stage of data storage, within the context of BI.

Research Delimitations

Time frame. Thiesse, Floerkemeir, Harrison, Michahelles, and Roduner (2009) suggest that the large number of publications in recent years could all be potentially viewed as contributing to the field of BI; however, due to recent advances in BI, it is best to focus on references published within the last five to ten years. Thus, the references provided in this study are limited to publishing dates between 2001 through 2011. The focus on this period is seen as covering the rapid changes in information quality assurance and data mining that are shaping BI as it has evolved today (Davenport & Harris, 2007). Older research results are rendered obsolete by the myriad changes in the aspects aligning BI maxims with IT, thus this time frame excludes

older research and practices that may not reflect the current advancements in BI (Seng & Chen, 2010). In order to reduce the likelihood of obsolete information becoming a part of the focus, resources published prior to January 2001 are not used in this literature review.

Selection criteria. Literature is selected from peer-reviewed scholarly resources, and others that have been deemed to be authored by an authority on the topic, including business publications, whitepapers, online academic journals, and books, using keyword searches, controlled terms, and scope notes (Bell & Smith, 2007; Ormondroyd, Engle, & Cosgrave, 2009). References selected for this literature review are directly relevant to information quality assurance and the dimensions that most influence it. Additional references are used to establish the framework for this review, such as those that associate data mining with the decision-making process of BI. They are then carefully evaluated to gauge both relevancy and credibility (Bell & Smith, 2007; Creswell, 2009). References meeting the requirements are compared to identify commonalities among dimensions that have been proven to have the most influence on information quality assurance.

According to Creswell (2009), scholarly material provides a practical and theoretical context for the study and is useful for developing a framework for comparing results of this study with other studies. Thus, all literature is reviewed for quality of methods, results, and conclusions and is included in or excluded from this study based on usefulness, breadth of scope, quality, publishing date, accessibility, and language (Bell & Smith, 2007; Creswell, 2009; Leedy & Ormrod, 2010). This study only includes literature that is available for viewing online or reproducible in hard copy.

Outcome/Audience. This literature review is designed to produce a framework that addresses identifying and prioritizing key dimensions of information quality in the pre-

processing stage of data storage that align with unique BI goals and objectives (McGilvray, 2008). The intended outcome is based on identification of dimensions that most influence information quality assurance for consideration by those who are responsible for data storage and data mining processes, specifically in an IT business environment, such as knowledge workers, knowledge technologists, and knowledge professionals, including project managers, project team members, executives, and IT professionals and others (Haag et al., 2006; Zhao, Chen, & Yao, 2006). The audience should be familiar with the requirements for data mining within the context of a business environment in order to accurately determine how the identified dimensions should be applied for information quality assurance (Lefebvre, 2007). This study is not designed to benefit other audiences such as educational and non-profit agencies who may have an interest in implementing a BI system but may not be directly involved in the data mining or analysis processes (Lefebvre, 2007).

Topic focus. In order to generalize this study, literature is selected that includes the larger context of data mining for BI; however, the study is limited to one aspect of BI, namely information quality assurance in the pre-processing stage of data storage (Obenzinger, 2005). The scope of this study is further limited to the identification of dimensions of information quality that have the most influence on information quality assurance, specifically in terms of completeness, consistency, and trustworthiness, in the pre-processing stage of data storage to ensure that information can be successfully retrieved and correctly analyzed by data mining tools (Wixom & Watson, 2001).

The search for dimensions of information quality that have the most influence on information quality assurance is limited to those that are relevant to the pre-processing stage of data storage within the context of BI. However, in order to address this aspect it is necessary to

address its association with data storage, mining, and analysis processes for BI. There are a number of other related issues in BI systems that are excluded from this literature review. Specifically, the excluded subject areas are those that address data mining patterns and particular decision-making strategies for gaining competitive advantages in BI. Additionally, this study excludes the requirements and steps taken for data mining and analysis processes. Also excluded are the detailed decision-making processes that are necessary for successful BI systems. The decision to exclude these areas is designed specifically to place the focus of the inquiry on the dimensions that influence information quality assurance and lead to the success of BI, not on the outcomes. Additionally, the inclusion of these areas would have been beyond the reach of this study, given the limited time for conducting the study.

Inquiry context. The problem, sub-topic, and audience selection are framed based upon real-world challenges to retrieving unstructured information that is stored in data warehouses (Piatetsky-Shapiro et al., 2009). For example, many companies compete on the basis of their analytical capabilities by using BI to make better decisions and to extract maximum value from their data warehouses; the value of information, however, is not in the information itself but in how it affects the business (Davenport & Harris, 2007). Selected literature explores the role of information quality within the context of BI and the role of the dimensions of information quality for the purpose of assuring quality of content in the pre-processing stage of data storage (Knight & Burn, 2005).

Data Analysis Plan Preview

Resources that satisfy the evaluation criteria are analyzed using a qualitative approach known as content analysis (Busch, De Maret, Flynn, Kellum, Le, Meyers, Saunders, & White, 2005; Obenzinger, 2005; Ormondroyd et al., 2009). Content analysis is a widely used research

tool used to determine the presence of certain words or concepts within selected resources (Busch et al., 2005; Hsieh & Shannon, 2005). The approach begins with identifying research questions, selecting resources, and classifying and coding selected text into manageable categories to enable the researcher to "focus on, and code for, specific words or patterns that are indicative of the research question" (Busch et al., 2005, para. 1). The coding process is divided into eight steps and is detailed in the Research Parameters section of this paper (Busch et al., 2005):

1. Decide the level of analysis.
2. Decide how many concepts will be coded.
3. Decide whether to code for existence or frequency of a concept.
4. Decide on how concepts will be distinguished from one another.
5. Develop rules for coding texts.
6. Decide what to do with irrelevant information.
7. Code the texts.
8. Analyze and report results.

Focus during coding is on identification of dimensions that influence the quality of information in the pre-processing stage of data storage.

Writing Plan Preview

The writing plan for the presentation of the results compiled during the data analysis process is designed to provide a framework of the topic (Obenzinger, 2005). The objective is to address common themes between resources for assuring information quality in terms of relevance, accuracy, timeliness, and completeness that are proven effective in real-world environments (Busch et al., 2005). Presentation of the information aligns with the thematic

pattern of organization (University of North Carolina, n.d.). This approach, which organizes literature around a topic, emphasizes the development of the most influential dimensions of information quality assurance rather than the chronological development of the materials (University of North Carolina, n.d.).

The goal of the writing plan is to present the data derived from the coding process in a way that addresses identifying and prioritizing key dimensions of information quality. Theme one presents a description and discussion of the role of information quality assurance in the pre-processing stage of data storage. Anticipated sub-themes include the role of information quality, information quality for data storage, the impact of the quality of content, and the effects of information quality assurance for BI. Theme two presents an identification of key information quality dimensions for assuring information quality in the pre-processing stage of data storage. Anticipated sub-themes include a discussion of the role of each identified dimension, including which are most common and which are key.

Definitions

Special terms that are unique to the field of information quality, data mining, and BI are used in this study. According to Lamont (2010), decision makers in BI systems often use terms that are obscurely defined and have various meanings to different team members; thus establishing clear definitions reduces miscommunication and costly mistakes. Definitions are provided in this section to ensure that readers clearly understand the contextual meaning of the terminology, as used throughout this study. Key terms are defined in-text at the point at which they are introduced; others are withheld to prevent interruptions in the flow of the document and are defined in this section. The following list of terms provides a helpful collection of definitions interpreted in the context of information and data quality.

Accessibility – Accessibility is a dimension of information quality; it is the extent to which information is quickly retrievable (Kahn et al., 2002).

Accuracy – Accuracy, for the purposes of this study, is defined as the degree of the correctness of the content of the data (Davenport & Harris, 2007). It is a dimension of information quality also referred to as validity (McGilvray, 2008).

Amount of Data – Amount of data, or quantity, is a dimension of information quality that refers to the volume of information appropriate for the task at hand (Kahn et al., 2002).

Assessment – Assessment is the comparison of the actual environment and data to requirements and expectations (McGilvray, 2008).

Attribute – An attribute is additional information included with a dimension that is not used in defining the levels of the dimension (Arkady, 2007).

Believability – Believability is a dimension of information quality; it is the extent to which information is regarded as credible (Kahn et al., 2002).

Business Intelligence (BI) – Business Intelligence (BI) is a decision support system that utilizes data mining techniques to extract information from data warehouses (Andersson et al., 2008).

Business Intelligence Tools– Three types of tools are referred to as BI tools: analytical software (dimensional variations in data), query tools (ask questions about patterns in data), and data mining tools (search for significant patterns in data) (Watson & Wixom, 2007).

Business Systems – Business systems combine data gathering, data storage, and knowledge management with analytical tools to present complex internal and competitive information to planners and decision makers (Negash, 2008).

Classification Scheme – A classification scheme is the descriptive information for an arrangement or division of objects into groups based on characteristics that the objects have in common (Kriegel et al., 2007).

Compatibility – Compatibility, an information quality dimension, refers to the extent to which data are combined with other information (English, 2005).

Competitive Advantages/Disadvantages - Competitive advantages and disadvantages are the strengths and weaknesses of a corporation relative to its present and likely competitors by constantly analyzing the existing market for any relevant changes, and adapting to the changes quickly (Lamont, 2010; Negash, 2008).

Competitive Intelligence – Competitive intelligence is the act of gaining perspective on developments and events aimed at yielding a competitive advantage (Lamont, 2010).

Completeness – Completeness, a dimension of information quality, is the extent to which the expected attributes of data are provided, as it meets the expectations of the user (Caro et al., 2008). Data is of good quality when it is complete: when the user has coverage for all

needed data, when all related pieces are intact, and when the content is updated to correct any mistakes (Negash, 2008; Olson, 2009).

Concise Representation – Concise representation is a dimension of information quality that refers to the extent to which information is compactly represented.

Conformity – Conformity, an information quality dimension, is the extent to which data values conform to specified formats (Jafar, 2010).

Consistency – Consistency of data means that data across the business should be in sync with each other and is a dimension of information quality (McGilvray, 2008).

Controlled Vocabulary – A controlled vocabulary provides a way to organize knowledge for subsequent retrieval (Stvilia et al., 2007).

Customer Relationship Management (CRM) – Customer relationship management (CRM) is a term for the methodologies, software, and Internet capabilities that help a business manage customer relationships in an organized fashion (Negash, 2008).

Data – Data consist of unconnected facts, numbers, names, codes, symbols, dates, words, and other items of that nature that are out of context, and that only acquire meaning through association; it is what a computer records, stores, and processes (Negash, 2008).

Data Analysis – Data analysis is an approach in which data is organized so that useful information can be extracted from it (Jafar, 2010). It embeds predictive analytics into frontline applications to improve decision making (IBM, 2010).

Database – A database consists of an organized collection of data for one or more uses (Berkley, Bowers, Jones, Madin, & Schildhauer, 2009).

Data Capture – Data capture is the extraction of or access to data (Forcada el al., 2010).

Data Categories – Data categories are groupings of data with common characteristics or features (McGilvray, 2008).

Data Cleansing – Data cleansing is updating data that are imported to the warehouse, such as cleansing or removing errors and inconsistencies (McGilvray, 2008).

Data Decay – Data decay refers to a measure of the rate of negative change to the data (McGilvray, 2008).

Data Integrity – Data integrity is the process of ensuring consistency throughout all information systems, and providing end-to-end management of all metadata (Gallo, 2010).

Data Mapping – Data mapping is the process of determining where the data in a source data store is moved to another target data store (Seng & Chen, 2010).

Data Mining – Data mining refers to the technology that allows the user to efficiently retrieve information from the data warehouse (Sen & Sinha, 2007). Data mining technology is used to discover hidden relationships, patterns, and interdependencies, and generate rules to predict the correlations in data warehouse (Su et al., 2009).

Data Quality – See information quality.

Data Warehouse – A data warehouse is defined as a repository of historical data used to support decision making that allows centralized analysis, security, and control over data (Sen & Sinha, 2007; Web4All, 2010).

Decision Support – Decision Support is information that is generated to support decision makers in the decision-making process (Andersson et al., 2008).

Decision Support Systems (DSS) – Decision Support Systems (DSS) are systems used to directly support specific decision-making processes (Parameter, 2010).

De-duplication – De-duplication is a feature of data cleansing tools or processes that identifies multiple records representing the same real-world object (McGilvray, 2008).

Dimension – A dimension is one of the perspectives that can be used to analyze the data (Kanal, 2009).

Document Management System – A document management system (DMS) is a computer system used to track and store electronic documents or images of paper documents (Parameter, 2010).

Duplication – Duplication is an information quality dimension that refers to maintaining a single representation of similar data within the data set (Jafar, 2010).

Ease of Use – Ease of use refers to the degree to which data can be accessed and is a dimension of information quality (McGilvray, 2008).

Enhancement – Enhancement refers to a feature of data cleansing tools that updates or corrects data or adds new information to existing data (Berkley et al., 2009; McGilvray, 2008).

Entity – An entity is a person, place, or thing that is of interest to the business (Negash, 2008).

Enterprise Content Management – Enterprise content management refers to the use of appropriate technology and software to collect, manage, store, and retrieve content of any kind, including documents and unstructured information within an organization in order to better achieve the aims and goals of the business (Negash, 2008; Olson, 2003).

Environment – The environment refers to the conditions within a company that affect the way employees work and act (Panin, 2006).

Enterprise Resource Planning (ERP) – In the case of structured data, analysts use Enterprise Resource Planning (ERP) to create BI information by searching, analyzing, and delivering information to the decision maker (English, 2005).

Extract, Transform and Load (ETL) - Extract, Transform and Load (ETL) is a common three step process designed for data transformation and integration; it is used in data mining for patterns to extract data from a source system, transform and aggregate them to meet target system requirements, and load them into a target database (Keeton, et al., 2009; McGilvray, 2008).

Free of Error – Free of error is a dimension of information quality that refers to the extent to which information is correct and reliable (Kahn et al., 2002).

Gigabyte - A gigabyte is a unit of information equal to one billion bytes of memory (Popovic et al., 2009).

Geographic Information Systems (GIS) – A geographic information system (GIS) integrates hardware, software, and data for capturing, managing, analyzing, and displaying all forms of geographically referenced information (Negash, 2008).

Indexing – Indexing refers to a list of records arranged in order of some attribute (Wang & Wang, 2007).

Information – Information is the meaning given to data or the interpretation of data based on its context (English, 2005).

Information Quality – Information quality is the degree to which information and data can be a trusted source for any and/or all required users (McGilvray, 2008). While there is no single definition for information quality, researchers agree that it quantifies whether the correct information is being used to make a decision or take an action, and whether that information is good enough for the purpose of making a decision (Keeton et al., 2009; Popovic et al., 2009).

Information Quality Dimensions – Information quality dimensions are the minimum desired qualities that data should have to be considered effective for data mining techniques (English, 2009).

Information Quality Assurance – Information quality assurance is a methodology of assuring that the data retrieved is relative for BI (Stvilia, et al., 2007).

Integrity – Integrity, an information quality dimension, is when data generated by BI information systems are protected from deliberate bias or manipulation for political or personal reasons (Kahn et al., 2002).

Interpretability – Interpretability is a dimension of information quality that refers to the extent to which information is in appropriate languages, symbols, and units, and the definitions are clear (Kahn et al., 2002).

Key Performance Indicators (KPI) – KPIs are a set of quantifiable, long-term goals, which are measurable and key to the success of the company, that determine if a company is reaching its performance and operational goals (Parameter, 2010).

Keyword – Keyword is a substantive word in the title of a document or a record in a database that can be used to classify or index content (Arkady, 2007).

Knowledge – Knowledge is data that has been organized, synthesized, and made useful; it is what a business uses to make decisions (McGilvray, 2008).

Knowledge Discovery in Databases (KDD) – Knowledge Discovery in Databases (KDD) is the process of extracting useful knowledge from volumes of data (English, 2005; Lupu et al., 2007; Web4All, 2010).

