University of Arkansas, Fayetteville
ScholarWorks@UARK

Industrial Engineering Undergraduate Honors Theses

Industrial Engineering

12-2015

Modeling Information Reliability and Maintenance: A Systematic Literature Review

Daysi A. Guerra Garcia
University of Arkansas, Fayetteville

Recommended Citation
Guerra Garcia, Daysi A., "Modeling Information Reliability and Maintenance: A Systematic Literature Review" (2015). *Industrial Engineering Undergraduate Honors Theses*. 37.
http://scholarworks.uark.edu/ineguht/37

Modeling Information Reliability and Maintenance: A Systematic Literature Review

An Undergraduate Honors College Thesis

Department of Industrial Engineering

College of Engineering

University of Arkansas

Fayetteville, AR

By

Daysi Guerra Garcia

Thesis Advisor: Dr. Ed Pohl

Thesis Reader: Dr. Chase Rainwater

Abstract

Operating a business efficiently depends on effective everyday decision-making. In turn, those decisions are influenced by the quality of data used in the decision-making process, and maintaining good data quality becomes more challenging as a business expands. Protecting the quality of the data and the information it generates is a challenge faced by many companies across all industrial sectors. As companies begin to use data from these large data bases they will need to begin to develop strategies for maintaining and assessing the reliability of the information they generate using this data. A considerable amount of literature exists on data quality but minimal amounts of information exist on information reliability and maintenance. For these reasons, it is important to study the current methods used by businesses to maintain desired levels of data and information quality. The purpose of this paper is to describe and assess the available knowledge relating data quality, reliability and maintenance by performing a systematic literature review in these areas.

Keywords

Information Quality; Data Quality; Information Maintenance; Data Maintenance; Information Reliability; Data Reliability

Table of Contents

1 Introduction and Background

In recent years, technological developments in computational power and data analytics have helped companies across all sectors to gather massive amounts of data about their customers, suppliers and operations, and use data analytics to gather meaningful information from their data to make important strategic decisions. The data revolution has begun, and it refers to the "transformative actions" brought about by the outstanding growth rates in data production (Data Revolution Group, n.d.) (48). The new challenges brought by the data revolution have served to shed light on the data quality issues faced by companies around the world.

In recent years, more academic research has been performed in the area of data quality. In business processes, tasks are dependent on each other, and so is the data used in these tasks, given the nature of operational tasks and the inter-relationship of resources involved. Because of the complexity of business environments, the current landscape of data quality is a multidisciplinary field, and the extant literature on data quality covers a wide range of topics. The main objective of this research is to assess how recent studies performed since 2000 address the issues created by dirty data, and assess new methods focused on improving the quality of data used in business processes.

The first part of this honors thesis describes the method used to approach the current knowledge on the topics of data and information quality in order to gather a first set of related research papers. This thesis also describes how the relevant research papers are classified into different categories based on the respective data lifecycle stages. Furthermore, this paper describes how a taxonomy for the relevant papers was established. The first set of subject categories was later broken down into additional subcategories based on the topics they discussed and the research methods they used to contribute to the data quality research field. In addition, this paper will then cite and summarize the most representative papers of each category and subcategory to illustrate the topics discussed and the methods used. Lastly, this paper will explore opportunities for further academic research in this area.

2 Research Methodology

The research method utilized in this undergraduate thesis is a structured approach to a literature reviewed based on a formal research methodology. The objective of this research project is to assess the available knowledge relating to the topic of data and information quality, and provide a general overview of the current academic research landscape. This research is not an attempt to provide a comprehensive review of all papers on the topic or of any research work performed prior to the year 2000.

This section describes how the approach to understanding the available knowledge on data quality was divided into stages. During the first stage of the research, a very broad approach was taken to gather an initial set of relevant papers that could help to gain a better understanding of how the extant literature addressed data quality, and help to identify the main issues and concerns related to the quality of data used in business processes. Because the term "data quality" was a common and very lightly used term in the available academic research literature, three research constraints were established from the very beginning. The first research constraint consisted of gathering only journal articles and conference proceedings, and the second research constraint required that the papers be published during or after the year 2000. Lastly, the third required that the papers study data quality as a main topic and not as a secondary issue of their research. Examples of papers that could be of interest were ones that discussed different problems caused by dirty data, and papers that developed algorithms or mathematical models to deal with data quality problems. The research was not interested in gathering articles that aimed at quantifying or developing guidelines to measure data quality dimensions.

