Presented to the Interdisciplinary Studies Program:

UNIVERSITY OF OREGON
APPLIED INFORMATION MANAGEMENT

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

Management of Data Quality in Information Systems

CAPSTONE 1 Bibliography

Bill Worth
Engineer in Charge
Applied Materials Field Operations

University of Oregon
Applied Information
Management
Program

December 2013

Academic Extension
1277 University of Oregon
Eugene, OR 97403-1277
(800) 824-2714

Approved by

Dr. Linda F. Ettinger, Capstone Instructor

Dr. Kara McFall, Capstone Instructor

Management of Data Quality in Information Systems

Bill Worth

Applied Materials

Table of Contents

List of Tables and Figures

Introduction

Problem

Parssian, Sarkar, and Jacob (2004) suggest that information systems are critical to supporting organizational strategic, tactical and operational decisions. Popovic, Coelho and Jaklic (2009) identify that the contributions of information technology (IT) cannot be immediately seen, though IT solutions offer benefits of time saving as well as improved information quality. While providing a consolidated source for data reporting is important, the information provided is only as good as the quality of the data within the system. The importance of managing the quality of data is widely documented in the information economy (Parssian, Sarkar, & Jacob, 2004). The *quality* of products and services is the single most important factor in the long-term success of a business and this quality is dependent on the information that an organization's employees use to create the products and services (Nelson, Todd & Wixom, 2005).

Data dimensions. *Information quality* is a vaguely defined concept according to Popovic, Coelho and Jaklic (2009), though it is closely identified as the content and accessibility of data. The dimensions of data quality can be described as being complete, accurate, consistent and current, (also known as timeliness) (Batini, Cappiello, Francalanci, & Maurino, 2009). Completeness can be defined as availability of all relevant data to satisfy the user (Parssian, Sarkar & Jacob, 2004). Accuracy can be defined as the extent to which the data conforms to the real world (Parssian, Sarkar & Jacob, 2004). Consistency can be defined as data being consistent across all research contributions (Cappiello, Francalanci & Pernici, 2003) such as multiple data sources containing similar information. Information such as business names can appear to be multiple entries in different data sources and can be difficult to identify correctly and efficiently by users when performing a search (Madnick, Wang, & Xian, 2003). Currency can be defined as the time point in which the data was stored and how relevant the information is at the time of the access (Cappiello, Francalanci & Pernici, 2003). Caballero, Caro, Calero, and Piattini (2008) suggest *fitness for use* is a more appropriate definition of information quality; as only the persons performing the search have the ability to judge if the results match the query (Guerra-Garcia, Caballero, Piattini & Springer, 2013).

Poor data quality. The concept of poor data quality or incomplete data is known as *dirty data*. Dirty data can be defined as missing data, wrong data, a non-standard representation of the same data, or unusable data typically found in legacy systems (Kim, Choi, Hong, & Lee, 2003). Ryu, Park, and Park (2006) describes data quality as being one of the most powerful competition advantages for companies that rely on information to make decisions. So, it is no surprise that having poor data in the information system can cost the organization a great deal of time and money (Heinrich & Klier, 2011).Low quality of data can have several negative effects on business users through the loss of customer satisfaction, increased running costs, inefficient decision making processes and lower performance output (Ryu, Park, & Park, 2006). According to an international survey, 75 percent of the respondents made wrong decisions due to incorrect or outdated data; survey respondents also claim their staff spent up to 30 percent of their working time on checking the quality of data provided (Heinrich & Klier, 2011). Cappiello, Francalanci and Pernici, (2003) project that costs associated with poor data quality have been estimated to be as much as 8% to 12% of the revenue and 40%-60% of a service organization's expenses. Gartner Inc. forecasts that poor data quality affects one quarter of all Fortune 1000 companies at an estimated cost of over $600 billion each year (Jiang, Sarkar, De & Dey, 2007). Batini, Cappiello, Francalanci and Maurino (2009) claim that data quality is recognized as a performance factor of operating processes of decision making activities.

Parssian, Sarkar and Jacob (2004) cite that the management of data quality and the quality of associated data management processes are critical issues for organizations. For example, it is difficult for the search algorithms to produce accurate results if the data being accessed is of poor quality (Kim, Choi, Hong, Kim, & Lee, 2003). Cappiello, Francalanci and Pernici (2003) find the quality of information depends on the quality of the data that can be extracted from the organization’s data sources. The importance of data quality is a widespread concern in all economic sectors, and relevance is constantly growing (Cappiello, Francalanci & Pernici, 2003). The relationship between data accuracy and the resulting information accuracy is one of the greatest issues companies are trying to solve within their organization (Gelman, 2010). Dirty data can be caused by poor data entered into the source database by users, inadequate data management procedures, software errors and issues with the system's ability to read the information correctly, such as accessing information from legacy systems (Cappiello, Francalanci & Pernici, 2003).

Purpose

Information is not an isolated resource that can be departmentalized; information flows throughout the company, and information quality is an organizational issue (Cabellero, Caro, Calero, & Piattini, 2008). Figure 1 provides an example of how data flows through an organization.