Knowledge Management – Knowledge management is used to address technologies employed for the management and analysis of unstructured information (Halonen & Thomander, 2008).

Knowledge Worker – A knowledge worker is one who uses data or information to perform his or her work or to complete job responsibilities (Halonen & Thomander, 2008; McGilvray, 2008).

Linking – Linking is a feature of data cleansing tools that matches, or links, associated records through a user-defined or common algorithm (McGilvray, 2008).

Management Information Systems – A management information system is a system that provides information needed to manage organization effectively (Gallo, 2010).

Master Data – A data category that describes the people, places, and things that are involved in an organization's business (McGilvray, 2008; Rodriguez et al., 2010).

Matching – Matching is a feature of data cleansing tools or processes that matches, or links, associated records, through user-defined or common algorithm (McGilvray, 2008).

Measure – A measure refers to an indicator that is an indirect predictor of performance (Kanal, 2009; McGilvray, 2008).

Media – Media refers to the various means of communication, such as user guides, Web surveys, hardcopy forms, and database entry interfaces (Lupu et al., 2007; McGilvray, 2008).

Megabyte – A megabyte is a unit of information equal to one million bytes of memory (Popovic et al., 2009).

Metadata – A data category that describes data in the warehouse that labels, describes, or characterizes other data for ease of use in retrieving, interpreting, or using information; it literally means "data about data" (McGilvray, 2008, p. 294).

Objectivity - Objectivity, a dimension of information quality, is the extent to which information is unbiased, unprejudiced, and impartial (Kahn et al., 2002).

On-line Analytical Processing Tools (OLAP) – On-line analytical processing tools (OLAP) are computer applications designed to search, analyze, and deliver data to assist in the decision-making process of BI (English, 2005).

Parsing – Parsing refers to the separation of character strings or free-form text fields into component parts, meaningful patterns, or attributes, which are then moved into clearly labeled and distinct fields (McGilvray, 2008; Popovic et al., 2009).

Petabyte – A petabyte is a unit of information equal to one quadrillion bytes of memory (Popovic et al., 2009).

Precision – Precision, an information quality dimension, means that data have sufficient detail (McGilvray, 2008).

Predictive Analytics – Predictive analytics is a tool used in data mining to predict future probabilities and trends (Kriegel et al., 2007; Davenport & Harris, 2007; Forcada et al., 2010).

Process – Process refers to any functions, activities, actions, tasks, or procedures that touch the data or information (English, 2005; Berkley et al., 2009; McGilvray, 2008).

Profiling – Profiling is the use of analytical techniques to discover the structure, content, and quality of data (Olson, 2003).

Reference Data – A data category that are sets of values or classification schemas referred to by systems, applications, data stores, processes, and reports (McGilvray, 2008).

Relevancy – Relevancy is a standard for determining if what is being considered in the project is associated with and meaningful to the business issue to be resolved (McGilvray, 2008).

Reliability – Reliability, a dimension of information quality, is the extent to which data are measured and collected consistently (McGilvray, 2008).

Reputation – Reputation, an information quality dimension, is the extent to which information is highly regarded in terms of its source or content (Kahn et al., 2002).

Return on Investment (ROI) – Return on investment (ROI) is a means of showing benefit from investing in data quality (English, 2005; McGilvray, 2008).

Root Cause Analysis – Root cause analysis is the study of all possible causes of a problem, issue, or condition to determine its actual cause (Cong et al., 2007; McGilvray, 2008).

Sample – Sample refers to a subset of a population or a group under study that is representative of the entire population (Arkady, 2007).

Schema – The schema refers to the local organization of data in a database (McGilvray, 2008).

Search Engines – Search engines are software programs capable of successfully retrieving information from computer networks or databases in order to match the needs of searchers (English, 2005; Negash, 2008; Zhao et al., 2006).

Security – Security is an information quality dimension that refers to the extent to which access to information is restricted appropriately to main its security (Kahn et al., 2002).

Serviceability – Serviceability, an information quality dimension, is the extent to which data are consistent and follow a predictable revisions plan (Olson, 2003).

Stakeholder – A stakeholder is any individual or group that has a direct interest, or some level of involvement, in the success of an organization and would be affected by the outcome of any decisions (Popovic, 2009).

Standardization – Standardization refers to converting data into standard formats to facilitate parsing and, thus, matching, linking, and de-duplication (McGilvray, 2008).

Strategic Early Warning – Strategic early warning is the process of monitoring the business environment for weak signals and early trends that may reveal potential changes before they become obvious to others (Gallo, 2010; Lupu et al., 2007; Rodriguez et al., 2010).

Strategic Group Analysis – Strategic group analysis identifies groups or clusters of businesses that adopt similar strategies and that tend to be affected by, and respond to, competitive actions and external events in similar ways (Gallo, 2010; McGilvray, 2008).

Strategic Intelligence – Strategic intelligence is knowledge about an organization's business environment that has implications for its long-term viability and success, usually extending several years into the future (Gallo, 2010; McGilvray, 2008).

Strategic Research – Strategic research is mission-oriented and involves the application of established scientific knowledge and methods to broad social or economic objectives, often extending over a considerable period (Gallo, 2010; McGilvray, 2008).

Synthesis – Synthesis is the process of combining data, information, and existing knowledge in order to produce a connected whole, such as a hypothesis, theory, or system (Arkady, 2008).

Target Audience – The target audience is a group of people for whom a specific study is directed (Lefebvre, 2007).

Terabyte – A terabyte is a unit of information equal to one trillion bytes (Popovic et al., 2009).

Timeliness – Timeliness, an information quality dimension, refers to the degree to which data are sufficiently up to date for the task at hand (Kahn et al., 2002; McGilvray, 2008).

Transactional Data – A data category that describes an internal or external event or transaction that takes place as an organization conducts its business (McGilvray, 2008).

Transformation – Transformation is any change to the data, such as during parsing and standardization (Gallo, 2010; Parameter, 2010).

Trust – Trust refers to confidence in data quality (McGilvray, 2008; Rodriguez et al., 2010).

Unstructured Data – Unstructured data is information that has no defined or standard structure such as would allow for its convenient storage and retrieval (Popovic et al., 2009).

Usage – Usage is a technique that inventories the current and/or future uses of the data (McGilvray, 2008).

Validity – Validity, a dimension of information quality also referred to as accuracy, refers to the determination that values in the field are or are not within a set of allowed or valid values (McGilvray, 2008).

Research Parameters

Literature reviews are beneficial because they provide a meaningful context of a study within the framework of existing research to broaden the understanding of the problem (Obenzinger, 2005). This section presents the framework and methods in which the literature review is designed and conducted (Geist, 2008). A detailed search strategy is established and outlined as a guide for continued methods to search, locate, and retrieve literature (Obenzinger, 2005). A method is defined by which resources are deemed credible and relevant to the information search (Luckey, 2009). Evaluation criteria are conveyed in terms of credibility and relevance to the topic (Obenzinger, 2005).

The initial set of research questions and sub-questions, the search strategy report, the documentation approach, and the full descriptions of the data analysis and writing plans are presented (Creswell, 2009). The documentation approach outlines and details the processes used to record, classify, code, and capture all resources found.

Research Design

Obenzinger (2005) describes a literature review as a method of "providing meaningful context for a project within the universe of already existing research" (p.1). A methodological review of past literature is a crucial endeavor for any academic research work (Levy & Ellis, 2006). Indeed, it is through the literature review that previous perspectives are synthesized, and new ones are gained (Obenzinger, 2005).

This inquiry is structured as a review of literature that evaluates and summarizes the most relevant information on and provides meaningful context of the topic in order to set the basis for an indication toward further research (Obenzinger, 2005). According to Obenzinger (2005),

providing direction for further research requires the analysis of a large body of selected literature to present a "picture of current knowledge, identifying gaps or holes in the field" (p. 5). The direction for further research is expressed by focusing on factors that most influence information quality assurance in the pre-processing stage of data storage within the context of BI.

Research Questions and Sub-questions

The design of this study is framed by a series of research questions that guide the development of both content and research process (Creswell, 2009). The questions are each formulated to focus on identification of the dimensions that most influence information quality assurance. The overarching research question is: What dimensions most influence information quality assurance in the pre-processing stage of data storage, in an effort to support data mining, where the goal is to produce competitive advantage information for BI (Andersson et al., 2008; Creswell, 2009; IBM, 2010; Rodriguez et al., 2010)? The guiding questions and sub-questions are listed below.

1. What is the role of information quality assurance in the pre-processing stage of data storage within the context of BI?
    a. What is the role of information quality?
    b. What are the benefits of information quality for data storage?
    c. What is the impact of quality of content?
2. How are key dimensions identified and prioritized to assure information quality in the pre-processing stage of data storage?
    a. What are the key information quality dimensions for BI?
    b. How are key dimensions identified and prioritized?

c. How do key dimensions contribute to assuring information quality in the pre-processing stage for data storage?

Search Strategy Report

Although the collection of literature is focused on materials pertaining to information quality assurance in the pre-processing stage of data storage, other areas relevant to the topic are addressed in order to provide a basis for understanding within the larger context of BI (Creswell, 2009; Fink, 2010; Leedy & Ormrod, 2010). The process of literature collection focuses on information quality assurance and the role it plays in data mining for BI decision-making processes.

Search terms. Exploratory keyword searches are derived from a variety of types of literature, including books published on information quality assurance, data storage, data mining, and BI. The whitepaper "Business Intelligence: From data collection to data mining and analysis" published by Web4All (2010) refers to keywords as IT industry standard terminology. For example, the term "data mining" dates back to the 1950s, with even earlier methods of identifying patterns in data and regression analysis in the 1700s (Web4All, 2010).

The predominant search method is to identify co-referential terms and links in books, peer-reviewed journal articles, conference proceedings, and reports that align with the role of information quality for successful data mining in BI (Creswell, 2009; Fink, 2010; Tang, Jin, & Zhang, 2008). Controlled terms and scope notes help focus the keyword searches by referring to other terms, concepts, or links connected to the much broader database search (Tang et al., 2008).

Key search terms initially used are:

- data mining;

- business intelligence;
- information quality;
- information quality assurance;
- knowledge discovery;
- competitive advantage;
- data analytics; and,
- data warehouse.

Literature resources. This study is designed as a literature review, with the goal of expanding knowledge and adding clarification of the topic (Creswell, 2009; Fink, 2010; Leedy & Ormrod, 2010). The initial search for literature includes the following types of references: peer reviewed academic research, business publications, whitepapers, online academic journals, and books.

Search engines. Initial specific sites searched include the UO Libraries, Web of Science, Google Scholar, Sage Journals Online, Academic Search Premier, ERIC, and Google. As the topic focus began to reveal itself, database searches expanded to include CiteSeer Index, ACM Digital Library, IEEE Computer Science Digital Library, JSTOR, and Project Muse.

Search strategies. The first search strategy is to utilize the University of Oregon's (UO) libraries Web site, using a combination of controlled term and keyword searching. Creswell (2009) suggests that a good search method is to follow leads to the specific article, or to the database with a large number of relevant results, and refine and search again until articles relevant to the topic focus are located. The UO libraries Web site provides search access to several relevant academic search indexes, such as Academic Search Premier, JSTOR, Project Muse, and Web of Science**.**

According to Berkley et al. (2009), another effective strategy is to perform both keyword and controlled term searching to cover a wider range of possible results and to avoid false hits. This can be done on most search sites via the use of thesauri (Creswell, 2009). Kanal (2009) suggests employing an iterative process as another effective strategy. Thus some searches are conducted using relevant fixed fields; text queries are added in subsequent searches to narrow the results (Berkley et al., 2009).

Search Results

Appendix A, *Detailed Record of Searches,* illustrates which search engines and databases are utilized, search terms that are applied, and results that are obtained. Categories of information include:

- Search Engine/Database, which lists the resource used for the search;
- Search Terms, denoting keywords used that are related to topic, subtopic, and research questions;
- Number of Search Results, indicating the number of hits resulting from the search;
- Number of Eligible Titles Found, referring to the number of relevant pieces of literature that are eligible for inclusion in this study; and
- Comments, stating the rationale for continuing with or abandoning specific search engines, databases, and search terms.

Results from searches of the ACM Digital Library, Academic Search Premier Index/EBSCO HOST (UO Libraries), and CiteSeer^{X} Search Index produced very good to excellent results, yielding some of the most relevant, quality literature pertaining to the topic. Searches of Google Scholar Advanced, IEEE Computer Science Digital Library, Project Muse

(UO Libraries), and Web of Science (UO Libraries) produced adequate to good results, although some required membership to view the full article. Several other search engines, including ERIC, SAGE Journals Online, and Google, were abandoned because search results were consistently non-productive to the topic; it was deemed not worth continuing the effort to use these databases or search engines. Several other search engines and databases were abandoned due to duplication of results or lack of authority for the source.

A summary of search results is illustrated in Table 2.

Table 2

Summary of Search Engine and Database Results

Search Engine/Database	Eligible Titles Found
ACM Digital Library	69
Academic Search Premier Index/EBSCO HOST	74
CiteSeer^{x} Search Index	79
ERIC	1
Google Scholar Advanced	34
IEEE Computer Science Digital Library	31
Project Muse	43
Sage Journals Online	17
Web of Science	45

Evaluation Criteria

The literature selected for this study comes from a variety of resources in order to focus on dimensions that most influence information quality (Creswell, 2009; Obenzinger, 2005). All resources are collected using keyword searches from online search engines and databases. The majority of resources are drawn from CiteSeer^{x} Search Index, which is a scientific literature

digital library and search engine indexing over 750,000 documents that focuses primarily on the literature in computer and information science (CiteSeer^{x}, n.d.).

Results are restricted to articles, papers, conference proceedings and books published after January 1, 2001 in an effort to reference the most current and updated information (Bell & Smith, 2007). Excluding information older than ten years is critical to this literature review in order to focus on the most influential dimensions affecting information quality (English, 2009).

Keyword search and date parameter filters are set within each search engine or database. The resulting list of matches is reviewed to determine validity and value to this topic. Abstracts are reviewed to further determine the significance of the match. Matches that meet the criteria are considered relevant and are added to BibMe, an online automatic citation creator that supports APA formatting. If the search does not produce any relevant hits, the keywords are revised and the search is repeated. If results are still not relevant to the topic after multiple search attempts with various controlled terms, the decision is made to abandon the search engine altogether.

After relevant resources are identified, credibility is examined to determine the authority of the document (Bell & Smith, 2007). Three steps are then followed to evaluate the credibility of the author, the validity of the research, and the relevance of articles (Bell & Smith, 2007; University of Colorado at Boulder, n.d.).

The first step is to evaluate authority and trustworthiness by determining credentials, education, and experience of the author (University of Colorado at Boulder, n.d.). Next, the second step, according to the University of Colorado at Boulder (n.d.), is to determine the validity of the research based on the author's use of citations and references, and whether or not the literature is classified as peer-reviewed or a refereed publication by Ulrich's Periodicals

Directory (Ulrichsweb™, n.d.). Ulrichsweb™ is the authoritative source of bibliographic and publisher information on more than 300,000 periodicals of all types: academic and scholarly journals, Open Access publications, peer-reviewed titles, popular magazines, newspapers, newsletters, and more from around the world (Ulrichsweb™, n.d.). Finally, the third step is to evaluate relevance to the topic by determining how broad or narrow the article is to the topic; is the information applicable or generalizable to the topic (University of Colorado at Boulder, n.d.).

Documentation Approach

Search results are captured in an electronic database using the software tool, BibMe. This method stores document information, including abstracts and other bibliographic detail, in APA format.

Resources are hand-coded and electronically stored using the Zoho® Creator software tool. Zoho® provides the ability to create forms and fields, and to conveniently sort by author, title, date, topic area, or any other named convention assigned to the information. Full-text articles are uploaded directly or scanned and saved in Zoho©, along with research notes and relevant keywords used to find the resource. This system allows all resources to be in one location and provides a quick, efficient tool to search through documents when reviewing the literature.