The main search engines used for this research experience were Compendex and Web of Science, through the University of Arkansas databases. Compendex was primarily used to identify the first set of articles from academic journals and conference proceedings. This resource permitted the search string to be modified by allowing the addition of search terms, and by having those terms search in specific fields. Table 1 lists the main search terms, and their respective search fields, applied to obtain the first set of research articles, and the total number of search results obtained from each combination of search terms. These modifications were used in addition to the three research constraints explained previously in this section.

Table 1. Compendex Search Terms

Search Terms	**Search Fields**	**Search Results**
Data Reduction Data Quality	Main Heading Subject/Title/Abstract	621
Data Reduction Data Quality	Controlled Term Controlled Term	450
Data Quality Errors Quality Management	Subject/Title/Abstract Controlled Term Controlled Term	79

To sort through the extensive Compendex search results, the titles and controlled terms for each of the search results were reviewed, and gradually, the first set of relevant articles was collected, consisting of 120 papers. In addition, the search engine Web of Science served as a useful resource to learn about the number of times certain papers had been cited by other authors, and helpful for finding links to those articles that referenced them. Other relevant papers were also found by looking at the References section of the articles gathered.

From the first set of papers gathered, it was understood that data errors could enter a system in many different ways throughout the data lifecycle from the very beginning in the data entry stage, to data storage in data warehouses, to data sharing across organizational boundaries, to data cleaning, and all the way to data analysis. The next stage of the research consisted of trying to classify the papers gathered according to how they discussed and modeled data quality. This was done by carefully reading the abstract, and if necessary, the introduction and conclusion of each of the papers. Whether the papers discussed potential data errors associated with data entry forms or data corruption risks associated with inter-organizational data exchanges, this stage in the research process tried to identify general groupings that could be used to categorize the possible ways in which the available knowledge on data quality might be studied.

The way in which data quality errors could enter a system was used to develop the first tier of categorization for this research. Based on the papers gathered, three general areas associated with data lifecycle stages were identified. The first category is “data entry”, and is concerned with data quality issues when data is first created and enters a system through different means, such as data entry forms or Excel spreadsheets. The next general grouping is “data flow”, and is concerned with data quality issues caused by data sharing within and across organizational boundaries. Lastly, the third grouping is “data maintenance”, and is concerned with activities directed towards attaining a desired level of data quality in data warehouses, whether by cleaning data before it could be integrated into an enterprise-wide data warehouse, or when data is already stored in a data warehouse, or through the improved design of a database. Having identified these three groupings helped to screen out papers outside of the focus areas of this research project, and boil down the initial set of papers from one hundred and twenty to approximately thirty articles. Although the existing academic literature studies other aspects of data quality, based on the representative set of relevant papers, it was concluded that these were the general ways that we were interested in studying how the quality of data is used in business processes. As this research experience progressed, these three categories were the focus as additional papers were uncovered and collected.

The next step in the research consisted of reviewing more in depth each one of the papers that had already been gathered, and further reduce the number of articles from thirty to twenty two. This was done by rereading the abstract, introduction, background and conclusion of each one of the gathered papers, and summarizing them around three basic points. The goal of the first point was to recognize the data lifecycle stages that each particular paper was focused on. The second step sought to identify the main issue that each article was trying to solve. Lastly, the intention of the third step was to identify how each research paper added value to the data quality research field, and pinpoint their recommendations for dealing with data quality issues. Moving forward, this was the basis for which newly gathered conference papers or journal articles were compared against.