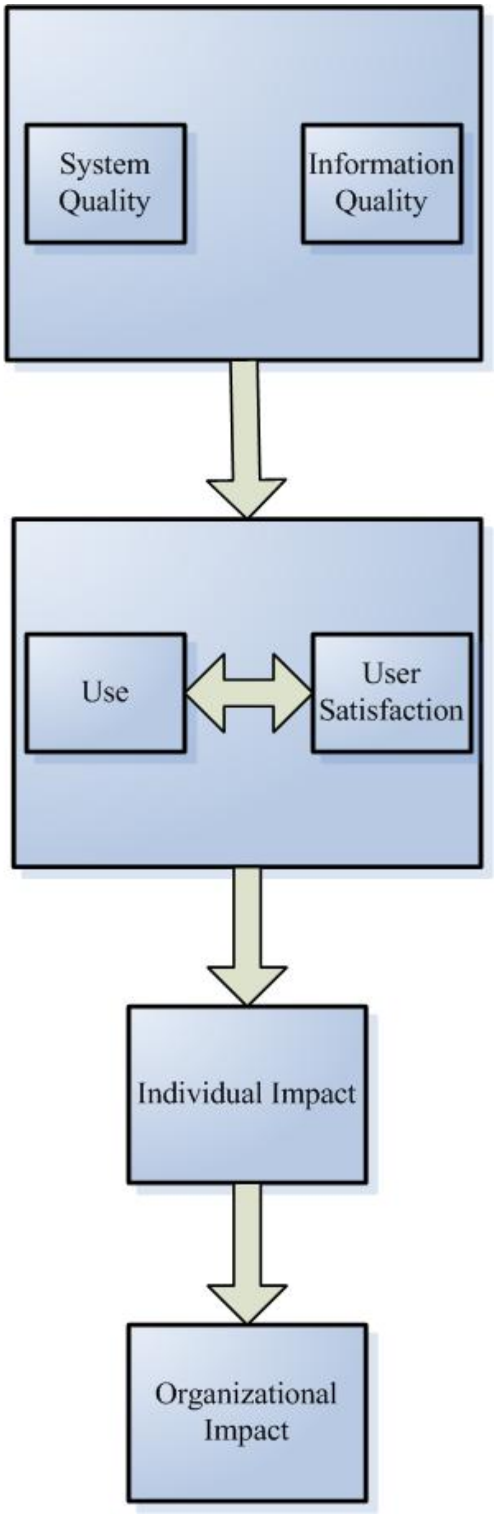

Figure 1. Information Workflow

Despite the short history of data quality as a distinct field of inquiry, data quality professionals have been developing, revising and creating rules for data quality problem solving for a long time, and are often forced by urgency to do so in unconventional ways that bypass standard organizational procedures (Lee, 2003). According to Cabellero, Caro, Calero, and Piattini (2008) companies must assess and improve the quality of information using data quality standards within their systems in order to improve business decisions. With the transfer of information constantly taking place, it is critical that the data within the systems have an adequate level of quality (Guerra-Garcia, Caballero, Piattini & Springer, 2013).

The purpose of this annotated bibliography is to identify literature that addresses how to improve management of data quality within information systems (Cappiello, Francalanci & Pernici, 2003). Organizations continue to be increasingly dependent on data-driven technologies, which makes the importance of managing the quality of data critical (Parssian, Sarkar & Jacob, 2004). Following the perspective presented by Cappiello, Francalanci and Pernici (2003), focus in this study is on how data quality can be better managed (and measured) in relation to the four categories they suggest: accuracy, consistency, currency, and completeness.

Audience

This annotated bibliography is written for system developers who create and manage information systems and the managers who determine the strategy an organization pursues for capturing and storing data. Information quality must be managed throughout the organization; managers and system developers are responsible for information quality made available to system users through business procedures and with the use of technology (Ryu, Park, & Park, 2006). Popovic, Coelho, and Jaklic (2009) suggest that quality data in information systems is dependent on a *well designed* system with the software tools that provide workers with timely access, effective analysis and intuitive presentation of the right information, enabling them to make the right business decisions.

Research Question

Main question. How do managers in today's organizations ensure that information systems are storing and reporting quality information? How can systems developer and managers improve *data quality* within information systems?

Search Report

Search strategy. References on the topic of information quality including the dimensions of data quality, impacts on an organization and possible solutions to repair existing information systems or developing new information systems are found using the UO Libraries website and Google Scholar. Unfamiliar words within the articles are clarified in order to understand their context within the text. The following key words and combination phrases are used to search for journal articles: *dirty data*, *data quality, strategic information systems, information system structure, evaluation of information systems, information quality, use of information systems, user satisfaction, information system success, and system quality*. Supporting references within the selected journal articles are analyzed and retrieved within the databases in order to find additional articles that support the topic of information quality management.

Key words used to extend the search to find relevant articles include: *dirty data*, *information quality and data quality.* Keywords found within reference articles are also used to search for additional journal articles to support the accumulation of reference articles. Creswell (2009) recommends locating recent journal articles and documents that support the topic; journal articles from a respected national journal that are peer reviewed are an excellent source to find supporting information.

Evaluation criteria. The evaluation of the articles located in the annotated bibliography rely upon Bell and Frantz (2013) suggested key areas to evaluate articles based on (a) the authority of the author, (b) the objectivity of the author, (c) the quality of the work, (d) the currency of the work and (e) the relevancy of the work. Methods such as the use of keywords and combination of phrases are used to search for relevant articles that discuss the importance of quality within organizations' information systems. Many articles are available on the individual components of information systems; though very few recent articles exist that encompass the sum of the many parts that make up data quality in an information system. Reviewing the title, abstract and a publication date (within the last 10 years) determines if the information within the article is relevant. Selected articles are reviewed, looking for supporting information that describes data quality dimensions, particularly in relation to four categories: accuracy, consistency, currency, and completeness (Cappiello, Francalanci and Pernici, 2003). Information related to the management of data quality within information systems is highlighted.

Search engines and databases. The following databases are accessed during the information search:

- Academic Search Primer
- Annual Reviews
- ArticleFirst
- PAIS
- Sociological Abstracts
- JSTOR
- Project Muse
- Web of Science

- Google Scholar
- UO Catalog

The academic databases produced a large number of articles covering the topic of data quality within information systems and managing information systems. Not all articles found during the research have *full text* available without additional costs and those articles requiring payment were eliminated. *Pay for articles* that were found using Google Scholar is searched in the UO catalog in order to gain access without additional costs. The reverse was also true; some articles in the UO catalog are not free; these articles are accessed using Google Scholar in an effort to find a source that does not charge for access.

Documentation approach. Full text articles selected are downloaded and reviewed in Adobe *pdf* format. Tools within Adobe Reader are used to highlight key points of the article that are documented in a Microsoft Word *references* document saved on a local computer. The published abstract and reference citation is documented in APA format, using the search results found in the UO Library search functions; *digital object identifier* links are copied into the Microsoft Word *reference* document. Key points found within journal articles related to the main research question are reviewed and summarized in the Annotated Bibliography section of this document.