Resources are coded with the following naming convention: Topic_Year_Author_Word or Phrase (01-10 for 10 identified words or phrases) _Page Number (page on which information is found). For example, the article published by Kahn et al. (2002) is coded for information quality and data dimensions as follows: IQ_2002_KAH_02_187, where the phrase *dimensions of information quality* is coded as 02. The topic field is either assigned IQ, which refers to information quality; DM, which refers to data mining; DS, which refers to data storage; or BI,

which refers to business intelligence. The field for year refers to the year the reference was published. The first three letters of the author's last name are added as another method to quickly locate the article. The file naming convention and the research notes allow articles cited to be found quickly when more information is needed. Text for the words or phrases will be coded on the implication level of concepts with similar meaning to distinguish among concepts. Within each resource, level of analysis, relevant categories, existence of concepts, and level of implication are coded as outlined in the data analysis plan.

A summary of the documentation coding plan is illustrated in Table 3.

Table 3

Summary of Documentations Method and Coding Plan

Column	Variable	Code Description
1-2	Topic	BI = Business Intelligence, IQ = Information Quality, DM = Data Mining
3	(space marker)	__
4-7	Year	2001 through 2011
8	(space marker)	__
9-11	Author	First three letters of author's last name
12	(space marker)	__
13-14	Word or phrase (implication level of concepts with similar meaning coded the same)	01 = assurance, 02 = dimensions of information quality, 03 = decision-making, 04 = guidelines for information quality, 05 = framework for data management, 06 = data mining processes, 07 = business intelligence systems, 08 = benchmarks for effectiveness, 09 = information quality recommendations
15	(space marker)	__
16-18	Page Number	Page on which specific information is found

Data Analysis Plan

Creswell (2009) and Obenzinger (2005) describe qualitative data analysis as a form of content analysis that is a non-linear, iterative, progressive process in which the key is coding, sorting, and sifting through resources that satisfy evaluation criteria. The key components of an analysis plan, appropriate for application in a literature review, synthesizes old perspectives with

new ones by identifying research questions, selecting resources, and coding text into manageable categories (Busch et al., 2005; Ormondroyd et al., 2009; Obenzinger, 2005). The process of coding is basically one of selective reduction; reducing the text to categories consisting of a word, set of words, or phrases, focuses on and codes for, specific words or patterns that are indicative of the research question (Busch et al., 2005).

Key concepts addressing information quality assurance in the pre-processing stage of data storage to ensure the quality of content of information before reaching the warehouses are identified, classified, and coded in selected literature found in the Annotated Bibliography. The results are analyzed and synthesized to address identifying and prioritizing key dimensions of information quality for assurance in the pre-processing stage of data storage, with the underlying goal of increasing competitive advantages in the decision-making process across the broader context of BI (Negash, 2008).

The data analysis process is conducted in two stages on a single set of literature listed for coding in the Annotated Bibliography, according to the eight coding steps for conceptual analysis (Busch et al., 2005). Stage one is the process of selecting literature from the Annotated Bibliography that is most applicable to the development of this study (Ormondroyd et al., 2009). Stage two is reading and coding the selected literature to identify terms and phrases that are relevant to the purpose and research goals (Obenzinger, 2005). After references are identified, classified, and coded, the results are presented in the Conclusions section (Busch et al., 2005; Ormondroyd, 2009; Obenzinger, 2005).

Coding process. An eight-step process for coding text is used to identify the dimensions that most influence information quality assurance in the pre-processing stage of data storage within the context of BI (Busch et al., 2005).

1. Determine level of analysis – Analysis is conducted using words or phrases. The concept of BI is identified by the following sets of words or phrases: BI, decision makers, decision making, decision support system (DSS), informational advantage, competitive intelligence, competitive intelligence advantages, and competitive advantages. The concept of data or information quality is identified by the following words and sets of words or phrases: data, information, profiling, information systems, accuracy, completeness, consistency, integrity, and relevancy. The concept of data mining is identified by the following sets of words or phrases: data mining, patterns, extracting information, and transformation.
2. Decide how many concepts will be coded – A pre-defined set of three concepts is created, as detailed in step one, Level of Analysis. Concepts include: BI, data or information quality, and data mining. Additional concepts may be added to introduce a level of coding flexibility that permits important new material to be incorporated into the coding process (Busch et al., 2005).
3. Decide whether to code for existence or frequency of a concept – The text is coded for existence. For example, each dimension coded in relation to the concept of information quality assurance is counted once, no matter how many times it appears.
4. Decide on how to distinguish among concepts – Text is coded on the implication level of concepts with similar meaning. For example, information quality and data quality are similar enough to be coded as implying the same thing and thus will not require a separate category.

5. Develop rules for coding text – The coding process is streamlined and organized to ensure consistent and coherent coding throughout the text. For example, information is referred to as data, and both are coded in the same category.
6. Decide what to do with irrelevant information – Information deemed irrelevant to this study is considered immaterial and will be disregarded without impacting the outcome of the coding.
7. Code the texts – Terms and phrases are coded by hand first, and then entered in Zoho©, a free qualitative research software. Once key terms and phrases are established and entered, the program examines texts for data matching the parameters.
8. Analyze results – After the data is coded, conclusions and generalizations are summarized as specified in the Writing Plan section of this study and reported in the Review of Literature section.

Writing Plan

The final step in presenting the results of the data analysis process is to reflect on the themes in relation to the needs of the audience and describe the outcome of the study (Busch et al., 2005). This writing plan is designed to present and discuss the results compiled during the data analysis process through the use of a rhetorical pattern, as suggested by Obenzinger (2005). In particular, the writing plan in this study is designed thematically (Busch et al., 2005; Creswell, 2009; Obenzinger, 2005). An overview of current research is presented, as revealed by the conceptual analysis, followed by context based inferences to create meaning (Obenzinger, 2005). The objective is the development of a set of guidelines addressing dimensions with the most influence on information quality assurance in real-world BI environments in relation to two

overarching themes: (a) the role of information quality assurance in the pre-processing stage of data storage, and (b) key dimensions of information quality for assurance within the context of BI. Relevant dimensions identified during the identification, classification, and coding process will be discussed individually.

Two themes have been identified as most important to the study. Theme one examines the critical nature of the role of quality information in the pre-processing stage of data storage within the context of BI. The first theme is informed by broad searches of previous research on the importance of information quality in the pre-processing stage of data storage (Watson & Wixom, 2007). Popovic et al. (2009) propose and test a model of the relationship between BI and information quality, and investigate in more detail the potential differential impact of BI on two dimensions of information quality: the quality of content and the effects of quality assurance. Su et al. (2009) examine a methodology to determine key information quality dimensions, and provide models to examine how the precision, timeliness, and integrity of source data affect information quality in the pre-processing stage of data storage. Negash (2008) discusses a BI framework for the cleanup, search, analysis, and delivery of unstructured data, and explores a matrix of on-line analytical processing (OLAP) tools for use in the case of unstructured or semi-structured data for BI systems.

The second theme is informed by previous research from Cong et al. (2007) and McGilvray (2008) among others, and identifies the role of dimensions with the most influence on information quality in the pre-processing stage of data storage. Research is focused on identifying and describing dimensions of information quality that influence quality of content in the early stages of data preparation for storage and management (Negash, 2008; Olson, 2009). The role of dimensions is further analyzed as a way to identify and prioritize key dimensions for

assuring information quality early on in data storage within the context of BI. The objectives of each key dimension, along with those that should be considered by decision makers to gain competitive advantages, are described.

The goal of the writing plan is to organize the presentation of the results of the coding process in such a way as to address identifying and prioritizing key dimensions of information quality for assurance in the pre-processing stage of data storage (English, 2009; McGilvray, 2008). An outline of the thematic presentation format is as follows:

1. Theme one: The role of information quality assurance in the pre-processing stage of data storage within the context of BI.
    1.1. Examining the role of information quality.
    1.2. Discussing information quality for data storage.
    1.3. Discussing the impact of the quality of content.
2. Theme two: Key information quality dimensions for assurance in the pre-processing stage of data storage.
    2.1. Examining key information quality dimensions.
    2.2. Identifying and prioritizing key dimensions.
    2.3. Examining key dimensions for assurance in the pre-processing stage of data storage.

Annotated Bibliography

The references presented in this annotated bibliography are those that are judged to be the most significant to the identification of the dimensions that influence information quality assurance for data storage within the context of BI, specifically addressing problems prior to storage in data warehouses (Obenzinger, 2005; Ormondroyd et al., n.d.). This section, which consists of 24 key references, provides citations selected for use in the Review of Literature, and represents the core data set for coding as part of the larger content analysis (Luckey, 2009). A few additional references are coded (see Appendix B). References represent current knowledge about dimensions that influence information quality assurance in the pre-processing stage of data storage in an effort to support data mining, where the goal is to produce competitive advantages for BI systems (Obenzinger, 2005; Ormondroyd et al., n.d.). The annotated bibliography contains an abstract pulled directly from the reference, along with a content summary, a credibility assessment, and consideration of the relevance to this study (Stacks & Karper, 2008).

Caro, A., Calero, C., Caballero, I., & Piattini, M. (2008) A proposal for a set of attributes relevant for web portal data quality, *Software Quality Journal, 16*(4), 513-542. doi:10.1007/s11219-008-9046-7

Abstract. Data Quality is a critical issue in today's interconnected society. Advances in technology are making the use of the Internet an ever-growing phenomenon with the creation of applications such as Web Portals. These applications are important data resources and means of accessing information which many users employ to make decisions. Quality is a very important factor in any data software. As quality is a broad concept, quality models are typically used to assess the quality of a software product.

From the software point of view, there is a widely accepted standard proposed by ISO/IEC (the ISO/IEC 9126) for a quality model for data software products. Similar proposals for data quality are non-existent. Although proposals of data quality models exist, none focus specifically on web portal data quality and the user's perspective. In this paper, the authors propose a set of 33 attributes which are relevant for portal data quality. These have been obtained from a literature review and through a validation process carried out by means of a survey. Although these attributes do not conform to a usable model, it might be a good starting point for constructing one.

Comments. This article provides a framework for understanding the evolution of data quality. The authors discuss a variety of applications available for quality assurance, and examine the broad topic of quality overall. Caro, Calero, Caballero, and Piattini are professors of computer science, with a combined peer-reviewed publication count of over 350 articles pertaining to data quality software product applications. The article supports content development of the study by focusing on identifying dimensions that most influence information quality. *Software Quality Journal* is listed as an academic/scholarly refereed journal on Ulrichsweb™, and thus is considered to be a credible resource. It is classified within theme one to support the context of information quality.

Cong, G., Fan, W., Geerts, F., Jia, X., & Shuai, M. (2007). Improving data quality: Consistency and accuracy. *Proceedings of the 33rd International Conference on Very Large Databases (VLDB), Vienna, Austria, 2007,* 315-326. Retrieved from http://www.vldb.org/conf/2007/papers/research/p315-cong.pdf

Abstract. Two central criteria for data quality are consistency and accuracy. Inconsistencies and errors in a database often emerge as violations of integrity constraints. Given a dirty, or inconsistent, database *D*, applying automated methods make it consistent, i.e., find a repair *D′* that satisfies the constraints and minimally differs from *D*. Equally important is to ensure that the automatically-generated repair *D′* is accurate, or makes sense, i.e., *D′* differs from the correct data within predefined boundaries. This paper studies effective methods for improving both data consistency and accuracy. A class of conditional functional dependencies (CFDs) proposed to specify the consistency of the data is examined, which are able to capture inconsistencies and errors beyond what their traditional counterparts can catch. To improve the consistency of the data, two algorithms are proposed: one for automatically computing a repair *D′* that satisfies a given set of CFDs, and the other for incrementally finding a repair in response to updates to a clean database. Both problems are intractable. The resulting algorithms develop a statistical method that guarantees that the repairs found by the algorithms *are accurate above a predefined rate* without incurring excessive user interaction.

Comments. This article helps to clarify data mining techniques as well as the need for information quality; thus it is critical to the study as it presents a framework for data mining of information for BI. This article is highly technical, but is heavily cited with

references that inform the need for information quality assurance for data mining in BI. The authors discuss real-world data situations in which inconsistencies, conflicts, and errors affect data quality. The authors are professors of Web data management and are also all software engineers for The Database Group at the University of Edinburgh. The authors have over 50 years of combined experience in data cleansing and information quality. The Proceedings of the VLDB is a scholarly peer reviewed journal listed on Ulrichsweb™. Based on these criteria, this article is deemed a credible resource, and is generalizable to the broad topic of information quality.

Davenport, T.H., & Harris, J.G. (2007). The architecture of business intelligence. In *Competing on analytics: The new science of winning.* (chapter 8). Boston, MA: Harvard Business School Press. Retrieved from http://www.accenture.com/NR/rdonlyres/15DCFF6A-4DE0-44D8-B778-630BE3A677A2/0/ArchBIAIMS.pdf

Abstract. Many companies today are collecting and storing a mind-boggling quantity of data. The numbers are hard to fathom: in just a few years, the common terminology for data volumes has grown past projected amounts. However, while organizations have more data than ever at their disposal, they rarely impose sufficient order on it and thus get limited value from all that information. Further, many IT departments lack the capabilities to do more than support and maintain basic transactional and reporting capabilities. In short, while improvements in technology's ability to store data have been astonishing, most organizations struggle to manage, analyze, and apply it.

Comments. The authors provide an overview of the need for information quality assurance practices, and indicate that with the large volume of increased data, it is crucial to have an information quality management system in place in the beginning. Besides

authoring 13 books (including the first books written on knowledge management) and hundreds of articles for refereed journals, Davenport was named one of the top 25 consultants in the world by *Consulting Magazine* in 2003. In 2007 and 2008, he was named one of the 100 most influential people in the IT industry by Ziff-Davis publishers, one of the world's premier publishers of technology-based digital content products. Harris is a senior executive research fellow; her work is published in numerous refereed journals and is quoted extensively by the *Wall Street Journal, Forbes Magazine, CIO Magazine*, and many others. This book is considered a credible resource based on the professional and academic achievements of both authors and the extensive use of citations from and references to peer reviewed published works. The content in this book provides a framework for the critical need for information quality in the pre-processing stages of data storage. It is classified within theme two and is generalizable to the topic of BI.

English, L. (2009). *Information quality applied: Best practices for improving business information, processes and systems.* New York, NY: John Wiley & Sons, Inc.

Abstract. In this book, the author takes a hands-on approach, showing how to apply the concepts outlined in his first book, *Improving Data Warehouse and Business Information Quality,* to specific business areas like marketing, sales, finance, and human resources. The book presents real-world scenarios with examples for melding data quality concepts to specific business areas such as supply chain management, product and service development, customer care, and others. Step-by-step instructions, practical techniques, and helpful templates from the author enable the application of best practices for businesses to begin immediate modeling of quality initiatives.

Comments. The author's explanation of how to ensure information quality for BI and data mining focuses on maintaining the quality and accuracy of business data by conducting information quality assessments in the pre-processing stage to allow time for correction initiatives and adequate preparation for mining. He offers IT, database, and business managers step-by-step instructions for setting up methodical and effective procedures. Templates are included for businesses to model their own quality initiatives. A companion Web site provides templates, updates to the book, and links to related sites. English has extensive academic experience with information quality systems and management and is an internationally recognized speaker, educator, author, and consultant in knowledge management and information quality improvement. He developed an information quality system that was the basis for Six Sigma, and was awarded the 1998 Individual Achievement Award for his contributions to Information Management. His refereed published works are regularly cited in peer reviewed journals in over 40 countries on six continents. Thus this is deemed to be a valid resource and is considered to be significantly relevant to framing the context of information quality assurance for BI. This article provides a framework for assuring quality information in the pre-processing stage of data mining and is classified within theme one.