This structured approach helped to systematically categorize each one of the gathered papers as belonging to one, two, or three of the general groupings. This method also helped to distinguish the papers that were more pertinent to our research scope, and reduce the set of research papers to approximately twenty two articles. Figure 1 uses a Venn Diagram to describe how these twenty two research papers were classified according to the three general groupings: “data entry”, “data flow“, and “data maintenance”. Two papers were associated with “data entry”, six papers with “data flow”, eight papers with “data maintenance”, three papers with “data flow” and “data maintenance”, and lastly, three papers were associated with all three categories.

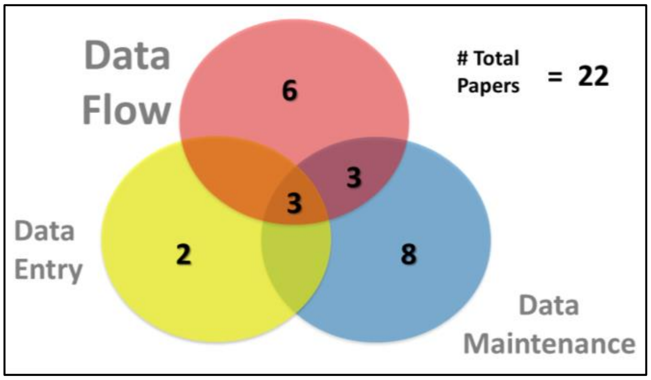

Figure 1. Venn Diagram of General Categories

The next stage of the research focused on performing a more in-depth review of each one of the twenty two articles collected, to better understand the recent work that had been performed in each one of the three general categories. During the next stage of the research project, the articles were subcategorized based on their proposed solutions for dealing with data quality problems, and on the research methods used, for which earlier work done by Madnick and Wang (2009), titled “Overview and Framework for Data and Information Quality Research” provided a solid foundation.

3 Literature Review

Companies and organizations all around the world are gathering massive amounts of data about their customers, suppliers and operations, and this is shaping the way businesses make decisions and operate today. According to the Computer Sciences Corporation (CSC) (2012), the production of data is growing so rapidly that by the year 2020, the amount of data produced will be 44 times larger than what it was in 2009. CSC (2012) also explains that together with a rapid increase in the amount of data created, the data revolution is also bringing about a change in the way data is produced, processed and analyzed. As data continues to evolve, this field of research will continue to be a rich environment for future challenges. Right now, companies and organizations are interested in translating data into meaningful insights that can be used to improve business processes, and make important strategic decisions, (Madnick & Wang, 2009; Ransbotham, 2015). However, dirty data can lead to incorrect decision-making, legal penalties and inefficient operations. Chiang (2008) described that dirty data costs businesses in the United States above $600 billion a year. In 2012, Bai, Nunez, and Kalagnanam stated that the Institute of Medicine reported 1.5 million people injured by medication errors, costing billions of dollars annually. Additionally, according to Ransbotham (2015) (35), because of issues surrounding the quality of their data, companies and organizations are not able to take advantage of emerging opportunities before their competitors. From these examples it can be seen that, in the past few decades, there has been both an increased awareness of data and information quality issues. For these reasons, it is important to understand how the available knowledge on the topic is trying to solve many data and information quality problems (Madnick & Wang, 2009).

As explained above in the Methodology portion of this honors thesis, the knowledge on data quality encompasses a wide variety of topics. The goal of this research effort is to gain a holistic view of this broad area of study, and identify categories that could describe the different ways in which new models and solutions could be utilized to improve the quality and effectiveness of data-driven decisions in large organizations. This research focused on “data entry”, “data flow”, and “data maintenance”, and the following portion of this paper provides a brief overview of the reviewed articles that were deemed most relevant, and identifies the research methods employed in each article based on the framework for data and information quality research developed by Madnick and Wang (2009).