Annotated Bibliography

This section of the document presents 15 references selected to address the main research questions posed in this study: How do managers in today's organizations ensure that information systems are storing and reporting quality information? How can systems developers and managers improve *data quality* within information systems? Each of the 15 references includes an annotation that contains: (a) a full bibliographic citation, (b) a published abstract, and (c) a summary. Summaries include descriptions of how the reference relates to the research questions presented by (a) introducing methods that existing managers use to manage information system data quality and (b) presenting methodologies for managers and developers to analyze, prepare, and implement changes to existing systems to improve data quality. Common data quality dimensions provided by Cappiell, Francalanci and Pernici (2003), including (a) completeness, (b) accuracy, (c) consistency and (d) currency, are used to define data qualities in an information system.

Batini, C., Cappiello, C., Francalanci, C. & , Maurino, A., (2009). Methodologies for data quality assessment and improvement. *Acm Computing Surveys, 41*(3). Retrieved from: http://delivery.acm.org.libproxy.uoregon.edu/10.1145/1550000/1541883/a16-batini.pdf?ip=128.223.86.31&id=1541883&acc=ACTIVE%20SERVICE&key=C2716FEBFA981EF15542EBFCB1385A8FF2B5F7F13CB63901&CFID=264232693&CFTOKEN=56203701&__acm__=1385324173_574458b8753628896710c0845ddb9332

Abstract. The literature provides a wide range of techniques to assess and improve the quality of data. Due to the diversity and complexity of these techniques, research has recently focused on defining methodologies that help the selection, customization, and application of data quality assessment and improvement techniques. The goal of this article is to provide a systematic and comparative description of such methodologies. Methodologies are compared along several dimensions, including the methodological phases and steps, the strategies and techniques, the data quality dimensions, the types of data, and, finally, the types of information systems addressed by each methodology. The article concludes with a summary description of each methodology.

Summary. The article covers both (a) how managers in today's organizations ensure their systems are housing quality data and (b) how system developers and managers can improve *data quality* within information systems using techniques identified within the article. This was one of multiple articles that introduced the dimensions of data quality that included *accuracy*, *timeliness*, *consistency* and *completeness*. The goal of this article is to compare and explain existing data quality techniques as they apply to different information systems. Two main techniques by which data systems can be analyzed are (a) data driven strategies and (b) process driven strategies. The article discusses basic data quality issues that can be found in all systems and some common methods to improve data quality within those systems. Common methods include (a) data-driven strategies that improve the quality of data by reviewing existing data and updating data that is found to be inaccurate and (b) process-driven strategies that improve data quality by redesigning the processes that create or modify data.

Caballero, I., Caro, A., Calero, C. & Piattini, M., (2008). Iqm3: Information quality management maturity model. *Journal of Universal Computer Science, 14*(22), 3658-3685. Retrieved from: http://www.jucs.org/jucs_14_22/iqm3_information_quality_management

Abstract. In order to enhance their global business performance, organizations must be careful with the quality of their information since it is one of their main assets. Analogies to quality management of classical products demonstrate that Information Quality is also preferably attainable through management by integrating some corresponding Information Quality management activities into the organizational processes. To achieve this goal we have developed an Information Quality Management Framework (IQMF). It is articulated on the concept of Information Management Process (IMP), based on the idea of Software Process. An IMP is a combination of two sub-processes: the first, a production process, aimed to manufacture information from raw data, and the second to adequately manage the required Information Quality level of the first. IQMF consists of two main components: an Information Quality Management Maturity Model (IQM3), and a Methodology for the Assessment and Improvement of Information Quality Management (MAIMIQ), which uses IQM3 as a reference model for the assessment and for the improvement goal of an IMP. Therefore, as a result of an assessment with MAIMIQ, an IMP can be said to have raised one of the maturity levels described in IQM3, and as improvement goal, it would be desirable to achieve a higher maturity level. Since an Information System can be seen as a set of several IMPs sharing several resources, it is possible to improve the Information Quality level of the entire Information System by improving the most critical IMPs. This paper is focused only on describing the foundations and structure of IQM3, which is based on staged CMMI.

Summary. This article addresses how system developers and mangers can improve *data quality* within information systems by using the IQM3 model to analyze the present system and develop an improvement plan. The authors stress the impact of information on business processes and present a model of information quality management. A role for the information quality manager is recommended within the organization to ensure information quality practices are standardized throughout the company. There are two models introduced to assist in the management of information quality. The first model is Information Quality Management Maturity Model, which treats the data in an information system as a product; this model analyzes each variable such as users and developers who create data within the system. The second model introduced is used for assessing and developing an improvement plan for information quality within the organization. The models presented in the article provide a system to analyze and solve information quality problems.

Cappiello, C., Francalanci, C. & Pernici, B. (2003). Time-related factors of data quality in multichannel information systems. *Journal of Management Information Systems, 20*(3), 71-91. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/pdfplus/40398641.pdf?acceptTC=true&acceptTC=true&jpdConfirm=true

Abstract. Modern organizations offer services through multiple channels, such as branches, ATMs, telephones, and Internet sites, and are supported by multifunctional software architectures. Different functional modules share data, which are typically stored in multiple local databases. Functional modules are usually not integrated across channels, as channels are implemented at different times within independent software projects and are subject to varying requirements of availability and performance. This lack of channel and functional integration raises data quality problems that can impact the quality of the products and services of an organization. In particular, in complex systems in which data are managed in multiple databases, timeliness is critical. This paper focuses on time-related factors of data quality and provides a model that can help companies to evaluate data currency, accuracy, and completeness in software architectures with different degrees of integration across channels and functionalities. The model is validated through simulation based on empirical data on financial information systems. Results indicate how architectural choices on the degree of data integration have a varying impact on currency, accuracy, and completeness depending on the type of financial institution and on customer profiles.