Fisher, C., Lauria, E., Chengalur-Smith, S., & Wang, R. (2008). Introduction to information quality (4th ed.). Cambridge,MA: MIT Press.

Abstract. This book educates readers about the critical issues in data and information quality that have been plaguing information systems for many years. Researchers have only recently begun to address data quality as a discipline in its own right, and a body of data quality literature has just begun to appear. Researchers at Massachusetts Institute of

Technology (MIT) began a total data quality management program and have hosted ten international conferences on information quality aimed at practitioners, academicians, and researchers. This book is built on two primary sources. After an extensive literature review and study, an importance of data quality knowledge and skills survey was completed by 110 data quality researchers and practitioners, all data quality leaders in their own right, at the International Conference on Information Quality held at MIT. The results of these studies led to a consensus of the most critical skills necessary to begin performing information quality work. An introduction to those critical skills and knowledge areas are the primary topics of this book. The second source is the research into data and information quality of the four authors who collectively have published over 100 articles.

Comments. The book discusses the need to address data quality practices in businesses and organizations. The authors are convinced that an organized discipline for data and information quality is required. The contents of this book provide a broad basis for understanding the concepts and philosophy of data and information quality. Tools and techniques are introduced that are essential for a data quality analyst to make improvements. Authority is established based on the credentials, education, and experience of the authors: all hold a Ph.D. in information science; all are regularly published and frequently cited in peer reviewed journals; and all have at least 20 years of experience each in the IT arena working for multinational firms including Microsoft, IBM, and Hewlett Packard. Validity is established by the use of multiple citations and references from refereed publications. The book is relevant to the topic of information quality assurance and is classified within theme one.

Hakim, L. (2007a). *Information quality management: Theory and applications*. Hershey, PA: Idea Group Publishing.

Abstract. This book provides insights and support for professionals and researchers working in the field of information and knowledge management, information quality, practitioners and managers of manufacturing, and service industries concerned with the management of information.

Comments. This book offers tips for information quality assurance, and helps to structure and inform key dimensions influential for information quality that are identified in this study. It suggests ways in which different professionals working in information quality management can manage information effectively. It offers advice and recommendations, and describes best practices beneficial to knowledge management professional. The author's education and professional experiences in information quality and management span industry, research, and development over various academic institutions. His research is extensively published in refereed journals; he is the author of more than 60 papers published in peer reviewed journals and books. He is considered to be an expert in the field. His use of citations and references to peer reviewed work establishes validity. Thus, this book is a trusted resource based on the author's extensive professional and academic experience with knowledge management. Relevant text and key information is classified within theme one; results are extrapolated and coded in the dataset.

Jafar, M.J. (2010). A tools-based approach to teaching data mining. *Journal of Information Technology Education: Innovations in Practice, 9*, 2-24. Retrieved from http://jite.org/documents/Vol9/JITEv9IIPp001-024Jafar740.pdf

Abstract. Data mining is an emerging field of study in Information Systems programs. Although the content has been streamlined, the underlying technology is still in a state of flux. The paper describes how Microsoft Excel's data mining add-ins as a front-end to Microsoft's Cloud Computing and SQL Server 2008 BI platforms as back-ends is used to teach a senior level data mining methods class. The content presented and the hands on experience gained have broader applications in other areas, such as accounting, finance, general business, and marketing. Business students benefit from learning data mining methods and the usage of data mining tools and algorithms to analyze data for the purpose of decision support in their areas of specialization. Newly introduced capabilities to faculty currently teaching a BI course are highlighted. This set of integrated tools allowed focus on teaching the analytical aspects of data mining and the usage of algorithms through practical hands-on demonstrations, homework assignments, and projects. As a result, students gained a conceptual understanding of data mining and the application of data mining algorithms for the purpose of decision support. Without such a set of integrated tools, it would have been prohibitive for faculty to provide comprehensive coverage of the topic with practical hands-on experience. The availability of this set of tools transformed the role of a student from a programmer of data mining algorithms to a BI analyst. Students now understand the algorithms and use tools to perform (1) elementary data analysis, (2) configure and use data mining computing engines to build, test, compare and evaluate various mining models, and (3) use the mining models to analyze data and predict outcomes for the purpose of decision support. If it was not for the underlying technologies that were used, it would have been impossible to cover such material in a one-semester course and provide students with

much needed hands-on experience in data mining. Finally, utilizing the cloud as a computing platform that transformed the role of a student from "doing low-level IT" in a data mining course to a BI analyst using tools to analyze data for the purpose of decision support is described.

Comments. The authors teach students how to analyze data and use data mining tools to predict outcomes for the purpose of decision support. This informs the broader context of BI in the literature review, and is relevant to the study because it focuses on using a set of integrated tools together with the analytical aspects of data mining to benefit BI. The author is a professor of computer information systems and is published extensively in refereed publications. He is considered an expert in the field of data mining by the academic community of higher education. The validity of this article is established because it is listed as an academic, scholarly refereed journal on Ulrichsweb™, and it is heavily cited with prior research in the area of data mining and data analysis. The relevance of the article is to data mining in general, and it is classified within theme one.

Keeton, K., Mehra, P., & Wilkes, J. (2009). Do you know your IQ: A research agenda for information quality in systems. *ACM Sigmetrics Performance Evaluation Review, 37*(3), 1-6. Retrieved from http://www.sigmetrics.org/sigmetrics/workshops/papers_hotmetrics/session1_4.pdf

Abstract. Information quality (IQ) is a measure of how fit information is for a purpose. Sometimes called Quality of Information (QoI) by analogy with Quality of Service (QoS), it quantifies whether the correct information is being used to make a decision or take an action. Not understanding when information is of adequate quality can lead to bad decisions and catastrophic effects, including system outages, increased costs, lost

revenue – and worse. Quantifying information quality can help improve decision making, but the ultimate goal should be to select or construct information producers that have the appropriate balance between information quality and the cost of providing it. In this paper, a brief introduction to the field of data mining is presented, the case for applying information quality metrics in the systems domain is argued, and a research agenda to explore this space is proposed.

Comments. The authors indicate the need for determining whether information is good enough to lead to results that will help decision makers inform their process. They note that poor information can lead to bad results, but good information may be costly to acquire. As such, the authors introduce the field of information quality and suggest ways it can be measured and used. Keeton is a Senior Researcher in the Storage and Information Management Platform Lab at HP Labs. Her research focuses on simplifying the management of enterprise information systems. Mehra has over 20 years of large-scale systems and software design experience with HP Labs, and has won numerous awards and honors for articles published in refereed journals. Wilkes was with HP Labs for 26 years, before he left to join Google. He has written three books, and his publications are found in refereed journals. His research in self-managing storage systems paved the way for open cloud-computing. This article is published in a peer reviewed journal listed on Ulrichsweb™; thus it meets evaluation criteria for validity. This is a pivotal article for establishing the context of the decision-making process, and is classified within theme one.

Kriegel, H. P., Borgwardt, K.M., Kroger, P., Pryakhin, A., Schubert, M., & Zimek, A., (2007). Future trends in data mining. *Data Mining and Knowledge Discovery, 15*(1) 87-97. doi:10.1007/s10618-007-0067-9

Abstract. Over recent years data mining has been establishing itself as one of the major disciplines in computer science with growing industrial impact. Undoubtedly, research in data mining will continue and even increase over coming decades. In this article, we sketch a vision of the future of data mining. Starting from the classic definition of data mining, topics that will set trends in data mining are discussed.

Comments. The authors provide an excellent classic description of data mining in this article. They address data mining approaches to complex objects as well as dynamic real-world systems. Furthermore, they discuss pre-processing as the most important and essential part of data mining. Pre-processing is a critical part of information quality, and the authors conclude that the techniques used in pre-processing can deeply influence the results of the actual data mining analysis. The authors are all part of an international research group that focuses on database and information management systems. The group is ranked by the ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) among the top-10 in the world, second in Europe, and top-ranked in Germany. *Data Mining and Knowledge Discovery* is listed as an academic/scholarly refereed journal on Ulrichsweb™, and is heavily cited with prior research studies; thus it meets the evaluation criteria for validity. The context of the article is relevant to the study and is classified within theme one.

Lee, Y.W., Pipino, L.L., Funk, J.D., & Wang, R.Y. (2009). *Journey to data quality*. Cambridge, MA: MIT Press.

Abstract. All organizations today confront data quality problems. Neither ad hoc approaches nor fixes at the systems level installing the latest software or developing an expensive data warehouse solve the basic problem of bad data quality practices. *Journey to Data Quality* offers a roadmap that can be used by practitioners, executives, and students for planning and implementing a viable data and information quality management program. This practical guide, based on rigorous research and informed by real-world examples, describes the challenges of data management and provides the principles, strategies, tools, and techniques necessary to meet them. The authors, all leaders in the data quality field for many years, discuss how to make the economic case for data quality and the importance of getting an organization's leaders on board. They outline different approaches for assessing data, both subjectively (by users) and objectively (using sampling and other techniques). They describe real problems and solutions, including efforts to find the root causes of data quality problems at a healthcare organization and data quality initiatives taken by a large teaching hospital. They address setting company policy on data quality and, finally, they consider future challenges on the journey to data quality.

Comments. This book is a practical guide that offers strategies for planning and implementing a viable data and information quality management program. According to the *ACM Journal of Data and Information Quality*, the authors are leaders in the data quality field and have many years of combined experience with different approaches to assess data both subjectively and objectively. Analysis of their research in conducting in-depth analyses of the role of data security in enterprise information quality at Massachusetts Institute of Technology (MIT) is published in numerous books and

refereed journals. This book is deemed valid to the study based on authority, and myriad citations and peer reviewed references throughout the book. It is relevant to and supports the framework of data quality, and is classified within theme one.

Lupu, A.R., Razvan, B., Sabau, G., & Muntean, M. (2007). Influence factors of business intelligence in the context of ERP projects, *International Journal of Education and Information Technologies, 2*(1), 90-94. Retrieved from http://www.naun.org/journals/educationinformation/eit-15.pdf

Abstract. BI projects are very dynamic and during their development may encounter many environmental, technological, and personnel changes. All of these changes determine the need for progressive planning and an iterative development approach. This article presents the development of a real industry BI project in a company that used an ERP system. It focuses on the main factors that influence and affect project development and it analyses the system evolution from technical point of view. The description of this particular experience is useful to all those who are involved in building BI solutions to reveal success factors.

Comments. The authors establish the context of BI in an integrated business environment and present a case study involving a real experience of developing a large BI project, along with the analysis of difficulties and problems. Technical solutions are provided along with direction for future research in BI. The real-world examples clarify relationships and further the understanding of BI systems. The authors are professors of data and information systems integration and are internationally recognized leaders in the field, as noted by the *International Journal of Education and Information Technologies.* They have published numerous articles in refereed journals, and are respected speakers

on the topic world-wide. This resource is considered credible because it is published in a peer reviewed journal listed on Ulrichsweb™, and the authority of the authors establishes validity. The relevance of the article is generalizable to BI in a broad sense, and is classified within theme two because it helps to establish the framework for the overall context of information quality in BI.

McGilvray, D.M. (2008). *Executing data quality projects: Ten steps to quality data and trusted information.* Burlington, MA: Morgan Kaufmann Publishers.

Abstract. In this book the author presents a thorough understanding of significance of information quality in the world today. She describes the impact of information quality on the ability to make effective business decisions, and notes that with flawed, incomplete, or misleading data, information cannot be trusted to further business goals and objectives.

Comments. This book provides a systematic approach for improving and creating data and information quality within businesses. It provides a central role in identifying dimensions that influence information quality. It explains a methodology that combines a conceptual framework for understanding information quality with the techniques, tools, and instructions for improving and creating information quality. The author presents a ten-step process for implementing the concepts she describes. McGilvray has extensive professional experience in information quality management and data governance and is recognized as a leader in the field by Fortune 50 organizations. She is an accomplished program manager and facilitator, and is an internationally respected expert on data profiling, metrics, quality, audits, benchmarking, and tool acquisition and implementation. The use of citations and references from peer reviewed journals

throughout her book establishes the validity of this resource. Relevant text provides a basis for understanding the necessity of information quality assurance in the pre-processing stage. The context of the book is relevant to the topic of information quality and is classified within theme one.

Negash, S. (2008). Handbook on decision support systems 1: Business intelligence. In *International handbooks on information systems*. (chapter 45). Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-540-48713-5

Abstract. Business intelligence (BI) is a data-driven decision support system (DSS) that combines data gathering, data storage, and knowledge management with analysis to provide input to the decision making process. The term originated in 1989; prior to that many of its characteristics were part of executive information systems. BI emphasizes analysis of large volumes of data about the company and its operations. It includes competitive intelligence (monitoring competitors) as a subset. In computer-based environments, BI uses a large database, typically stored in a data warehouse or data mart, as its source of information and as the basis for sophisticated analysis. Analyses ranges from simple reporting to slice-and-dice, drill down, answering ad hoc queries, real-time analysis, and forecasting. A large number of vendors provide analysis tools. Perhaps the most useful of these is the dashboard. Recent developments in BI include business performance measurement (BPM), business activity monitoring (BAM), and the expansion of BI from being a staff tool to being used by people throughout the organization (BI for the masses). In the long-term, BI techniques and findings will be imbedded into business processes.

Comments. The author presents a definition of BI, describes its purpose, and provides an architectural framework. The costs and benefits of BI systems are weighed, and competitive analyses are presented. Focus is placed on techniques and applications that support informed actions by decision makers. The author is well respected in the field of BI as an expert, with over 75 published articles in peer reviewed journals and refereed conference proceedings. The author cites peer reviewed references through the article. The information provided in this article is classified within theme two and is generalizable to a broad framework of BI. This resource is considered credible because it is published in a peer reviewed journal listed on Ulrichsweb™, and the authority of the author establishes validity. This article adds to knowledge of BI in general and therefore is deemed relevant to the study.

Olson, J.E. (2003). *Data quality: The accuracy dimension*. San Francisco, CA: Morgan Kaufmann Publishers.

Abstract. This book describes techniques for assessing the quality of corporate data and improving its accuracy using the data profiling method. Corporate data is increasingly important as companies continue to find new ways to use it. Likewise, improving the accuracy of data in information systems is becoming a major goal as companies realize how much it affects their bottom line. Data profiling is a new technology that supports and enhances the accuracy of databases throughout major IT shops. The author explains data profiling and shows how it fits into the larger picture of data quality.

Comments. This book provides a thorough understanding of data accuracy in real-world environments and provides a framework for data profiling. It describes analytical tools appropriate for assessing data accuracy. The author has over 36 years of experience

developing commercial software and tools for data management systems. He is an early pioneer of data profiling and has developed concepts for building an understanding of databases at the content, structure, and quality levels. He is considered an expert in the field of database management systems by publishers and the data management arena. The book is heavily cited and references to peer reviewed journals appear throughout it; thus this is deemed to be a valid and credible resource for this study. The content is relevant to data quality and mining and is classified within theme one.

Olson, J.E. (2009). *Database archiving: How to keep lots of data for a very long time*. San Francisco, CA: Morgan Kaufmann Publishers.

Abstract. This book is about database archiving for large database applications. The types of organizations that benefit from building a database archiving practice are any that have long-term retention requirements and lots of data. This includes most public companies and those that are private but that work in industries requiring retention of data (such as medical, insurance, or banking fields). It also includes educational and government organizations.

Comments. This book represents the author's view of the current state of thinking on the topic of database archiving. Database archiving is a new and growing field within data management. The author points out that data archived today will take years to grow old enough to expose some of the flaws in current thinking, and that it is critical to establish database archiving practices now. Olson has over 36 years experience developing commercial software and tools for data management systems, and is considered an expert in the field of database management systems. Similar to *Data quality: The accuracy dimension,* this book is heavily cited and refereed journal references appear throughout it.

The content is relevant to data quality and mining; thus this is deemed to be a valid and credible resource for this study, and is classified within theme one.