3.1 Data Entry

The first general category, referred to as "data entry", was concerned with the quality of data as it first enters a system. As can be seen in Figure 1, of all the twenty two research papers gathered, only two papers were characterized as belonging in this group. In 2011, Chen, Chen, Conway, Hellerstein, and Parikh conducted research motivated by the "United Nations Development Program's Millennium Villages Project" (p. 1152) in Uganda, and a community health care program in Tanzania. According to the authors, there is little research into automatic methods for improving the quality of data at the moment of data entry, even though, according to the authors, data entry forms represent the best opportunity for detecting and mitigating data errors. The study proposed a system for form design, data entry, and data quality assurance called "Usher" that applies statistical analysis techniques at the moment of data entry to improve the quality of a data. Figure 2 provides an illustration developed by the authors that graphically represents the flow of data through Usher and its system components.

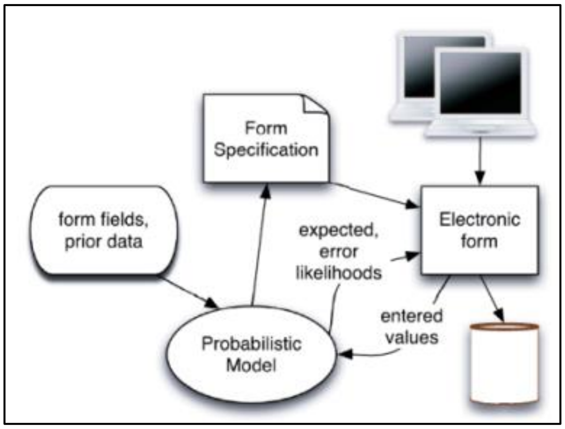

Figure 2. Components of Usher (Chen et al., 2011)

Chen et al. (2011) described how Usher is able to learn a probabilistic model from existing data over the questions of the form, and is able to dynamically adapt forms to values being entered by the user during data entry, and provide immediate feedback to the user by repeating questions with “dubious” responses, or by reformulating them. By doing this, the authors described how Usher could potentially improve the quality of data at every step of the data entry process at a reduced cost, and improve the speed of entry.

3.2 Data Flow

The second general category, referred to as “data flow”, is concerned with the quality of data as it is shared within different areas of an organization, or as it flows across organizational boundaries. As can be seen in Figure 1, of the twenty two research papers gathered, a total of twelve articles were characterized as belonging in this category, with six corresponding only to “data flow”, three overlapping with “data maintenance” and three others overlapping with both “data entry” and “ data maintenance”. Figure 3 shows a Venn Diagram that describes how the papers associated with “data flow” were also sub-categorized based on their topics and research methods. The general topics of concern in this area were data networks in inter-organizational data exchanges, the cost and benefit of implementing control policies to improve the quality of data, and its impact on the overall performance of a company or organization. These sub-categories were identified based on the framework for data and information quality research developed by Madnick and Wang (2009).