Summary. Data dimensions are defined and discussed as contributors to data quality within information systems. The focus of currency, accuracy, and completeness of data is further researched when comparing multiple systems used by companies today. The article addresses both research questions covered in this paper. By comparing systems from multiple organizations, the article identifies what some organizations do to ensure quality data is present in their systems. The author offers a model using mathematics to determine the time-related accuracy of stored data, and identifies a method for improving data quality in the future. Data currency has a direct impact to the other quality dimensions of accuracy, completeness and consistency. For example, delays in propagating data changes across multiple systems can cause inaccuracy and incomplete data to be returned in a search. To improve the currency of the data, linked databases must be kept up-to-date so that the same information stored in multiple sources will be consistent. The financial industry is used as a model of how multiple systems must work together and how critical it is to have the data within the system be accurate and up-to-date. The article suggests that organizations increase the integration of their information systems so that information is updated across all systems at an increased frequency in order to decrease the likelihood that data currency will affect accuracy and completeness.

Gelman, I. (2010). Setting priorities for data accuracy improvements in satisficing decision-making scenarios: A guiding theory. *Decision Support Systems, 48*(4), 507-520. Retrieved from: http://www.sciencedirect.com.libproxy.uoregon.edu/science/article/pii/S0167923609002218#

Abstract. This study introduces a mathematical-statistical theory that illustrates the effect of input errors on the accuracy of dichotomous decisions which are implemented through logical conjunction and disjunction of selected criteria. Decision-making instances in this category are often labeled "satisficing." Mainly, our theory provides criteria for ranking the effect of errors in different inputs on decision accuracy. This ranking can be used to improve the efficiency and effectiveness of resource allocation decisions in data quality management settings. All other things being equal, inputs in which errors exhibit a higher negative effect on the output would naturally earn higher priority.

Summary. This article offers a literature review of the relationship between input and output accuracy of an information system. Using existing models such *as data quality algebra,* which uses mathematical calculations to compare raw data and the quality of query outputs, the author, establishes the need to improve the quality of data entry into the system. Real world examples are provided that address both research questions of how existing managers ensure data quality, as well as how managers and developers can improve data quality management in the future by using some of the tools provided in the article. Managers and developers can further analyze the data results performed from a query and target specific areas that need improvement. Poor data within the system should not be treated equally and the models presented in the article help managers identify key areas to improve data quality.

Guerra-Garcia, C., Caballero, I., Piattini, M. & Springer. (2013). A framework for designing data quality aware web applications. *Information Systems Frontiers, 15*(3), 433-445. Retrieved from: http://ssmvm030.mit.edu/ICIQ/Documents/IQ%20Conference%202011/Papers/02_01_ICIQ2011.pdf

Abstract. The number of Web applications which are part of Business Intelligence (BI) applications has grown exponentially in recent years, as has their complexity. Consequently, the amount of data used by these applications has also increased. The larger the number of data used, the greater the chance to make errors is. That being the case, managing data with an acceptable level of quality is paramount to success in any organizational business process. In order to raise and maintain adequate levels of Data Quality (DQ), it is indispensable for Web applications to be able to satisfy specific DQ requirements. To do so, DQ requirements should be captured and introduced into the development process of the web application, together with the other software requirements needed in the applications. In the field of web application development, however, there appears to us to exist a lack of proposals aimed at managing specific DQ software requirements. This paper considers the MDA (Model Driven Architecture) approach and, principally, the benefits provided by Model Driven Web Engineering (MDWE), putting forward a proposal for two artifacts. These consist of a meta model and a UML profile for the management of Data Quality Software Requirements for Web Applications.

Summary. While this article is meant to address the second research question of how managers and developers can improve data quality within an information system, it largely focuses on Web applications design, though the recommendations can apply to any information system. Organizations increasingly rely on Web applications to access data stored in multiple data sources, as well as allow data entry into Web applications. A case study is used to provide an example of how the use of Web engineering using metadata analysis can improve data quality. Data quality dimensions are provided and data entered into the system is checked against these dimensions to ensure the accuracy of information entered into the system. There is an increased need for organizations to manage data quality, which starts by ensuring only quality data can enter the system. The article identifies current deficiencies in Web design applications and offers a model to improve Web application design. The author suggests that future work will include his own approach at solving this problem by creating a tool that will support all phases of the development cycle.

Heinrich, B. & Klier, M., (2011). Assessing data currency - a probabilistic approach. *Journal of Information Science, 37*(1), 86-100. Retrieved from: http://jis.sagepub.com.libproxy.uoregon.edu/content/37/1/86.full.pdf+html

Abstract. The growing relevance of data quality has revealed the need for adequate measurement. As time aspects are extremely important in data quality management, we propose a novel approach to assess data currency. Our metric, which is founded on probability theory, enables an objective and widely automated assessment for data liable to temporal decline. Its values are easy to interpret by business users. Moreover, the metric makes it possible to analyze the economic impacts of data quality measures like data cleansing and can therefore build a basis for an economic management of data quality. The approach can be applied in various fields of application where the currency of data is important. To illustrate the practical benefit and the applicability of the novel metric, we provide an extensive real world example. In cooperation with a major German mobile services provider, the approach was successfully applied in campaign management and led to an improved decision support.

Summary. This article identifies the key components that make up *data quality dimensions* which include: completeness, correctness and currency of the data within the information system. The article supports the second research question of how managers and developers can improve *data quality* within an information system. Ensuring that the information within the system is up to date and relevant to the person performing the search is a key aspect of data quality. Organizations and employees rely on data to aid them in making critical decisions. The article details the definition of *currency*; the relevancy of the data within the information system and how the age of the data becomes a factor. The authors offer a data quality metric in the form of mathematical equations to aid in determining when the data is no longer useful.

Jiang, Z., Sarkar, S., De, P. & Dey, D. (2007). A framework for reconciling attribute values from multiple data sources. *Management Science, 53*(12), 1946-1963. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/20122350

Abstract. Because of the heterogeneous nature of different data sources, data integration is often one of the most challenging tasks in managing modern information systems.

While the existing literature has focused on problems such as schema integration and entity identification, it has largely overlooked a basic question: When an attribute value for a real-world entity is recorded differently in different databases, how should the "best" value be chosen from the set of possible values? This paper provides an answer to this question. We first show how a probability distribution over a set of possible values can be derived. We then demonstrate how these probabilities can be used to solve a given decision problem by minimizing the total cost of type I, type II, and misrepresentation errors. Finally, we propose a framework for integrating multiple data sources when a single "best" value has to be chosen and stored for every attribute of an entity.