Piatetsky-Shapiro, G., Djeraba, C., Getoor, L., Grossman, R., Feldman, R., & Zaki, M. (2009). What are the grand challenges for data mining? *SIGKDD Explorations, 8*(2), 70-77. doi:10.1145/1233321.1233330

Abstract. The authors create grand challenge problems for data mining, and then propose criteria for solutions. They consider possible grand challenge problems from multimedia mining, link mining, large-scale modeling, text mining, and proteomics.

Comments. This article builds a framework for understanding problems facing data mining processes. The authors take a real-world perspective to create potential problems and then consider solutions on a broad scale. Their research spans many different approaches to data mining that address the need for tools and techniques for intelligence data understanding. Piatetsky-Shapiro is considered to be one of the founders of data mining and knowledge discovery fields and has extensive experience developing data analysis models for banks, insurance companies, and pharmaceutical companies. He has served as an expert witness and provided expert opinions in several cases. He has over 60 publications in refereed journals, including two best-selling books and several edited collections on topics related to data mining and knowledge discovery. Djeraba has produced over 150 publications in book chapters, conferences, and peer reviewed journals. Getoor's research interests are in machine learning, databases and artificial intelligence, with over 150 publications in refereed arenas. Grossman is involved in open source project in data intensive computing. Research accomplishments include developing scaled tree-based classifiers to very large data sets and the introduction of

infrastructures for deploying statistical data mining models for BI. Feldman and Zaki are researchers specializing in the development of text mining tools and applications; they have over 75 combined papers on the topic published in refereed journals. This resource is considered credible because it is published in a peer reviewed journal listed on Ulrichsweb™. Authority and validity are established. It is deemed relevant to the topic of data mining and is classified within theme one.

Popovic, A., Coelho, P.S., & Jaklic, J. (2009). The impact of business intelligence system maturity on information quality. *Information Research, 14*(4), 1-14. Retrieved from http://informationr.net/ir/14-4/paper417.html

Abstract. A model of the relationship between BI systems and information quality is proposed and tested. The potential differential impact of BI systems' maturity on two aspects of information quality, content quality and media quality, is investigated in more detail. The results indicate that the implementation of a BI system positively affects both aspects of information quality as conceptualized in the model. However, the effect of BI systems' maturity is greater on media quality than on content quality. Since most of the information quality problems in knowledge-intensive activities relate to content quality, it is reasonable to expect that the implementation of BI systems would adequately address these problems. However, the effects of implementing such systems seem to be more focused on media quality outcomes. Based on the findings, it is suggested that projects implementing BI systems need to focus more on ensuring content quality.

Comments. The authors discuss the implementation of BI systems and whether or not BI adequately addresses all the information quality problems that knowledge workers most often encounter. The focus is on whether the implementation of BI technologies

and related data management activities contribute to the ability to access information, and whether it focuses adequately on the content aspects of information quality. The authors have published over 60 papers in refereed journals, with main research focuses on Web-based information systems applications, techniques, and tools for decision makers. The article is heavily cited and references peer reviewed journals. *Information Research* is listed as an academic, scholarly refereed journal on Ulrichsweb™; therefore, this article is deemed credible for use in the study. This pivotal article focuses on the problems with which decision makers are most faced, is relevant to the study, and is classified within theme two.

Rodriguez, C., Daniel, F., Casati, F., & Cappiello, C. (2010). Toward uncertain business intelligence: The case of key indicators. *IEEE Internet Computing, 14*(4), 32-40. http://doi.ieeecomputersociety.org/10.1109/MIC.2010.59

Abstract. Enterprises widely use decision support systems (DSS) and, in particular, BI techniques for monitoring and analyzing operations to understand areas where the business is not performing well. These tools are often unsuitable in scenarios involving Web-enabled, intercompany cooperation and IT outsourcing, however. The authors analyze how these scenarios impact information quality in BI applications and lead to nontrivial research challenges. They describe the idea of uncertain events and key indicators and present a model to express and store uncertainty and a tool to compute and visualize uncertain key indicators.

Comments. The authors summarize the factors that are critical to a company's performance, and how those key indicators can be used to detect problems and trigger business decisions. The specificity of the indicators increases knowledge, which in turn

leads to ensuring information quality for effective BI. The authors are well-known researchers, particularly for their work with Intelligent Business Operations Management in the Information Services and Process Innovation Lab at HP Labs. Combined, they have over 250 papers published in books, in conference proceedings, and in refereed journals. *IEEE Internet Computing* is listed as an academic, scholarly refereed journal on Ulrichsweb™ and thus this is considered to be a credible and valid resource for the study. The article's content and extensive bibliographic information is generalizable to the topic of BI, and is classified within theme two.

Sen, A., & Sinha, A.P. (2007). Toward developing data warehousing process standards: An ontology-based review of existing methodologies. *IEEE Transactions on Systems, Man and Cybernetics: Part C, Applications and Reviews, 37*(1), 17-31. http://dx.doi.org/10.1109/TSMCC.2006.886966

Abstract. A data warehouse is developed using a data warehousing process (DWP) methodology. Currently, there are a large number of methodologies available in the data warehousing market, in part due to the lack of any centralized attempts at creating platform-independent DWP standards. For the development of such standards, it is very important that current practices being followed by the data warehousing industry are first examined. In this study, 30 commercial data warehousing methodologies are reviewed and the standard practices they have adopted with respect to DWP are analyzed. The study provides valuable insights into the prevailing standard practices for different DWP task-system development, requirements analysis, architecture design, data modeling, ETL, data extraction, and end-user application design-and identifies important directions for future research on DWP standardization.

Comments. In this article, the authors provide a framework for understanding data mining and data warehouses. The authors foresee the need to develop a methodology that standardizes current practices. The authors have over 100 papers between them that are published in refereed journals. They are both well-known in the field and are considered respected researchers in data mining standards. The journal is listed as an academic, scholarly refereed journal on Ulrichsweb™; therefore, it is deemed credible for use in the study and is classified within theme one.

Seng, J.L., & Chen, T.C. (2010). An analytic approach to select data mining for business decisions. *Expert Systems with Applications, 37*(12), 8042-8057. http://dx.doi.org/10.1016/j.eswa.2010.05.083

Abstract. Due to the information technology improvement and the growth of the internet, businesses are able to collect and store huge amounts of data. Using data mining technology to aid the data processing, information retrieval, and knowledge generation process has become one of the critical missions to businesses. Proper use of data mining tools properly is now the primary user concern. Since not every user completely understands the theory of data mining, choosing the best solution from the functions data mining tools provides is not easy. A selection model of data mining algorithms is proposed. By analyzing the content of business decision and applications, user requirements will map to certain data mining category and algorithm. This method makes algorithm selection faster and reasonable to improve the efficiency of applying data mining tools to solve business problems.

Comments. The authors present a selection model of data mining designed to save users time and money by analyzing the content of a business decision and presenting a specific

data mining strategy. They believe that their method improves efficiency in applying data mining tools to solve business problems. This article clarifies the relationship between data mining and information quality, which is central to the study as it defines the framework for data mining strategies. *Expert Systems with Applications* is listed as an academic, scholarly refereed journal on Ulrichsweb™, establishing validity and authority. This article is relevant to the broader topic of the BI decision-making process and is classified within theme two.

Stvilia, B., Gasser, L., Twidale, M.B., & Smith, L.C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*(12), 1720-1733. doi:10.1002/asi.20652

Abstract. One of the main components in information quality (IQ) assurance is an IQ measurement model design. One cannot manage information quality without first being able to measure it meaningfully and establishing a causal connection between the source of IQ change, the IQ problem types, the types of activities affected, and their implications. A better understanding is needed of the roots of IQ change through the development of a systematic, predictive, reusable IQ assessment framework. The framework should enable effective IQ reasoning through the disambiguation of IQ problem resources, and through the rapid and inexpensive development of context-specific IQ measurement models. A general IQ assessment framework is proposed in contrast to context-specific IQ assessment models, which usually focus on a few variables determined by local needs. The proposed model's framework consists of comprehensive typologies of IQ problems, related activities, and a taxonomy of IQ dimensions organized in a systematic way based on sound theories and practices. The

framework can be used as a knowledge resource and as a guide for developing IQ measurement models for many different settings.

Comments. Sources of information quality problems are analyzed and solutions are identified with the use of decision-tree models. Types of activities affected by information quality problems are discussed, and direction for future research is presented in the form of case studies. Specific research interests for the authors include information quality, metadata and ontologies, information retrieval, and digital data curation. Together they contributed to the design of the Theory of Information Quality, and are known as experts in the field of information quality. The *Journal of the American Society for Information Science and Technology* is listed as an academic, scholarly refereed journal on Ulrichsweb™; thus validity is established and the article is deemed credible. The authors present a framework for information quality assessment that contributes to the understanding of the focus of this study. It is central to the framework for describing and identifying information quality measures and therefore is classified within theme one.

Su, Y., Peng, J., & Jin, Z. (2009). Modeling information quality risk for data mining in data warehouses. *Human & Ecological Risk Assessment, 15*(2), 332-350. doi: 10.1109/ICISE.2009.755

Abstract. Information Quality (IQ) is a critical factor for the success of many activities in the information age, including the development of data warehouses and implementation of data mining. The issue of IQ risk is recognized during the process of data mining; however, there is no formal methodological approach to dealing with such issues. Consequently, it is essential to measure the risk of IQ in a data warehouse to

ensure success in implementing data mining. This article presents a methodology to determine three IQ risk characteristics: accuracy, comprehensiveness, and non-membership. The methodology provides a set of quantitative models to examine how the quality risks of source information affect the quality for information outputs produced. It can be used to determine how quality risks associated with diverse data resources affect the derived data.

Comments. The authors discuss their development of quantitative models to confirm information quality risks for data mining in data warehouses. This establishes a connection between information quality and data mining. The connection helps to describe the larger context within which decision making resides. The study also proposes that two important system design factors, control transparency and outcome feedback, will incrementally influence perceived information quality. The quality checks listed in this paper are presented in the form of risk measures to have in place prior to data mining. The analysis process is usable in business data mining environments to determine how information that is mined identifies datasets with acceptable quality. The authors have extensive experience designing models for information quality assurance and have published over 100 papers in refereed journals. *Human and Ecological Assessment* is listed as an academic, scholarly refereed journal on Ulrichsweb™, establishing validity for this resource. Based on this criteria, the article is deemed credible and is classified within theme one.

Watson, H.J., & Wixom, B.H. (2007). The current state of business intelligence. *Computer Society, 40*(9), 96-99. http://dx.doi.org/10.1109/MC.2007.331

Abstract. BI is now widely used to describe analytic applications. BI has become a strategic initiative and is now recognized by CIOs and business leaders as instrumental in driving business effectiveness and innovation. BI is a process that includes two primary activities: getting data in and getting data out. Getting data in, traditionally referred to as data warehousing, involves moving data from a set of source systems into an integrated data warehouse. Getting data in delivers limited value to an enterprise; only when users and applications access the data and use it to make decisions does the organization realize the full value from its data warehouse. Thus, getting data out receives most attention from organizations. This second activity, which is commonly referred to as BI, consists of business users and applications accessing data from the data warehouse to perform enterprise reporting, OLAP, querying, and predictive analytics.

Comments. This article describes the BI process, beginning with a description of the role, characteristics, benefits, and suitability of data warehouses. Successes and failures of data warehouses are presented, and data analysis and knowledge discovery are defined. Data mining is described in detail, and a sample of data mining applications is presented. Watson helped develop much of the conceptual foundation for decision support systems (DSS) in the 1970's and applied his knowledge and expertise to executive information systems in the 1980's, making him a recognized leader in information management, and one of the world's leading scholars and authorities on decision support. He is the author of over 25 books and over 100 scholarly refereed journals. Wixom research is also recognized a s a leader in the industry, and has published over 70 papers in peer reviewed journals. *Computer Society* is listed as an academic, scholarly refereed journal on

Ulrichsweb™; therefore the article is deemed a credible resource for this study. It provides a framework of the BI process that is central to the framework of this study.

Zhao, Y., Chen, Y., & Yao, Y. (2006). *User-centered interactive data mining*. IEEE International Conference on Cognitive Informatics 2006, 457-466. http://dx.doi.org/10.1109/COGINF.2006.365532

Abstract. While many data mining models concentrate on automation and efficiency, interactive data mining models should focus on adaptive and effective communications between human users and computer systems. The crucial point is not how intelligent users are, or how efficient systems are, but how well these two parts can be connected, adapted, understood, and trusted. Some fundamental issues including processes and forms of interactive data mining, roles, requirements, as well as complexities of interactive data mining systems are discussed in this paper.

Comments. This article provides a framework for the efficiency of data mining systems. The authors explore the requirements and forms of different data mining systems, with a focus on the connection between users and systems. Zhao has published over 70 papers in peer reviewed journals and refereed conference proceedings; his research interests are in data analysis and computational engineering. Chen and Yao have published numerous papers in refereed journals; their interests include data mining methodologies and conceptual data analyses. This article is deemed credible because it is published in a peer-reviewed journal. It is classified within theme two to establish a connection between data mining and the decision-making process of BI.

Review of the Literature

The underlying assumption of this study is that establishing effective information quality in the pre-processing stage assures capitalization of advantages and opportunities in the form of increased ROI and competitive advantage gains for BI (Keeton, Mehra, & Wilkes, 2009). Thus the review of the literature begins by examining the impact of information quality assurance. Next, two primary themes are examined: the first theme frames the context of the importance of information quality assurance in the pre-processing stage of data storage; the second theme describes the key dimensions with the most influence on information quality in the pre-processing stage of data storage.

Information Quality Assurance

Business decisions are based on data regardless of whether that information is poor or high-quality (McGilvray, 2008). However, according to English (2008), Keeton et al. (2009), McGilvray (2008) and others, effective business decisions and actions are made when they are based on high-quality information. The concept of information quality is the degree to which information and data are a trusted source for decision makers to effectively run the business, to serve customers, and to achieve and meet goals and objectives (McGilvray, 2008). Thus assuring information quality for decision makers is essential to successful BI (Davenport & Harris, 2007). Figure 2 depicts the concept of information quality assurance for competitive advantages in BI as data passing through an information quality dimension filter; the resulting information aids in the decision-making process to ensure BI goals and objectives are met (K. Brown, AIM Program instructor, personal communication, November 28, 2010).

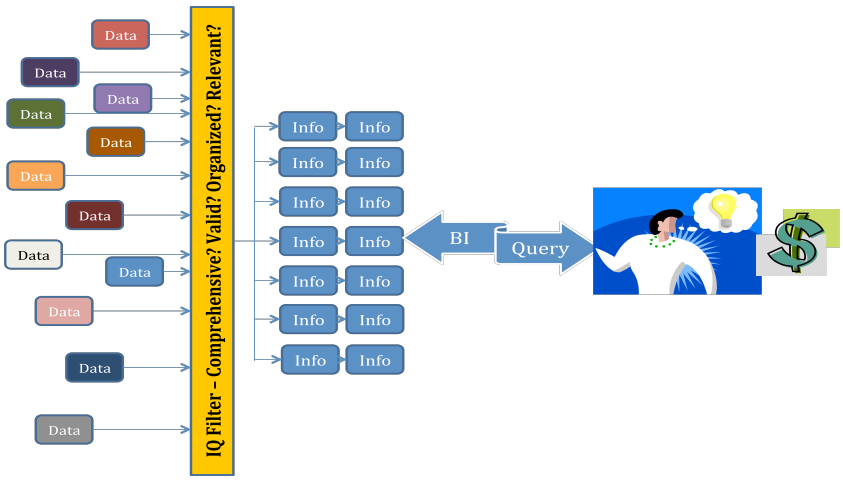


Figure 2. The concept of information quality as a trusted source for decision makers to meet BI goals and objectives (K. Brown, AIM Program instructor, personal communication, November 28, 2010)

Lefebvre (2007) contends that successful decision makers are familiar with information quality assurance and data mining techniques in the business environment in order benefit from focusing on the dimensions that most influence information quality assurance. Moreover, according to Kriegel et al. (2007) and Lefebvre (2007), the degree to which BI is successful depends on the objective characteristics of the audience and the focus placed on identifying and prioritizing specific key dimensions that align with goals and objectives. For example, media consumption habits, attitudes, and personal Web site preferences are characteristics of audiences that must be systematically and quantifiably identified and prioritized in order to have a higher degree of BI success (Kriegel et al., 2007; Lefebvre, 2007). Thus, the specific audience for this study is broadly described as business and IT professionals, managers, and non-management specialists who are involved in increasing competitive advantages for BI through informed decision making (Lefebvre, 2007).