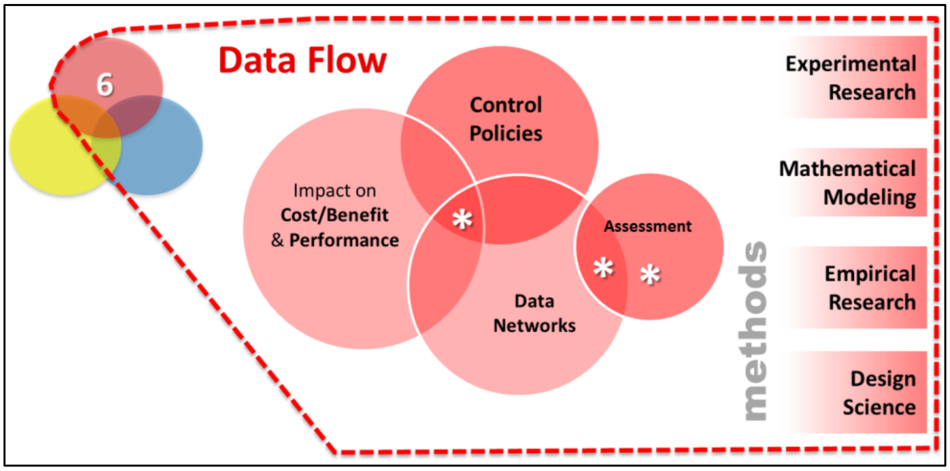

Figure 3. Data Flow Venn Diagram

In 2005, Krishnan, James, Padman, and Kaplan addressed the issue of reliability of data in organizations, motivated by a field study of a major international accounting firm, and the legislative requirements of the Sarbanes-Oxley Act, which requires that a company's CEO and CFO certify the reliability of the data reported in financial statements, as well as the reliability of the information system that produced that data. This research addressed the practical issue of data quality assurance by auditors, and developed an interdisciplinary approach for data reliability assessment using the components, interrelationships and functions of an accounting information system to help an auditor improve the efficiency and effectiveness of reliability assessments. To achieve this, Krishnan et al. (2005) developed a basic ontological structure for modeling information flows in accounting information systems, and an optimization problem that seeks to determine the smallest set of controls that an auditor has to test to assess the presence of data errors in the accounts of an accounting information system. The contributions provided by Krishnan et al. (2005) laid the foundation for other research projects on the topic of data quality and data reliability.

Bai, Nunez, and Kalagnanam (2012) adopted the "process-oriented ontology" (p. 455) for an accounting information system developed by Krishnan et al. (2005). Similarly to Krishnan et al. (2005), this paper was also motivated by the Sarbanes-Oxley Act of 2002, addressed the issue of data quality management in accounting information systems, and proposed a methodology for managing the data quality risks in accounting information systems. The authors modeled the error evolution process in transactional data flow, and found an optimal control policy "at the task level" to mitigate data quality related risks. Figure 4 provides a graphical representation of a business process, from the transaction source (or the origination point in the process), up to the information repositories or "audit targets" located at the end of the process, where all the sub-processes in between represent a potential error source.

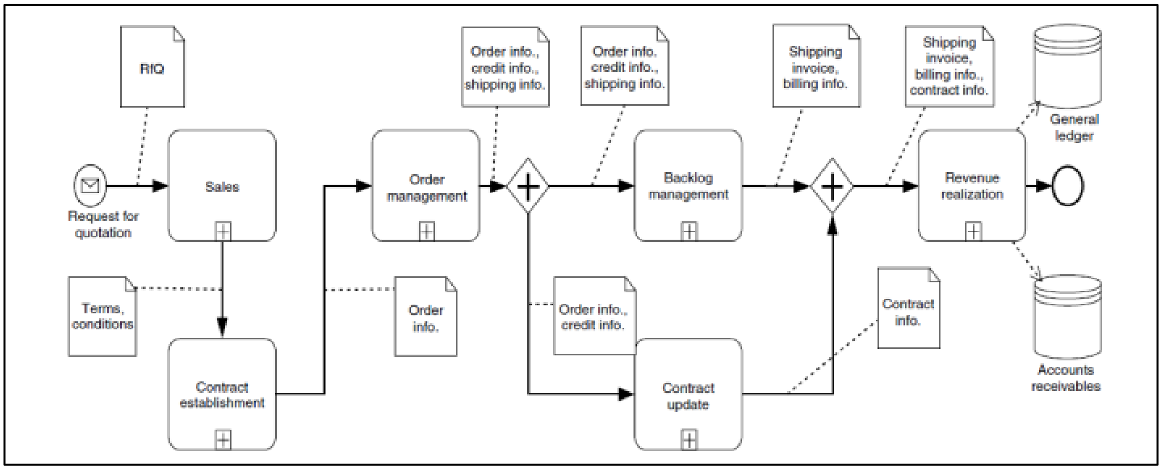

Figure 4. "The High Level Modeling Diagram of the Revenue Realization Process"

(Bai et al., 2012, p. 457)

Bai et al. (2012) proposed a constrained Markov decision model (CMDP) to model the control process for mitigating data quality related risks (or data errors). More specifically, the authors used a CMDP to determine the optimal control policy that minimizes the cost of implementing controls, while keeping risks below a threshold. In other words, the model determines the smallest set of controls that need to be tested to assess the absence or presence of target error classes in the accounts of an accounting information system. They also formulated a linear program to find an optimal “stationary” policy for the CMDP. The model captures the error introduction and propagation dynamics in the transactional data flow in business processes and, according to the authors, this research paper represented a first attempt to address data quality risks at the “root causes” (p. 454).