Summary. This journal article addresses both research questions presented in this paper. The article uses real world problems as an example of how current managers make business decisions based on data retrieved from multiple data sources. Inaccurate information can result in unnecessary costs to the organization, which is why there is a need to improve data quality within information systems. The research focuses on trying to identify conflicts within the system using a system that will detect similar information stored in multiple data sources and *output* to the user the *best* data. Whereas some off the shelf solutions exist, the author continues the research of how to improve nonsystematic causes of data conflicts within the system. Mathematical modeling is used to factor existing data quality health of the information system and is also used to aid the information system in determining which data to report out. Developers and managers can use the tools presented in this research article to cleanse inaccurate or irrelevant data from the system.

Kim, W., Choi, B., Hong, E., Kim, S. & Lee, D., (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery, 7*(1), 81-99. Retrieved from: http://link.springer.com.libproxy.uoregon.edu/article/10.1023/A:1021564703268

Abstract. Today large corporations are constructing enterprise data warehouses from disparate data sources in order to run enterprise-wide data analysis applications, including decision support systems, multidimensional online analytical applications, data mining, and customer relationship management systems. A major problem that is only beginning to be recognized is that the data in data sources are often “dirty”. Broadly, dirty data include missing data, wrong data, and non-standard representations of the same data. The results of analyzing a database/data warehouse of dirty data can be damaging and at best be unreliable. In this paper, a comprehensive classification of dirty data is developed for use as a framework for understanding how dirty data arise, manifest themselves, and may be cleansed to ensure proper construction of data warehouses and accurate data analysis. The impact of dirty data on data mining is also explored.

Summary. This article provides many definitions that support this research. This article introduces a taxonomy of *dirty data* which breaks up the different ways that dirty data can enter into the system. Many industries such as business, technology and healthcare use some type of data warehouse to store data and access the data with the use of business intelligence software. The data retrieved by either the user or the application will only be as good as the information stored within the data systems. A model to define and measure data quality within an information system is introduced as a tool to improve data quality. While the introductory challenges presented in the article provide an example of how today's organizations deal with data quality and some of the methods today's mangers use in an effort to store quality data within their systems; the taxonomy presented provides a solution for developers and mangers to analyze their existing systems in order to improve data quality in the future. The article covers both research questions covered in this paper.

Kumar, V. & Tharaja, R. (2013). A simplified approach for quality management in data warehouses. *n.p.* Retrieved from: http://arxiv.org/abs/1310.2066

Abstract. Data warehousing is continuously gaining importance as organizations are realizing the benefits of decision oriented data bases. However, the stumbling block to this rapid development is data quality issues at various stages of data warehousing. Quality can be defined as a measure of excellence or a state free from defects. Users appreciate quality products and available literature suggests that many organization`s have significant data quality problems that have substantial social and economic impacts. A metadata based quality system is introduced to manage quality of data in data warehouse. The approach is used to analyze the quality of data warehouse system by checking the expected value of quality parameters with that of actual values. The proposed approach is supported with a metadata framework that can store additional information to analyze the quality parameters, whenever required.

Summary. This article focuses on the second research question and is meant to aid managers and developers in improving data quality within information systems; specifically data warehouses. The issue of poor data quality within an information system is well documented in the article and a model is used to identify existing problem areas, define quality metrics to improve quality, develop a plan and verify the results. A list of *twenty-three* quality parameters that affect data quality is listed that explains the relationship between the parameter and the effects on the quality metric outputs. The article also lists the relevant stakeholders who are impacted by data quality issues or who play a key role in managing the data quality within the information system. Managers and system developers can improve data quality by proactively performing error detection and cleansing existing data by performing the following steps, (a) *prevention* of poor data entering the system with the use of improved entry fields, (b) *audit* of information that enters the system by checking the output of a query, (c) *filtering* incorrect data found during the audit process and, (d) *correcting* system rules that caused poor audit results.

Lee, Yang. (2003). Crafting rules: Context-reflective data quality problem solving. *Journal of Management Information Systems, 20*(3), 93-119. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/40398642

Abstract. Motivated by the growing importance of data quality in data-intensive, global business environments and by burgeoning data quality activities, this study builds a conceptual model of data quality problem solving. The study analyzes data quality activities at five organizations via a five-year longitudinal study. The study finds that experienced practitioners solve data quality problems by reflecting on and explicating knowledge about contexts embedded in, or missing from, data. Specifically, these individuals investigate how data problems are framed, analyzed, and resolved throughout the entire information discourse. Their discourse on contexts of data, therefore, connects otherwise separately managed data processes, that is, collection, storage, and use. Practitioners' context-reflective mode of problem solving plays a pivotal role in crafting data quality rules. These practitioners break old rules and revise actionable dominant logic embedded in work routines as a strategy for crafting rules in data quality problem solving.

Summary. The contents of this journal article identify how managers in a sample of organizations handled the issue of data quality, which supports the first research question of how today's managers ensure their information systems are storing and retrieving quality data. The study followed several companies over several years to understand how executives and managers make business decisions based on their information systems. Data and information quality have long been an issue within every business sector; businesses have changed the way they use data. Data continues to become more complex, organizations rely upon it more for making critical decisions and the problem of data quality is constantly changing. The journal article explores the understanding that business professionals must have of data quality issues within their organization and identifies ways to correct those issues. Several companies are used as subjects of a study to enable real-world research into data quality issues. The study helps identify the data quality issues, develop an action plan to correct the issue and uses reflection-in-action problem solving to observe the results from performing the corrective action.

Madnick, S., Wang, R. & Xian, X. (2003). The design and implementation of a corporate house holding knowledge processor to improve data quality. *Journal of Management Information Systems, 20*(3), 41-69. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/40398640

Abstract. Awareness of data and information quality issues has grown rapidly in light of the critical role played by the quality of information in our data-intensive, knowledge-based economy. Research in the past two decades has produced a large body of data quality knowledge and has expanded our ability to solve many data and information quality problems. In this article, we present an overview of the evolution and current landscape of data and information quality research. We introduce a framework to characterize the research along two dimensions: *topics* and *methods*. Representative papers are cited for purposes of illustrating the issues addressed and the methods used. We also identify and discuss challenges to be addressed in future research.