Although various aspects of quality and information exist, there is a critical need for a methodology that assures a uniquely consistent definition, identification, and prioritization of

quality of content for individual BI systems (Kahn et al., 2002). Key dimensions identified from such a methodology provide priorities for assessing and improving information quality procedures (Cong et al., 2007; Lupu et al., 2007).

Information quality assurance affects the level of success of a business, and thus is the most important aspect of any company (Davenport & Harris, 2007). By developing and improving quality of content, businesses gain an understanding of different policies and practices in information quality assurance followed by organizations across the world (Fisher et al., 2008). BI strategies can be formulated by keeping ahead of the competition through the framework used to develop assurance guidelines (Hakim, 2007a; Negash, 2008). Good data are needed to inform the design of the decision-making process and to monitor and evaluate the quantitative progress toward goals and objectives; poor or unstructured data can mislead decision makers and result in loss of competitive advantages (Jafar, 2010).

According to Fisher et al. (2008) and Negash (2008), attention to key information quality dimensions ensures that goals and objectives are informed by valid information and that those BI systems are collecting and organizing data in the same manner. Furthermore, Negash (2008) notes that data are unique for each business; thus, if data are correct and managed, competitive advantages increase. Information quality assurance requires true continuous assessment; as such, successfully planning and implementing information quality assurance is an iterative approach (McGilvray, 2008).

Assuring information quality means that data must adequately represent dimensions that are inherent to the BI system goals and objectives (Olson, 2003). The dimensions are ubiquitous and influence information quality regardless of the unique BI system plan (McGilvray, 2008). For example, in the real world, plans are implemented and processes are designed to produce

quantifiable results (Keeton et al., 2009). The BI system collects and analyzes the results for the decision-making process by identifying and prioritizing the dimensions that fundamentally influence and assure information quality within the context of its goals and objectives (Andersson et al., 2008). Thus information quality is defined as the accuracy with which the BI system represents the real world (McGilvray, 2008; Negash, 2008).

Information Quality Awareness

The quality of data and validity of results for BI systems rely on assuring information quality in the pre-processing stage of data storage (Lupu et al., 2007). However, the continued growth of data warehouse storage capabilities increases the volume of information available for decision makers, which may not always be of the highest quality; as a result, data mining processes and applications require a framework for assuring quality of content (Zhao et al., 2006). To remedy unstructured or low-quality data, Olsen (2003) calls for information quality awareness as a way to bridge the gap between unstructured and structured data.

The significant amount of research in information quality in the last decade is generating greater awareness of the importance of quality of content, particularly in the pre-processing stage of data storage (Popovic et al., 2009; Stvilia et al., 2007). BI technology is changing and expanding, both in the scope of the data it collects and analyzes, and in the range of employees using it (Rodriguez et al., 2010). Today, virtually every software application feeds data into warehouses, permitting focus on the current picture, rather than on something that took place months or years ago (Popovic et al., 2009).

McGilvray (2008) and Olson (2003) note two major trends towards an environment in which information quality assurance is commonplace. The first trend is the increasing number of

legal and regulatory data quality constraints on businesses that requires information quality assurance aligns with stated goals and objectives (Caro et al., 2008; Lee et al., 2009; Olson, 2003). According to Hakim (2007a), there is a direct correlation between the recent regulatory requirements for information quality standards and the increase in the number of assurance processes for BI systems, the results of which are significantly improved competitive advantages. For example, the Sarbanes-Oxley Act of 2002 requires that businesses protect investors by improving the accuracy and reliability of information that is produced, or face large fines and corporate disgrace (Sarbanes & Oxley, 2002). These standards reduce risks of incompatibility, incompetence, and promise conformity for compliance, accuracy, and best practices (Seng & Chen, 2010).

Another example of regulatory requirements for assuring information quality is the Capital Requirement Directive, which requires that data and information is accurate, complete, and appropriate for the task at hand (Rodriguez et al., 2010). BI systems deploy policies and procedures to manage and measure risk, as well as to meet standards critical to legal and regulatory compliance (Caro et al., 2008).

The second trend is based on the need for businesses to increase competitive advantages to make data available for decision support through BI and data warehousing (McGilvray, 2008). The emergence of data warehouses, the advances in data mining, the increased capabilities of hardware and software, and the growth of the Internet present complex competitive information to decision makers; the overarching goal to improve competitive advantages through quality of content (Lee et al., 2009). The basis for competition has changed from tangible products to intangible information, and that information represents collective knowledge used to produce and deliver products and services to meet goals and objectives (Stvilia et al., 2007).

BI systems, then, are tools for maximizing competitive advantages by reducing redundancy, increasing efficiency, and ensuring better data integrity by streamlining information assurance processes (Popovic et al., 2009). Increasing competitive advantages provides decision makers with current information to make effective, rapid decisions to maximize profit and decrease overhead (Rodriguez et al., 2010).

The Information Quality Challenge

Businesses need information that can be trusted to be correct and current to meet goals and objectives (Olson, 2003). Negash (2008) notes that increasing pressure in businesses to justify ROI is met with the challenge of competitive intelligence: it is not the amount of information available to decision makers that ensure competitive advantages as much as it is the ability to differentiate useful data from misinformation.

Information quality problems are caused by human, process, and systems issues, and are not restricted to older systems (McGilvray, 2008). For example, normal business activities such as correction activities, duplication of work, and handling returns are indicative of data quality problems (Olson, 2003). BI systems create, update and delete data, and while IT teams are responsible for the quality of the systems that store and move the data, they are not completely responsible for content (McGilvray, 2008). In fact, according to McGilvray (2008), both IT and BI need clearly articulated requirements for the development of quality processes for effective data management.

Quality information is the most valuable asset of a firm; thus capitalizing on information quality assurance from BI systems enables decision makers to understand the capabilities available in a company to increase ROI by meeting goals and needs (Negash, 2008; Popovic et

al., 2009). According to English (2005) and McGilvray (2008), investing in information quality assurance is a means of showing benefits in returns on investment (ROI). However, a business must first identify and prioritize dimensions of information quality that align with corporate needs and goals to reach the level of data accuracy within the critical data warehouses of the corporation, and then, keep them at that level (Olson, 2003).

The Role of Information Quality Assurance in the Pre-Processing Stage of Data Storage within the Context of Business Intelligence

Information quality assurance in the pre-processing stage of data storage guards against erroneous data or information of marginal quality becoming factors in data mining and analysis procedures (Olson, 2003). According to Watson and Wixom (2007), too much information can be as ineffective as unstructured or poor-quality data. Focusing on key information and ensuring it is of useful data quality is the role of assurance plans for data storage (Lee et al., 2006). By collecting more, businesses end up with less; too many fields to check mean many fields to define and rules to implement (Cong et al., 2007). Assuring quality of content in data storage redesigns the processes of building data warehouse applications and automates the processes of measuring significant and structured information (Caro et al., 2008). Ensuring that necessary data quality guidance is developed and implemented within BI structures for consistency and accuracy is a major role of information quality assurance in data storage, and according to Stvilia et al. (2007) one that indicates the effectiveness of the decision-making process for knowledge workers in BI.

A solid, scalable information quality assurance plan for data storage is the essence of effective BI (Popovic et al., 2009). Assuring quality of content for data storage and management

maintains integrity for BI decision makers by ensuring that inconsistencies and discrepancies are non-existent (Davenport & Harris, 2007). In particular, information quality assurance for data storage ensures that results for the decision-making process are factual, present solutions for achieving or exceeding BI goals and objective, and provide clarity for BI knowledge workers (English, 2009).

Information quality. Information quality produces a clear competitive advantage for companies in both the public and private sectors (Lee et al., 2009). According to Lee et al. (2009), Knight and Burn (2005), and Keeton et al., (2009), the role of information quality is to:

- maximize objectivity and integrity of information;
- adopt a basic standard of quality and implement criteria into information quality practices; and,
- ensure compliance with legal and regulatory standards.

Information quality ensures objective, unbiased, and consistent data for substantively accurate identification of information sources (Knight & Burn, 2005). According to Davenport and Harris (2007), strategies for information quality policies and programs support business needs, goals, and objective by defining, measuring, analyzing, and improving the quality of data. Assurance for data storage also prioritizes requirements so that resulting systems produce information that better serves the needs of knowledge workers in the decision-making process (Hakim, 2007a).

Information quality assurance benefits. Assuring quality of content ensures that results from the data mining process are high-level quality and meet or exceed BI goals and objectives (Jafar, 2010). Thus information quality assurance aids BI knowledge workers in ensuring that quality is effectively managed in the data storage process (Olson, 2003).

Evaluating the quality of information before using it in the decision-making process ensures integrity for BI (Kahn et al., 2002). Collection of high quality data requires planning in the pre-processing stage of data storage to ensure accurate, consistent, reliable results for the decision-making process (Su et al., 2009). According to Su et al. (2009), poor quality data are caused by human, process, and system issues and it is often difficult to perceive the extent to which these problems affect the business systems. However, poor quality data cost as much or more to produce than meaningful quality data: an under- or over-designed solution for a problem results in a considerable expenditure of time and wasted money for decision makers (English, 2009). Thus the level of the importance of information quality assurance before reaching the warehouse is the degree to which information and data are viewed as trusted sources for achieving company goals (Watson & Wixom, 2007).

Sen and Sinha (2007) note that many businesses are ensuring quality within decision-making processes but still struggle with the critical task of assuring information quality for data before it is stored in warehouses. Lee et al. (2009) point out that while IT teams are responsible for the quality of the systems that store and move the data, they are not responsible for the content. Moreover, Piatetsky-Shapiro et al. (2009) state that both IT and BI systems need clearly articulated information quality processes in the pre-processing stage of data storage for successful data mining and management.

The impact of quality of content. Information quality assurance impacts business decision and actions by providing data in the form of intangible information (McGilvray, 2008). According to Jafar (2010), the importance of assuring high quality information in the pre-processing stage is often misunderstood, with the implicit assumption that the data mining process correctly represents the business when in fact the quality of the final result are only

representative of the level of quality in the early stages of data storage. That is, data are mined to discover knowledge about a business, and ultimately afford competitive advantages for BI systems (Panin, 2006; Seng & Chen, 2010). Importantly, results for decision makers are a reflection of the quality of the data captured during the pre-processing stage of data storage (Sen & Sinha, 2007). Data mining tools and procedures, such as decision trees or neural networks, are only effective when information quality assurance procedures are in place in the pre-processing stage (Zhao et al., 2006). According to Lupu et al. (2007), an understanding of the processes that are used to capture, generate, use, and store data are essential to information quality assurance in the pre-processing stage of data storage.

The Need to Define and Prioritize Key Information Quality Dimensions for Assuring Quality of Content in the Pre-Processing Stage of Data Storage

Information quality is not linear and thus, has many dimensions (English, 2009; McGilvray, 2008). Information quality assurance initiatives combine information from different sources in such a way that new and better uses are made with the resulting information (Olson, 2003). A clear understanding of the dimensions of information quality that most correctly align with BI systems goals and objective provides ways to effectively measure and manage the quality of data and information in the early stages of storage in data warehouses (English, 2009; Fisher et al., 2008; McGilvray, 2008).

Defining key information quality dimensions. Information quality is a multi-dimensional concept in which dimensions, or elements used in assessing subjective quality of content, are its measures (Olson, 2003). Once identified, dimensions are prioritized by BI systems by determining suitability for goals and objectives (English, 2005). According to Olson

(2003), the measurement of information quality effectiveness via the use of dimensions enables BI systems to focus on success from a decision-making perspective.

A dimension is a way of classifying and prioritizing BI information and needs (McGilvray, 2008). According to McGilvray, 2008, dimensions are used to define, measure, and manage the quality of data and content for data storage. BI systems measure dimensions of information quality to establish procedures and standards for meeting needs, goals and objectives (Rodriguez et al., 2010). Oversimplifying dimensions or poorly implemented processes do not align with true BI needs and triggers false results for the decision-making process (Su et al., 2009). Thus it is critical that BI systems focus on key dimensions that benefit the information quality assurance process by identifying and prioritizing those in alignment with BI goals and objectives (Cong et al, 2007; English, 2009; Watson & Wixom, 2007)

An information quality dimension provides a way to measure and manage the quality of data and information (McGilvray, 2008). According to McGilvray (2008), each dimension requires different tools, techniques, and processes to measure it. Differentiating the dimensions of quality helps match and business needs and goals (Caro et al., 2008; Stvilia et al., 2007). Dimensions that are the most meaningful to goals and objectives should be the focus; however, if a business is unsure where to begin information quality efforts, the dimensions of perception, relevance, and trust provide insight into issues by surveying knowledge workers and obtaining their point of view (Fisher et al., 2008; McGilvray, 2008). Those results articulate the BI problem and enable prioritization of the information quality efforts (Davenport & Harris, 2007).

According to McGilvray (2008) and Stvilia et al. (2007), in order to plan the ways to assure information quality, understanding common information quality dimensions is requisite. Businesses begin with a list of common dimensions, such as those listed below, and prioritize

according to goals and objectives (Hakim, 2007a; McGilvray, 2008; Negash, 2008; Olson, 2003). According to McGilvray (2008), dimensions used to assess information quality are grouped into four categories, as follows:

- Intrinsic Information Quality: Accuracy, Objectivity, Believability, Reputation
- Contextual Information Quality: Relevancy, Value-Added, Timeliness, Completeness, Amount of Information
- Representational Information Quality: Interpretability, Ease of Understanding, Concise Representation, Consistent Representation
- Accessibility Information Quality: Accessibility, Access Security

Identifying and prioritizing key dimensions. Information quality occurs along dimensions and is defined by the needs of the customer (Cong et al., 2007; McGilvray, 2008). Knowledge workers must understand the dimensions and the dynamic nature of information quality to effectively use identify and prioritize those useful for components of their decision-making processes (Negash, 2008). Understanding the key information quality dimensions is the first step to data quality assurance (Olson, 2003). Segregating data flaws by dimension allows companies to apply improvement techniques using information quality assurance tools to improve both the data and the processes that create and manipulate that information before it reaches the warehouse (English, 2009).

Information quality assurance begins with understanding the dimensions and moreover, identifying the key dimensions that align with BI goals (Cong et al., 2007). The dimensions are absolute, but the perception of the dimensions defines information quality (Hakim, 2007a; Keeton et al., 2009). The potential success of BI strategies for improving and ensuring

successful decision-making processes lies in identifying, defining, and prioritizing information quality dimensions (McGilvray, 2008).

Keeton et al. (2009) state that understanding the key information quality dimensions is the first step towards information quality assurance. Keeton et al. (2009) and Olson (2003) note that the ability to segregate unstructured data by dimension or classification allows analysts to apply improvement techniques using information quality tools to improve the quality of the information and the processes that create and manipulate that information.

Selecting the dimensions of information quality to be quantified within the context of user, environment, and task is critical to information quality assurance within the context of BI (McGilvray, 2008; Olson, 2003). Dimensions are assigned a value and ranking for analyzing priorities, addressing limitations with the context of the unique BI system, and for realistically determining achievable goals for competitive advantages (Davenport & Harris, 2007; Keeton et al., 2009). Fisher et al. (2008) note that by assigning a dimension value and rank, a business can better manage information quality assurance to ensure that end-user needs are met in the pre-processing stage of data storage.