Bai (2012) addressed the data quality management issue in enterprise information systems. Like the research papers mentioned above, Bai (2012) was also motivated by the legislative requirements of the Sarbanes-Oxley Act of 2002, and also adopted the “basic ontological structure” (p. 649) developed by Krishnan et al. (2005) to model the information flow of an accounting information system, using a revenue realization process. The goal of the author was to design cost-effective strategies for placing controls at error sources, with the intention of eliminating errors in the flow of information and from reaching data repositories, and improve the quality and reliability of key performance indicators (KPIs) of an enterprise. Bai (2012) took the work developed by Krishnan et al. (2005) a step further, by including quantitative attributes about the flow of transactions, about the probabilities of errors at tasks, and about the effectiveness of controls for detecting and correcting data quality errors. For this, the author developed a two-stage multiple choice knapsack model” that seeks to minimize the total cost of placing controls while keeping the probability of errors below a certain threshold. The first stage solves a “group of local problems”, and the second stage solves a “global problem”. The objective of the first stage is to find the cost matrix for a given data quality configuration, where every entry in the cost matrix involves solving a “local problem” for an error source and a “pre-specified” error probability level. This total cost matrix holds the minimum total cost values incurred at every error source after applying a series of control units. The dimensions of the cost matrix then consist of the minimum costs resulting from the control selection decisions applied at each one of the error sources to achieve a desired level of error probability. The output of the first stage is then used as in input in the second stage, where the objective is to find the optimal control configuration in the overall process, such that a specific threshold risk is achieved at minimum control cost.

Nikolaou, Ibrahim and Heck (2012), examined how information quality can influence an information system design. It focused on the issue of data quality, but more specifically, on the data flow occurring in inter-organizational electronic information and data exchanges. The author studied how the quality of data and information exchanged can influence the perceived trustworthiness of the "exchange partner" (p. 986), and potentially impact the user's intent to continue the electronic data exchange. Nikolaou et al. (2012) based its conclusions on results obtained from a Survey to 221 business professionals. This paper does not provide a math model, algorithm or solution; nevertheless, it provides many useful insights, and puts a spotlight on current data quality issues.

The last paper categorized as belonging in the "data quality" grouping is Bai, Krishnan, Padman and Wang (2013), which brings together several of the researchers mentioned earlier in this paper. In this paper, the authors adopted the "process-oriented ontology" (p. 731) developed in Krishnan (2005), and developed a "process modeling-based methodology" (p. 730) for managing data errors in information flows. Similarly to Bai (2012), the authors included quantitative attributes about the flow of information and errors at the tasks, and about the effectiveness of controls for detecting those errors. To illustrate the methodology, the authors used the order fulfillment process of an online pharmacy, and used "Petri nets" (p. 732) to model the business process. To evaluate the propagation of errors in a business process, the model considers the probability that a particular information element could reach other nodes from each node. Also, the authors assumed that the effectiveness of control procedures, or their ability to detect errors, would depend on the level of "control resource" (p. 737) applied at a node, and on the location of the node where a control is applied. According to the authors, the methodology focuses on the impact that the structure of a business process can have on the propagation and mitigation of data errors, and can help to design optimal "control resource allocation" strategies (p. 746) to help manage data quality risks.

3.3 Data Maintenance

As mentioned earlier, the third category, "data maintenance", was defined as being concerned with activities directed towards attaining a desired level of data quality in data warehouses, whether by cleaning data before it could be integrated into an enterprise-wide data warehouse, or through the improved design of a database. Figure 5 shows a Venn Diagram that describes how the papers associated with "data maintenance" were also sub-categorized based on their topics and research methods. According to the relevant set of papers gathered, the general topics of concern in this area were dealing with uncertainty in data, the cost and benefit of monitoring and cleaning data in data repositories, database design to achieve improved levels of data quality, and data integration activities. Like "data entry" and "data flow", the sub-categories were based on the framework for data and information quality research developed by Madnick and Wang (2009).