Summary. The authors define the term *corporate house-holding* as the process of capturing, analyzing and managing corporate information. A simple description might be that corporate house-holding is a process of accumulating and maintaining the critical information held within an organization's data system that is used to make business decisions. The authors study *entity aggregation,* which is described as data that can be represented in multiple ways within the data system; one of the three elements that make up corporate house-holding. The authors assert that data within the system must be of high quality in order for employees to make the right business decisions. This article also presents data dimensions and discusses their relevancy to data quality in an information system.

This article supports both research questions in this paper. The first research question is supported by examples of how organizations utilize MIT's Total Data Quality Management program in an attempt to control data dimensions such as accuracy, accessibility, timeliness, believability and relevance of data. The article provides real world business examples on challenges that companies face when using dirty data within their information systems and the cost of doing business when decisions are made using incorrect data. A modeling technique is presented that will identify the information product being created, methods to evaluate data quality and capabilities to manage data quality. The article addresses the second research question of identifying what managers and system developers can do to improve data quality with a solution designed to use technology to cleanse and consolidate data found across multiple systems, thus reducing and/or eliminating the house-holding problem of entity aggregation.

Nelson, R., Todd, P. & Wixom, B. (2005). Antecedents of information and system quality: An empirical examination within the context of data warehousing. *Journal of Management Information Systems, 21*(4), 199-235. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/40398737

Abstract. Understanding the successful adoption of information technology is largely based upon understanding the linkages among quality, satisfaction, and usage. Although the satisfaction and usage constructs have been well studied in the information systems literature, there has been only limited attention to information and system quality over the past decade. To address this shortcoming, we developed a model consisting of nine fundamental determinants of quality in an information technology context, four under the rubric of information quality (the output of an information system) and five that describe system quality (the information processing system required to produce the output). We then empirically examined the aptness of our model using a sample of 465 data warehouse users from seven different organizations that employed report-based, query-based, and analytical business intelligence tools. The results suggest that our determinants are indeed predictive of overall information and system quality in data warehouse environments, and that our model strikes a balance between comprehensiveness and parsimony. We conclude with a discussion of the implications for both theory and the development and implementation of information technology applications in practice.

Summary. This article supports the research question of what today's managers are doing to ensure their information systems are using quality data by identifying some existing tools that measure data quality. Since the early 80's, organizations have been using tools such as Total Quality Management (TQM) and other quality techniques to measure data quality. The article addresses the second research question of how system developers and managers can improve data quality by concentrating on the design of the information system. The authors suggest that the quality of the system is equal in importance to the actual quality of the information within the system in order to develop information technology applications with the potential for successful use.

Data quality dimensions are defined in the article as being (a) accurate, (b) complete, (c) relative, and (d) current. Where the dimensions of data are well documented, the defining attributes of system quality are also listed. Successful information systems are typically defined as systems that (a) are accessible, (b) are reliable, (c) are flexible, (d) have good response time, and (e) are easily integrated. This information is particularly useful to managers and system developers that are trying to improve data quality as they can look at the overall quality of the system in which the data is stored. Information systems like data warehouses and business intelligence applications are tested in different industries in order to determine which systems are affected most by individual quality dimensions. The author has identified a model for testing the quality of information systems based on the data quality dimensions, but identifies that additional research must be completed before a solution can be developed.

Parssian, A., Sarkar, S. & Jacob, V. (2004). Assessing data quality for information products: Impact of selection, projection and Cartesian product. *Management Science, 50*(7), 967-982. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/30047952

Abstract. The cost associated with making decisions based on poor-quality data is quite high. Consequently, the management of data quality and the quality of associated data management processes has become critical for organizations. An important first step in managing data quality is the ability to measure the quality of information products (derived data) based on the quality of the source data and associated processes used to produce the information outputs. We present a methodology to determine two data quality characteristics-accuracy and completeness-that are of critical importance to decision makers. We examine how the quality metrics of source data affect the quality for information outputs produced using the relational algebra operations selection, projection, and Cartesian product. Our methodology is general, and can be used to determine how quality characteristics associated with diverse data sources affect the quality of the derived data.

Summary. The study performed in this article concentrates on two aspects of data quality, accuracy and completeness, which are two of the characteristics that make up data quality dimensions. The authors assert that accuracy and completeness are the two attributes of data quality that are cited the most and are the easiest to track in the information system. Algebraic equations are used to define a metric for analyzing accuracy and completeness in a relational database information system. Examples are provided of how this model can be applied to different scenarios to predict the impact of data collected from multiple sources using the process of (a) selection, (b) projection and (c) Cartesian product operations (data with two relations). The article supports the second research question by providing managers and system developers with another method of analyzing the quality of data within an organization's information system so that a plan can be made to improve data quality using the proposed methodology. The proposed methodology is meant to be applicable to other types of information systems and not just databases.

Popovic, A., Coelho, P. & Jaklic, J. (2009). The impact of business intelligence system maturity on information quality. *Information Research-an International Electronic Journal, 14*(4). Retrieved from: http://apps.webofknowledge.com.libproxy.uoregon.edu/InboundService.do?SID=1APmw4XFPJIgXn6Nd2d&product=WOS&UT=WOS%3A000274193700005&SrcApp=METALIBSearch&DestFail=http%3A%2F%2Fwww.webofknowledge.com&Init=Yes&action=retrieve&Func=Frame&customersID=EXLIBRIS&SrcAuth=EXLIBRIS&IsProductCode=Yes&mode=FullRecord

Abstract. We propose and test a model of the relationship between business intelligence systems and information quality and investigate in more detail the potential differential impact of business intelligence systems' maturity on two aspects of information quality: the quality of content and media quality.