Contextualizing key dimensions for assurance in the pre-processing stage.

Information is critical for successful decision-making, and is effective when the quality of content is assured (English, 2005). The concept seems obvious, but the definition of information quality varies significantly depending on the business, the goal, or the objective (Olson, 2009). It is important that the information quality dimensions that best address BI needs and goals be chosen for successful data management; the scope of the effort required for a particular project is better assessed this way (Popovic et al., 2009; Rodriguez et al., 2010). The ultimate objective of assuring quality of content is establishing a data warehouse that contains relevant and accurate

information of a business environment (Sen & Sinha, 2007). Assuring quality of content in the pre-processing stage of data storage involves, at a minimum, data integrity and accuracy (Pipino et al., 2002). Identifying and prioritizing key dimensions in order to evaluate information quality and assure quality of content is effective when prescribed for each unique business environment dimension (Stvilia et al., 2007). Thus data quality is assured when measured along several dimensions and contextualized by unique BI goals and objectives (Popovic et al., 2009; Rodriguez et al., 2010).

Conclusions

The purpose of this study is to address key dimensions of information quality, as identified in selected literature, necessary for data quality assurance within the context of BI (Hakim, 2007a; Jafar, 2010). The goal is to produce a framework for identifying and prioritizing key dimensions unique to each BI system's goals and objectives to ensure integrity and consistency of information for assurance in the pre-processing stage of data storage.

Kahn et al. (2002) provide a set of general guidelines for structuring a comprehensive information quality assurance framework, which includes the following steps:

1. Develop BI goals and objectives;
2. Identify and prioritize dimensions of information quality that align with goals and objectives;
3. Implement and maintain an assurance plan for all information quality processes and procedures;
4. Review and approval of documentation by appropriate knowledge workers; and,
5. Define and communicate each key dimension for data quality management to stakeholders.

This study focuses on Step 2, identifying and prioritizing key dimensions of information quality assurance for data storage and management, for use within the context of each uniquely distinct BI system.

Summary of 10 Widely Accepted Key Dimensions for Information Quality Assurance

The fundamental key dimensions of information quality are those that most closely align with unique BI goals and objectives (Kahn et al., 2002). In fact, Davenport and Harris (2007)

and McGilvray (2008) note that established companies build on existing strengths by transforming dimensions of information quality into strategies after identifying those key to goal alignment. Thus identifying key dimensions of information quality and continuously prioritizing them based on current BI goals and objectives significantly contributes to an effective decision-making process and increases competitive advantages for BI (Negash, 2008).

Panin (2006) and Piatetsky-Shapiro et al. (2009) note that investing in identifying and prioritizing dimensions of information quality distinguishes effective BI systems from ineffective ones. The value of information quality assurance, then, is not in the level of quality of content; rather, the value is in how it affects the decision-making and competitive advantage processes for BI (Kriegel et al., 2007).

Table 4 presents a summary of the 10 most widely accepted key information quality dimensions for consideration at the pre-processing stage, to meet BI needs and goals. There are over 30 widely accepted dimensions of information quality; however, most experts in the field agree that while the process of prioritizing dimensions is unique to a specific BI system set of goals and objectives, those listed in Table 4 are key (Keeton et al., 2009; Knight & Burn, 2005).

Table 4
Summary of Key Dimensions and Definitions of Information Quality

Dimension	Definitions
Accessibility	The extent to which data is retrieved as needed
Accuracy	A measure of correctness of the content of the data
Completeness	The extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Free of Error	The extent to which data is correct and reliable
Interpretability	The extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial
Relevancy	The extent to which data is applicable and helpful for the task at hand
Reliability	The extent to which data is regarded as true and credible
Timeliness	The extent to which the data is sufficiently up-to-date for the task at hand
Value Added	The extent to which data is beneficial and provides advantages from its use

Two Selected Processes for Aligning Key Dimensions with Business Goals and Objectives

Prioritizing the key dimensions, then, creates niches for BI systems based on timeliness and opportunity; being the first to determine and respond to market changes and needs increases competitive advantages (English, 2009). Selected literature indicates that aligning key dimensions with goals and objectives requires forethought by knowledge workers and decision makers (Cong et al., 2007; English, 2005; English, 2009; McGilvray, 2008; Olson, 2003; Stvilia et al., 2007). Awareness of fundamental key dimensions provides a logical structure for identifying and prioritizing the components that contribute to assuring information quality at the pre-processing stage for specific goals (Stvilia et al., 2007). Moreover, according to McGilvray (2008) it provides an understanding of a complex environment in which information quality problems are created, and enables organized thinking for BI systems to plan and create quality data and implement improvements as needed. Regardless of how data are structured, it is important that businesses are consistently clear on what dimensions are, and what dimensions are not, when defining BI needs during the assessment stage of the information quality assurance improvement cycle (Caro et al., 2008; Davenport & Harris, 2007; English, 2009).

McGilvray (2008) describes a process for identifying key dimensions first, and then prioritizing those in alignment with specific goals and objectives. Furthermore, McGilvray (2008) states that once key dimensions are in place the process for the continuous assessment, maintenance, and improvement of information is critical for producing assuring information quality.

The process consists of a set of concrete instructions for planning and implementing information and data quality improvement projects (McGilvray, 2008). According to McGilvray (2008), each step contains general principles, directions, advice, and examples for assessment,

awareness, and action. The first step is to define business need and identify and prioritize dimension to focus on what is relevant and critical to meet objectives (McGilvray, 2008).

Stvilia et al. (2007) present another process, and propose that key dimensions must align with and be connected to the BI system to assure information quality. Stvilia et al. (2007) claim that incomplete, ambiguous, inaccurate, inconsistent, or redundant data that is not corrected in the pre-processing stage of data storage is a result of not identifying and prioritizing key dimensions.

The central part of Stvilia et al.'s (2007) framework is a taxonomy of information quality dimensions. The taxonomy consists of 22 information quality dimensions organized into three categories based on information quality variance: intrinsic information quality (cultural norms and conventions); relational, or contextual, information quality (immediate context or object of information quality assessment); and, reputational information quality (cultural or community related structure). In addition to a taxonomy of information quality dimensions, the framework consists of a set of 41 general metric functions implemented as Java codes used to develop context-specific information quality metrics.

The framework serves as a valuable knowledge resource and guide for assuring information quality by establishing connections among information quality dimensions. Moreover, the framework provides a predictive mechanism to identify information quality problems early on (Stvilia et al., 2007). According to Stvilia et al. (2007), the first step is to identify the business goals and objectives. Next, a set of relevant information quality dimensions is selected from the framework that aligns with goals. Finally, the information quality dimensions are aggregated into an index for each information quality dimension for assuring high-level quality in the pre-processing stage of data storage.

References

Andersson, D., Fries, H., & Johansson, P. (2008). *Business intelligence: The impact on decision support and decision making processes* (Unpublished master's thesis). Jonkoping University, Norway. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:hj:diva-1159

Arkady, M. (2007). *Data quality assessment*. Bradley Beach, NJ: Technics Publications, LLC.

Bell, C., & Smith, T. (2007). *Critical evaluation of information sources*. Retrieved from http://libweb.uoregon.edu/guides/findarticles/credibility.html

Berkley, C., Bowers, S., Jones, M.B., Madin, J.S., & Schlidhauer, M. (2009). Improving data discovery for metadata repositories through semantic search. *Complex, Intelligent and Software Intensive Systems, 16*(19), 1152-1159. http://doi.ieeecomputersociety.org/10.1109/CISIS.2009.122

Busch, C., De Maret, P., Flynn, T., Kellum, R., Le, S., Meyers, B., Saunders, M., & White, R., (2005). *Content analysis*. Retrieved from http://writing.colostate.edu/guides/research/content

Caro, A., Calero, C., Caballero, I., & Piattini, M. (2008) A proposal for a set of attributes relevant for web portal data quality, *Software Quality Journal, 16*(4), 513-542. doi:10.1007/s11219-008-9046-7

CiteSeer. *About CiteSeer.* Retrieved from CiteSeer Web site: http://citeseer.ist.psu.edu/about/site;jsessionid=6D156B1258F653C3D4CC178ABC694380

Cong, G., Fan, W., Geerts, F., Jia, X., & Shuai, M. (2007). Improving data quality: Consistency and accuracy. *Proceedings of the 33rd International Conference on Very Large Databases*

(VLDB), Vienna, Austria, 2007, 315-326. Retrieved from http://www.vldb.org/conf/2007/papers/research/p315-cong.pdf

Creswell, J.W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Davenport, T.H., & Harris, J.G. (2007). The architecture of business intelligence. In *Competing on analytics: The new science of winning.* (chapter 8). Boston, MA: Harvard Business School Press. Retrieved from http://www.accenture.com/NR/rdonlyres/15DCFF6A-4DE0-44D8-B778-630BE3A677A2/0/ArchBIAIMS.pdf

English, L. (2005). Information quality for business intelligence and data mining: Assuring quality for strategic information uses. [White paper]. Retrieved from http://infoimpact.com/articles/IQBI&DataMining.pdf

English, L. (2009). *Information quality applied: Best practices for improving business information, processes and systems*. New York, NY: John Wiley & Sons, Inc.

Fayed, U.M., & Uthurusamy, R. (2002). Evolving data mining into solutions for insights. *Communications of the ACM, 45*(8), 28-21. Retrieved from http://sce.uhcl.edu/boetticher/ML_DataMining/p28-fayyad.pdf

Fink, A. (2010). *Conducting research literature reviews* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Fisher, C., Lauria, E., Chengalur-Smith, S., & Wang, R. (2008). Introduction to information quality (4th ed.). Cambridge,MA: MIT Press.

Forcada, N., Casals, M., Fuertes, A., Gangolells, M., & Roca, X. (2010). A web-based system for sharing and disseminating research results: The underground construction case study. *Automation in Construction*, *19*(4), 458-474. doi:10.1016/i.autocon.209.12.018

Gallo, J. (2010, September 21). A business context for agile business intelligence [Web log comment]. Retrieved from http://www.b-eye-network.com/view/14384

Geist, M. (2008). *Enhancing home computer user information security: Factors to consider in the design of anti-phishing applications*. Retrieved from http://aim.uoregon.edu/research/pdfs/2008-geist.pdf

Haag, S., Cummings, M., McCubbrey, D., Pinsonneault, A., & Donovan, R. (2006). *Management information systems for the information age* (3rd ed.). Whitby, Ontario, Canada: McGraw-Hill Ryerson.

Hakim, L. (2007a). *Information quality management: Theory and applications*. Hershey, PA: Idea Group Publishing.

Hakim, L. (2007b). *Challenges of managing information quality in service organizations.* Hershey, PA: Idea Group Publishing.

Halonen, R., & Thomander, H. (2008). Measuring knowledge transfer success by D&M. *Sprouts: Working Papers on Information Systems*, *8*(41). Retrieved from http://sprouts.aisnet.org/8-41

Hsieh, H.F., & Shannon, S.E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277-1288. doi: 10.1177/1049732305276687

IBM (2009). Business intelligence for business users: How IT can make business intelligence easy for everyone. [White paper]. Retrieved from http://public.dhe.ibm.com/software/data/sw-library/cognos/pdfs/whitepapers/wp_c8v4_bi_for_bus_users.pdf

IBM. (2010). The new promise of business intelligence. [White paper]. Retrieved from http://www.itbusinessedge.com/offer.aspx?o=00630554BIwp&pc

=defoffsliverbi

Jafar, M.J. (2010). A tools-based approach to teaching data mining. *Journal of Information Technology Education: Innovations in Practice, 9*, 2-24. Retrieved from http://jite.org/documents/Vol9/JITEv9IIPp001-024Jafar740.pdf

Kahn, B.K., Strong, D.M., & Wang, R.Y. (2002). Information quality benchmarks: Product and service performance, *Communications of the ACM*, 45(4), 184-192. doi: 10.1145/505999.56007

Kanal, L.N. (2009). Problem-solving models and search strategies for pattern recognition. *Pattern Analysis and Machine Intelligence, 1*(2), 193-201. doi:10.1109/TPAMI.1979.4766905

Keeton, K., Mehra, P., & Wilkes, J. (2009). Do you know your IQ: A research agenda for information quality in systems. *ACM Sigmetrics Performance Evaluation Review, 37*(3), 1-6. Retrieved from http://www.sigmetrics.org/sigmetrics/workshops/papers_hotmetrics/session1_4.pdf

Klein, B.D. (2002). When do users detect information quality problems on the World Wide Web? *American Conference in Information Systems, 41*(4), 9-18. Retrieved from http://sighci.org/amcis02/RIP/Klein.pdf

Knight, S., & Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. *Informing Science Journal,* 8(1), 159-172. Retrieved from http://inform.nu/Articles/Vol8/v8p159-172Knig.pdf

Kriegel, H. P., Borgwardt, K.M., Kroger, P., Pryakhin, A., Schubert, M., & Zimek, A., (2007). Future trends in data mining, *Data Mining and Knowledge Discovery, 15*(1), 87-97. doi:10.1007/s10618-007-0067-9

Lamont, J. (2010). Competitive intelligence: Capturing a wider view. *KM World, 19*(10), 14-15. Retrieved from http://www.kmworld.com/Articles/PrintArticle.aspx?ArticleID=70849

Lee, Y.W., Pipino, L.L., Funk, J.D., & Wang, R.Y. (2009). *Journey to Data Quality*. Cambridge, MA: MIT Press.

Leedy, P.D., & Ormrod, J.E. (2010). *Practical research: Planning and design* (9th ed.). Upper Saddle River, NJ: Allyn & Bacon.

Lefebvre, R. C. (2007). The new technology: The consumer as participant rather than target audience. *SMq, 13*(3), 31-42. Retrieved from http://www.scribd.com/doc/38464538/SMQ-The-Consumer-as-Participant-2007

Levy, Y., & Ellis, T.J. (2006). Towards a framework of literature review process in support of information systems research. *Proceedings of the 2006 Informing Science and IT Education Joint Conference*. Retrieved from http://www.informingscience.org/proceedings/InSITE2006/ProcLevy180.pdf

Luckey, T.S. (2009). *Key stages of disaster recovery planning for time-critical business information technology systems*. Retrieved from http://aim.uoregon.edu/research/pdfs/2009-luckey.pdf

Lupu, A.R., Razvan, B., Sabau, G., & Muntean, M. (2007). Influence factors of business intelligence in the context of ERP projects, *International Journal of Education and Information Technologies, 2*(1), 90-94. Retrieved from http://www.naun.org/journals/educationinformation/eit-15.pdf

McGarry, K., (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review, 1*(1), 1-24. doi:10.1017/S0269888905000408

McGilvray, D.M. (2008). *Executing data quality projects: Ten steps to quality data and trusted information.* Burlington, MA: Morgan Kaufmann Publishers.

Negash, S. (2008). Handbook on decision support systems 1: Business intelligence. In *International handbooks on information systems*. (chapter 45). Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-540-48713-5

Obenzinger, H. (2005). *What can a literature review do for me?* Retrieved from Stanford University: http//ual.stanford.edu/pdf/uar_literaturereviewhandout.pdf

Olson, J.E. (2003). *Data quality: The accuracy dimension*. San Francisco, CA: Morgan Kaufmann Publishers.

Olson, J.E. (2009). *Database archiving: How to keep lot of data for a very long time*. Burlington, MA: Morgan Kaufmann Publishers.

Ormondroyd, J., Engle, M., & Cosgrave, T., (2009). Critically analyzing information sources. Retrieved from Cornell University, Olin & Uris Libraries Web site: http://olinuris.library.cornell.edu/ref/research/skill26.htm

Panin, Z. (2006). Business intelligence in support of business strategy. *Proceedings of the 7th WSEAS International Conference on Mathematics & Computers in Business & Economics, Croatia, 6,* 19-23. Retrieved from http://www.wseas.us/e-library/conferences/2006cavtat/papers/528-109.pdf

Parameter, D. (2010). *Key performance indicators: Developing, implementing, and using winning KPIs.* Hoboken, NJ: John Wiley & Sons, Inc.