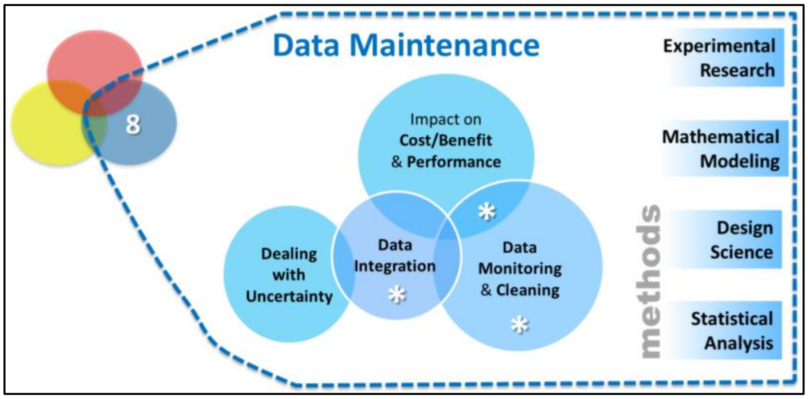

Figure 5. Data Maintenance Venn Diagram

As can be seen in Figure 5, a total of eight papers were associated with the "data maintenance" grouping. Even and Shankaranarayanan (2007) examined the economic challenges of maintaining data within organizational data repositories at high quality. The authors studied if the increased costs of maintain good quality data are worth the business benefits earned. According to the authors, the data quality considerations are insufficiently addressed in the design of large repositories and that even though it is a challenge to maintain data within organizational repositories at high quality levels, the business benefits of high quality data have "rarely been examined or quantified" (p. 23), and this research served to address that opportunity. They assessed the tradeoffs between utility/benefits and costs when trying to optimize the economic performance of large data repositories. The study proposed a mathematical model that helps to describe the economic effect of different data quality configurations in data repositories. The model developed by the authors views data quality measures as a design target, and focuses on the quality dimensions completeness and accuracy. The authors explained that modelling the economic effects of different data quality configurations can affect the design decisions of data-warehouse or data-repositories, and that these data quality considerations should be an integral part of the design process. With this model, the research proposed a more proactive approach, rather than reactive approach, to dealing with data quality issues.

Even and Shankaranarayanan (2007) identified two categories of data quality dimensions that are based on how design decisions impact the quality of data. The first is “deliberate choices” (p. 26), which refers to quality levels that are dependent on design choices. They defined the data quality dimension “completeness” (p. 23), represented as “R”, to examine its effect. The second is “random effects” (p. 26), which refers to data quality levels that depend on random factors/events, and not on design decisions. The authors defined the quality dimension “accuracy”, represented by “A”, to examine its effect. From those two data quality dimensions, the authors defined the two-dimensional vector $Q = (R, A)$ to evaluate data quality. Figure 6, and Equation 1 describe the tradeoffs and economic effects that design decisions related to completeness can have on utility, cost and benefit.

$$B(R) = U^{M} R^{\alpha} - (C^{F} + C^{V} R), \quad (\text{s.t. } 0 \le r \le 1).$$

Equation 1. (Even & Shankaranarayanan, 2007, p. 29)

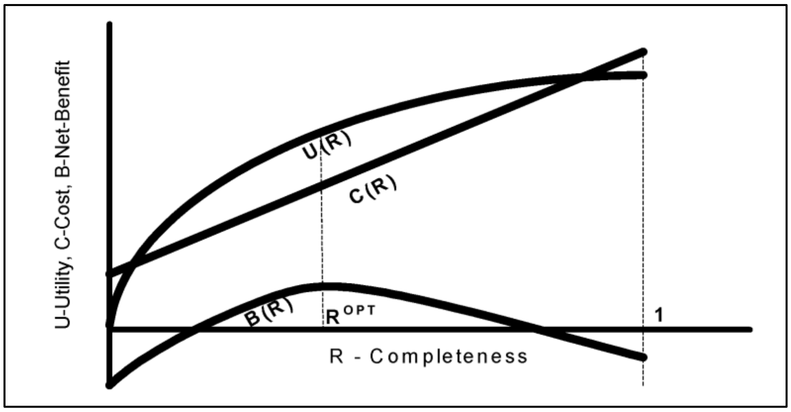

Figure 6. (Even & Shankaranarayanan, 2007, p. 28)