Summary. This article describes the impact that business intelligence (BI) applications can have on organizations when coupled with technology and organizational changes. BI applications have the ability to perform (a) data cleansing, (b) consolidation of business definitions, and (c) master data management. Many of today's organizations use some variation of BI software in order to provide their employees with a method to access consolidated data from multiple internal or external sources. This article covering BI supports the first research question of how today's managers are ensuring their information systems are storing and reporting quality information. BI applications rely on information quality, which is defined in this article as the *content of the information* and *its accessibility.* Information content can further be explained as the dimensions that make up data quality such as (a) accuracy, (b) consistency, (c) completeness and (d) currency. BI applications can offer improved information quality during performed queries and reporting, however BI applications often struggle when information stored across multiple sources does not match. For example, if IBM was a customer or supplier, different departments might list the name as IBM or International Business Machines. The maturity growth of the BI application would eventually have the ability to locate this discrepancy and either understand it or correct it. A model is presented to find the relationship between the maturity of BI systems and information quality, which results in identifying the need for higher data management in order to get the most out of the BI systems in place. Managers must ensure BI systems contain (a) relevant information, (b) sound information, (c) optimized processes and (d) reliable infrastructure. BI systems can be analyzed using the model to determine (1) how the available data is integrated and (2) whether data in the data sources are consistent across multiple sources. The success of the BI system relies upon the quality of the data that the system is accessing.

Ryu, K., Park, J. & Park, J. (2006). A data quality management maturity model. *Etri Journal, 28*(2)2, 191-204. Retrieved from: http://apps.webofknowledge.com.libproxy.uoregon.edu/InboundService.do?SID=3DDJA

CgCFoE5SgEqq7H&product=WOS&UT=WOS%3A000236754800007&SrcApp=METALIBSearch&DestFail=http%3A%2F%2Fwww.webofknowledge.com&Init=Yes&action=retrieve&Func=Frame&customersID=EXLIBRIS&SrcAuth=EXLIBRIS&IsProductCode=Yes&mode=FullRecord

Abstract. Many previous studies of data quality have focused on the realization and evaluation of both data value quality and data service quality. These studies revealed that poor data value quality and poor data service quality were caused by poor data structure. In this study we focus on metadata management, namely, data structure quality and introduce the data quality management maturity model as a preferred maturity model. We empirically show that data quality improves as data management matures.

Summary. This study addresses the second search question, how managers and developers of information systems can improve data quality. The article discusses how impactful poor data quality can be to an organization that utilizes data within its information systems to make critical business decision. A data management model is proposed using metadata as a way to evaluate existing data quality management practices. The model aids managers and system developers in introducing a higher level of data quality through the use of standardization practices throughout the company. The study concludes by introducing future ideas on how to manage data quality.

Conclusion

Heinrich and Klier (2011) conclude that executives and employees require high-quality data in order to make the critical business decisions. Data quality within information systems can affect any company in all business sectors and the relevance of poor data quality is constantly growing (Cappiello, Francalanci & Pernici, 2003). Information flows throughout the organization and the problem of information quality is an organizational issue that must be addressed and constantly monitored (Caballero, Caro & Piattini, 2008). Batini, Cappiello, Francalanci and Maurino (2009) suggest that poor information quality can increase unnecessary cost including (a*) process costs* that require the same task to be performed multiple times due to data errors and (b) *opportunity costs* due to lost and missed revenues; whereas the cost of a data quality program can be considered a preventive cost in order to reduce data errors.

Managing data quality. Data quality and the management of information systems affect virtually all companies in every business sector that rely on data to make business decisions (Cappiello, Francalanci & Pernici, 2003). There is no a single solution that can be put into place that is applicable to all data quality problems; however, there are multiple methods to analyze existing systems, determine problem areas, and develop a plan to resolve system faults (Batini, Cappiello, Francalanci & Maurino, 2009). According to Kim, Choi, Hong, Kim and Lee (2003), the information system itself must be of high quality in order for users to enter quality information into the system so that information system applications such as data warehouses and business intelligence software can produce accurate, relevant, current and timely data results.

Improvements can be made using models similar to Caballero, Caro and Piattini's (2008) Information Quality Management Framework, in which data is first treated like a *product*; this model analyzes each variable such as users and developers who create data within the system. A second model presented by Caballero, Caro and Piattini's (2008) is used for assessing and developing an improvement plan for information quality within the organization.

Guerra-Garcia, Caballero, Piattini, and Springer (2013) suggest that by managing the application during the design phase, developers can program mechanisms that will allow the application to check for accuracy, completeness, currency and consistency during data entry into the information system. Another methodology to analyze existing information system quality is presented by Jiang, Sarkar, De, and Dey (2007); by using algebra to assign inputs into a mathematical equation based on the source databases, data from multiple systems can be compared to determine the *best* value that meets data quality standards. Kumar and Tharaja (2013) demonstrate how metadata can be used to manage the quality of data in a data warehouse. The approach is used to analyze and improve the quality of data warehouse systems by (a) prevention, (b) auditing, (c) filtering, and (d) correcting system faults (Kumar & Tharaja, 2013). These illustrations of data quality system management offer ways to improve data quality in an information system.

Data quality dimensions. Table 1 provides a list of dimensions that make up data quality, as presented by Cappiello, Francalanci and Pernici (2003).

Table 1.

Data Quality Dimensions

Data Quality Dimension	Impact on Data Quality
Accuracy	The extent to which data is correct, reliable and certified.
Consistency	Similar information content found in multiple systems matches
Currency	The time point in which the data was entered into the system and the relevancy when retrieved.

Completeness	The degree to which a specific database includes all the values corresponding to a complete representation of a given set of *real world* events as database entities.

Poor data in an information system can be defined as *dirty data*, which is data that is (a) missing, (b) not missing but wrong, or (c) unusable (Kim, Choi, Hong, Kim & Lee, 2003). Due to the impact that data quality can have on an organization's ability to succeed, Heinrich and Klier (2011) stress the importance of analyzing the quality of the information systems and the data held within those systems, so that managers and developers can develop plans to improve data quality. Data quality within an information system is regarded as the most important factor because it is the basis of the information system (Ryu, Park & Park, 2006).