Piatetsky-Shapiro, G., Djeraba, C., Getoor, L., Grossman, R., Feldman, R., & Zaki, M. (2009). What are the grand challenges for data mining? *SIGKDD Explorations, 8*(2), 70-77. doi:10.1145/1233321.1233330

Pipino, L.L., Lee, Y.W., & Wang, R.Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211-218. doi: 10.1145/505248.506010

Popovic, A., Coelho, P.S., & Jaklic, J. (2009). The impact of business intelligence system maturity on information quality. *Information Research, 14*(4), 1-14. Retrieved from http://informationr.net/ir/14-4/paper417.html

Power, D.J. (2004). A brief history of decision support systems. *DSSRResources.com, 4*(1). Retrieved from http://dssresources.com/history/dsshistory.html

Redman, T., & Daugherty, M. (2001). *Data quality: The field guide.* Burlington, MA: Elsevier, Inc.

Rodriguez, C., Daniel, F., Casati, F., & Cappiello, C. (2010). Toward uncertain business intelligence: The case of key indicators. *IEEE Internet Computing, 14*(4), 32-40. http://doi.ieeecomputersociety.org/10.1109/MIC.2010.59

Sarbanes, P., & Oxley, M. (2002). *A guide to the Sarbanes-Oxley act*. Retrieved from http://www.soxlaw.com/

Sen, A., & Sinha, A.P. (2007). Toward developing data warehousing process standards: An ontology-based review of existing methodologies. *IEEE Transactions on Systems, Man and Cybernetics: Part C, Applications and Reviews, 37*(1), 17-31. http://dx.doi.org/10.1109/TSMCC.2006.886966

Seng, J.L., & Chen, T.C. (2010). An analytic approach to select data mining for business decisions. *Expert Systems with Applications, 37*(12), 8042-8057. http://dx.doi.org/10.1016/j.eswa.2010.05.083

Stacks, G., & Karper, E. (2008). *Annotated bibliographies*. Retrieved from The Owl at Purdue: http://owl.english.purdue.edu/owl/resource/614/01/

Stvilia, B., Gasser, L., Twidale, M.B., & Smith, L.C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*(12), 1720-1733. doi:10.1002/asi.20652

Su, Y., Peng, J., & Jin, Z. (2009). Modeling information quality risk for data mining in data warehouses. *Human & Ecological Risk Assessment, 15*(2), 332-350. doi: 10.1109/ICISE.2009.755

Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. *Proceedings from Eighth IEEE International Conference on ICDM*. doi:10.1109/ICDM.2008.71

Thiesse, F., Floerkemeir, C., Harrison, M., Michahelles, F., & Roduner, C. (2010). Technology, standards, and real-world deployments of the EPC network. *IEEE Internet Computing, 2*(9), 36-43. Retrieved from http://www.im.ethz.ch/publications/tech_standards_realworld_epc.pdf

Ulrichsweb™ (n.d.). *Ulrich's Periodicals Directory.* Retrieved from Ulrichsweb™ Web site: http://www.ulrichsweb.com.libproxy.uoregon.edu/ulrichsweb/Search/fullCitation.asp?navPage=1&tab=1&serial_uid=196771&vendor=SFX&

University of Colorado at Boulder (n.d.). *How do I…?* Retrieved from http://ucblibraries.colorado.edu/how/evaluate.htm

University of North Carolina (n.d.). *Writing center: Literature reviews.* Retrieved from University of North Carolina at Chapel Hill, Writing Center Web site: http://www.unc.edu/depts/wcweb/handouts/literature_review.html

Wang, S., & Wang, H. (2007). Mining data quality in completeness. *Proceedings of the 2007 International Conference on Information Quality (MIT IQ Conference Center),* 1-6. doi:10.1.1.90.4260

Watson, H.J., & Wixom, B.H. (2007). The current state of business intelligence. *Computer, 40*(9), 96-99. http://dx.doi.org/10.1109/MC.2007.331

Web4All. (2010). Business intelligence: From data collection to data mining and analysis. *Proceedings from the 7th International Cross-Disciplinary Conference on Web Accessibility.* Retrieved from http://wps.prenhall.com/wps/media/objects/2519/2580469/addit_chmatl/TURBMC04_01 31854615App.pdf

Wixom, B.H., & Watson, H.J. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly, 25*(1), 17-41. Retrieved from http://hinf551edwcase.wikispaces.com/file/view/3250957.pdf

Zhao, Y., Chen, Y., & Yao, Y. (2006). *User-centered interactive data mining*. IEEE International Conference on Cognitive Informatics 2006, 457-466. http://dx.doi.org/10.1109/COGINF.2006.365532

Appendix A – Search Record

Detailed Record of Searches

Search Engine / Database	Search Terms	Results: #	Eligible Titles Found	Comments
ACM Digital Library	Information + quality	117,347	12	This library is an excellent starting resource to search for the focused topic.
	Data + mining	57,595	14	
	Business + intelligence	26,091	9	
	Knowledge + discovery	44,302	8	
	Data + analytics	2,059	3	
	Data + warehouse	8,119	3	
	Competitive + advantage	4,221	6	
	Information + quality + mining + business + intelligence	3,159	9	
	Information + quality + assurance	9,084	11	
Academic Search Premier Index – EBSCO HOST (UO Libraries)	Information + quality	52,404	13	This index is a good resource for a starting point and is worth continued exploration with the focused topic.
	Data + mining	12,150	7	
	Business + intelligence	5,733	8	
	Knowledge + discovery	4,353	4	
	Data + analytics	934	3	
	Data + warehouse	1,661	9	
	Competitive + advantage	799	4	
	Information + quality + mining + business + intelligence	1,441	11	
	Information + quality	1,766	9	

	+ assurance			
Search Engine / Database	Search Terms	Results: #	Eligible Titles Found	Comments
CiteSeer[x] Search Index	Information + quality	218,812	12	This search engine is a very good resource the topic.
	Data + mining	42,737	5	
	Business + intelligence	22,023	11	
	Knowledge + discovery	35,208	11	
	Data + analytics	83,255	15	
	Data + warehouse	8,021	3	
	Competitive + advantage	8,221	9	
	Information + quality + data mining + business + intelligence	91,337	7	
	Information + quality + assurance	247,558	15	
ERIC	Information + quality	3,021	1	This is not a very helpful resource for the focused topic.
	Data + mining	124	0	
	Business + intelligence	89	0	
	Knowledge + discovery	186	0	
	Data + analytics	13	0	
	Data + warehouse	33	0	
	Competitive + advantage	19	1	
	Information + quality + mining + business + intelligence	0	0	
	Information + quality + assurance	134	0	
Google Scholar Advanced	Information + quality	473,000	6	Possibly a good resource; worth continuing effort with this search engine especially with more defined parameters.
	Data + mining	567,000	5	
	Business +	312,000	7	

	intelligence			
Search Engine / Database	Search Terms	Results: #	Eligible Titles Found	Comments
	Knowledge + discovery	689,000	9	
	Knowledge + discovery	142	2	
	Data + analytics	77	1	
	Data + warehouse	45	1	
	Competitive + advantage	291	2	
	Information + quality + mining + business + intelligence	105	1	
	Information + quality + assurance	180	2	
IEEE Computer Science Digital Library	Information + quality	106	6	This is a good resource for academic articles
	Data + mining	202	3	
	Business + intelligence	68	1	
	Knowledge + discovery	142	5	
	Data + analytics	77	1	
	Data + warehouse	45	1	
	Competitive + advantage	187	3	
	Information + quality + mining + business + intelligence	105	2	
	Information + quality + assurance	180	3	
	Information + quality + assurance	33,886	9	
Project Muse (UO Libraries)	Information + quality	27,082	8	This is a good resource for the topic. Worth further exploration, especially with refined parameters.
	Data + mining	1,612	6	
	Business + intelligence	4,891	7	
	Knowledge discovery	12,308	7	

Search Engine / Database	Search Terms	Results: #	Eligible Titles Found	Comments
	Data + analytics	178	3	
	Data + warehouse	287	3	
	Competitive + advantage	344	1	
	Information + quality + mining + business + intelligence	175	5	
	Information + quality + assurance	1,437	4	
Sage Journals Online	Information + quality	147	4	This search engine is not a productive website for articles related to this topic.
	Data + mining + techniques	1,934	2	
	Business + intelligence	4,472	1	
	Knowledge + discovery	9,079	2	
	Data + analytics	385	3	
	Data + warehouse	719	3	
	Competitive + advantage	211	1	
	Information + quality + mining + business + intelligence	17	2	
	Information + quality + assurance	3,467	0	
Web of Science (UO Libraries)	Information + quality	72,744	7	This index is a good resource for academic articles.
	Data + mining	16,470	11	
	Business + intelligence	903	9	
	Knowledge + discovery	8,644	3	
	Data + analytics	303	6	
	Data + warehouse	1,255	5	
	Competitive + advantage	988	2	
	Information + quality + assurance	2,613	4	

Appendix B – References Selected for Coding

Andersson, D., Fries, H., & Johansson, P. (2008). *Business intelligence: The impact on decision support and decision making processes* (Unpublished master's thesis). Jonkoping University, Norway. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:hj:diva-1159

Caro, A., Calero, C., Caballero, I., & Piattini, M. (2008) A proposal for a set of attributes relevant for web portal data quality, *Software Quality Journal, 16*(4), 513-542. doi:10.1007/s11219-008-9046-7

Cong, G., Fan, W., Geerts, F., Jia, X., & Shuai, M. (2007). Improving data quality: Consistency and accuracy. *Proceedings of the 33rd International Conference on Very Large Databases (VLDB), Vienna, Austria, 2007,* 315-326. Retrieved from http://www.vldb.org/conf/2007/papers/research/p315-cong.pdf

Davenport, T.H., & Harris, J.G. (2007). The architecture of business intelligence. In *Competing on analytics: The new science of winning.* (chapter 8). Boston, MA: Harvard Business School Press. Retrieved from http://www.accenture.com/NR/rdonlyres/15DCFF6A-4DE0-44D8-B778-630BE3A677A2/0/ArchBIAIMS.pdf

English, L. (2005). Information quality for business intelligence and data mining: Assuring quality for strategic information uses. [White paper]. Retrieved from http://infoimpact.com/articles/IQBI&DataMining.pdf

English, L. (2009). *Information quality applied: Best practices for improving business information, processes and systems*. New York, NY: John Wiley & Sons, Inc.

Fisher, C., Lauria, E., Chengalur-Smith, S., & Wang, R. (2008). Introduction to information quality (4th ed.). Cambridge,MA: MIT Press.

Hakim, L. (2007a). *Information quality management: Theory and applications*. Hershey, PA: Idea Group Publishing.

Jafar, M.J. (2010). A tools-based approach to teaching data mining. *Journal of Information Technology Education: Innovations in Practice, 9*, 2-24. Retrieved from http://jite.org/documents/Vol9/JITEv9IIPp001-024Jafar740.pdf

Kahn, B.K., Strong, D.M., & Wang, R.Y. (2002). Information quality benchmarks: Product and service performance, *Communications of the ACM*, 45(4), 184-192. doi: 10.1145/505999.56007

Keeton, K., Mehra, P., & Wilkes, J. (2009). Do you know your IQ: A research agenda for information quality in systems. *ACM Sigmetrics Performance Evaluation Review, 37*(3), 1-6. Retrieved from http://www.sigmetrics.org/sigmetrics/workshops/papers_hotmetrics/session1_4.pdf

Klein, B.D. (2002). When do users detect information quality problems on the World Wide Web? *American Conference in Information Systems, 41*(4), 9-18. Retrieved from http://sighci.org/amcis02/RIP/Klein.pdf

Knight, S., & Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. *Informing Science Journal,* 8(1), 159-172. Retrieved from http://inform.nu/Articles/Vol8/v8p159-172Knig.pdf

Kriegel, H. P., Borgwardt, K.M., Kroger, P., Pryakhin, A., Schubert, M., & Zimek, A., (2007). Future trends in data mining, *Data Mining and Knowledge Discovery, 15*(1), 87-97. doi:10.1007/s10618-007-0067-9

Lee, Y.W., Pipino, L.L., Funk, J.D., & Wang, R.Y. (2009). *Journey to Data Quality*. Cambridge, MA: MIT Press.

Lefebvre, R. C. (2007). The new technology: The consumer as participant rather than target audience. *SMq, 13*(3), 31-42. Retrieved from http://www.scribd.com/doc/38464538/SMQ-The-Consumer-as-Participant-2007

Lupu, A.R., Razvan, B., Sabau, G., & Muntean, M. (2007). Influence factors of business intelligence in the context of ERP projects, *International Journal of Education and Information Technologies, 2*(1), 90-94. Retrieved from http://www.naun.org/journals/educationinformation/eit-15.pdf

McGilvray, D.M. (2008). *Executing data quality projects: Ten steps to quality data and trusted information.* Burlington, MA: Morgan Kaufmann Publishers.

Negash, S. (2008). Handbook on decision support systems 1: Business intelligence. In *International handbooks on information systems*. (chapter 45). Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-540-48713-5

Olson, J.E. (2003). *Data quality: The accuracy dimension*. San Francisco, CA: Morgan Kaufmann Publishers.

Olson, J.E. (2009). *Database archiving: How to keep lot of data for a very long time*. Burlington, MA: Morgan Kaufmann Publishers.

Panin, Z. (2006). Business intelligence in support of business strategy. *Proceedings of the 7th WSEAS International Conference on Mathematics & Computers in Business & Economics, Croatia, 6,* 19-23. Retrieved from http://www.wseas.us/e-library/conferences/2006cavtat/papers/528-109.pdf

Piatetsky-Shapiro, G., Djeraba, C., Getoor, L., Grossman, R., Feldman, R., & Zaki, M. (2009). What are the grand challenges for data mining? *SIGKDD Explorations, 8*(2), 70-77. doi:10.1145/1233321.1233330

Pipino, L.L., Lee, Y.W., & Wang, R.Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211-218. doi: 10.1145/505248.506010

Popovic, A., Coelho, P.S., & Jaklic, J. (2009). The impact of business intelligence system maturity on information quality. *Information Research, 14*(4), 1-14. Retrieved from http://informationr.net/ir/14-4/paper417.html

Rodriguez, C., Daniel, F., Casati, F., & Cappiello, C. (2010). Toward uncertain business intelligence: The case of key indicators. *IEEE Internet Computing, 14*(4), 32-40. http://doi.ieeecomputersociety.org/10.1109/MIC.2010.59

Sen, A., & Sinha, A.P. (2007). Toward developing data warehousing process standards: An ontology-based review of existing methodologies. *IEEE Transactions on Systems, Man and Cybernetics: Part C, Applications and Reviews, 37*(1), 17-31. http://dx.doi.org/10.1109/TSMCC.2006.886966

Seng, J.L., & Chen, T.C. (2010). An analytic approach to select data mining for business decisions. *Expert Systems with Applications, 37*(12), 8042-8057. http://dx.doi.org/10.1016/j.eswa.2010.05.083

Stvilia, B., Gasser, L., Twidale, M.B., & Smith, L.C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*(12), 1720-1733. doi:10.1002/asi.20652

Su, Y., Peng, J., & Jin, Z. (2009). Modeling information quality risk for data mining in data warehouses. *Human & Ecological Risk Assessment, 15*(2), 332-350. doi: 10.1109/ICISE.2009.755

Watson, H.J., & Wixom, B.H. (2007). The current state of business intelligence. *Computer, 40*(9), 96-99. http://dx.doi.org/10.1109/MC.2007.331

Zhao, Y., Chen, Y., & Yao, Y. (2006). *User-centered interactive data mining*. IEEE International Conference on Cognitive Informatics 2006, 457-466. http://dx.doi.org/10.1109/COGINF.2006.365532