Improving data quality. Management of data quality and the quality of data management processes have been identified as critical issues for organizations (Parssian, Sakar & Jacob, 2004). Popovi, Coelho and Jaklic (2009) believe that managers and developers can improve data quality with the use of business intelligence (BI) applications that include (a) faster access to information, (b) easier querying and analysis, and (c) improved data consistency due to integration. The integration of data from different systems within BI applications is meant to provide employees at various levels in the organization with timely, relevant and easy to use information (Popovic, Coelho & Jaklic, 2009). Though BI applications offer some improvements in data quality due to advanced search methods, search results are only as good as the information stored within the system (Kim, Choi, Hong, Kim & Lee, 2003).

Although no single data quality management solution can be applied to an information system to eliminate all data quality problems, Batini, Cappiello, Francalanci and Maurino (2009) assert that data quality improvements can be achieved through (a) data driven strategies that improve the quality of existing data by reviewing and updating data that is found to be inaccurate within the information system, and (b) process driven changes that reduce the entry of dirty data into the system by redesigning the processes that create or modify data. Managers and developers of information systems must design and implement a system that reduces dirty data through the use of both data and process driven strategies. They must also continue to manage the quality of the data within the system as well as manage the information system itself.

References

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *Acm Computing Surveys, 41*(3). Retrieved from: http://delivery.acm.org.libproxy.uoregon.edu/10.1145/1550000/1541883/a16-batini.pdf?ip=128.223.86.31&id=1541883&acc=ACTIVE%20SERVICE&key=C2716FEBFA981EF15542EBFCB1385A8FF2B5F7F13CB63901&CFID=264232693&CFTOKEN=56203701&__acm__=1385324173_574458b8753628896710c0845ddb9332

Bell, C., & Frantz, P. (2013, July). Critical evaluation of information sources. University of Oregon Libraries. Retrieved from http://library.uoregon.edu/guides/findarticles/credibility.html

Caballero, I., Caro, A., Calero, C. & Piattini, M. (2008). Iqm3: Information quality management maturity model. *Journal of Universal Computer Science, 14*(22), 3658-3685. Retrieved from: http://www.jucs.org/jucs_14_22/iqm3_information_quality_management

Cappiello, C., Francalanci, C. & Pernici, B. (2003). Time-related factors of data quality in multichannel information systems. *Journal of Management Information Systems, 20*(3), 71-91. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/pdfplus/40398641.pdf?acceptTC=true&acceptTC=true&jpdConfirm=true

Creswell, J. (2009). Research design: Qualitative, quantitative, and mixed methods approaches (Kindle Edition) Sage Publications - A. Kindle Edition.

Gelman, I. (2010). Setting priorities for data accuracy improvements in satisficing decision-making scenarios: A guiding theory. *Decision Support Systems, 48*(4), 507-520. Retrieved from:

http://www.sciencedirect.com.libproxy.uoregon.edu/science/article/pii/S0167923609002218#

Guerra-Garcia, C., Caballero, I., Piattini, M. & Springer. (2013). Capturing data quality requirements for web applications by means of DQ WebRE. *Information Systems Frontiers, 15*(3), 433-445. Retrieved from: http://ssm-vm030.mit.edu/ICIQ/Documents/IQ%20Conference%202011/Papers/02_01_ICIQ2011.pdf

Heinrich, B. & Klier, M. (2011). Assessing data currency - a probabilistic approach. *Journal of Information Science, 37*(1), 86-100. Retrieved from: http://jis.sagepub.com.libproxy.uoregon.edu/content/37/1/86.full.pdf+html

Jiang, Z., Sarkar, S., De, P. & Dey, D. (2007). A framework for reconciling attribute values from multiple data sources. *Management Science, 53*(12), 1946-1963. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/20122350

Kim, W., Choi, B., Hong, E., Kim, S. & Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery, 7*(1), 81-99. Retrieved from: http://link.springer.com.libproxy.uoregon.edu/article/10.1023/A:1021564703268

Kumar, V. & Tharaja, R. (2013). A simplified approach for quality management in data warehouses. *n.p.* Retrieved from: http://arxiv.org/abs/1310.2066

Lee, Yang. (2003). Crafting rules: Context-reflective data quality problem solving. *Journal of Management Information Systems, 20*(3), 93-119. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/40398642

Madnick, S., Wang, R., & Xian, X. (2003). The design and implementation of a corporate house holding knowledge processor to improve data quality. *Journal of Management*

Information Systems, 20(3), 41-69. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/40398640

Nelson, R., Todd, P. & Wixom, B. (2005). Antecedents of information and system quality: An empirical examination within the context of data warehousing. *Journal of Management Information Systems, 21*(4), 199-235. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/40398737

Parssian, A., Sarkar, S. & Jacob, V. (2004). Assessing data quality for information products: Impact of selection, projection and Cartesian product. *Management Science, 50*(7), 967-982. Retrieved from: http://www.jstor.org.libproxy.uoregon.edu/stable/30047952

Popovic, A., Coelho, P. & Jaklic, J. (2009). The impact of business intelligence system maturity on information quality. *Information Research-an International Electronic Journal, 14*(4). Retrieved from: http://apps.webofknowledge.com.libproxy.uoregon.edu/InboundService.do?SID=1APmw4XFPJIgXn6Nd2d&product=WOS&UT=WOS%3A000274193700005&SrcApp=METALIBSearch&DestFail=http%3A%2F%2Fwww.webofknowledge.com&Init=Yes&action=retrieve&Func=Frame&customersID=EXLIBRIS&SrcAuth=EXLIBRIS&IsProductCode=Yes&mode=FullRecord

Ryu, K., Park, J. & Park, J. (2006). A data quality management maturity model. *Etri Journal, 28*(2)2, 191-204. Retrieved from: http://apps.webofknowledge.com.libproxy.uoregon.edu/InboundService.do?SID=3DDJACgCFoE5SgEqq7H&product=WOS&UT=WOS%3A000236754800007&SrcApp=METALIBSearch&DestFail=http%3A%2F%2Fwww.webofknowledge.com&Init=Yes&action

=retrieve&Func=Frame&customersID=EXLIBRIS&SrcAuth=EXLIBRIS&IsProductCode=Yes&mode=FullRecord