Figure 7 and Equation 2 describe the tradeoffs and economic effects that accuracy can have on utility, cost and benefit.

$$B(A) = U^{M} A - C^{0} \left(\frac{A}{A^{0}} \right)^{\beta} \quad (\text{s.t. } A^{0} \leq A \leq 1).$$

Equation 2. (Even & Shankaranarayanan, 2007, p. 31)

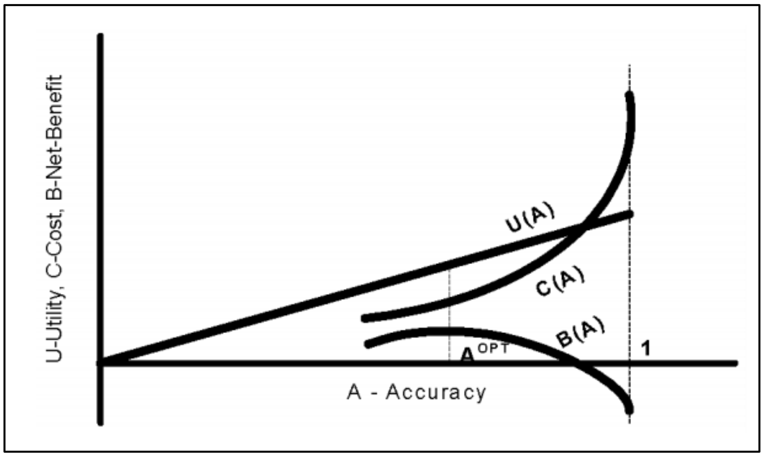

Figure 7. (Even & Shankaranarayanan, 2007, p. 30)

The study also developed an optimization problem that seeks to optimize the benefit for different scenarios of data quality configurations. From the sub-categories identified for "data maintenance" illustrated in Figure 5, this paper mostly focused on evaluating the cost and benefit of attaining desired levels of data quality in data repositories.

Chiang and Miller (2008) studied how database rules that hold over the data can impact the quality of the data that is being entered and stored. The authors describe how many organizations work with consultants to develop new integrity constraints and identify dirty data values, which can be an expensive process that requires people with knowledge of specific business policies and takes up a lot of time. According to the authors, deriving an organization's policies and domain semantics is a primary task towards improving data quality. The research proposes algorithms that can be used by an organization's data quality management process to search for conditional functional dependencies (CFDs) or "functional dependencies that hold only over a portion of the data" (p. 1166), and for "non-conformant records" (p. 1166) or dirty data records. The authors explained how the algorithms can help accelerate the "cleaning process" (p. 1167), and facilitate the work with a consultant looking for potential inconsistencies. The research also presents an experimental study showing the scalability of their techniques.

Along the same lines of Even and Shankaranarayanan (2007), Haug, Zachariassen, and Van Liempd (2011), attempted to describe what it meant to find the optimal data quality maintenance effort in terms of the costs caused as a consequence of poor quality data, and the costs caused from maintaining a certain level of data quality. The study is motivated by how companies are storing increasingly more data and are not paying enough attention to the quality of the data being stored, or do not have an efficient way to manage it. The authors categorized the types of costs that can be caused by poor quality data as "direct versus hidden" and "operational versus strategic" (p. 188), and explained the negative impact that these can have on any business. The authors explained how the art of managing data quality costs should focus on finding the balance between the data maintenance efforts and the costs incurred by poor quality data. Lastly, the authors proposed that the data quality should only be improved by a certain level, balancing the costs incurred from solving problems caused by it, and from working to have it the lowest possible. The idea proposed by this research paper is illustrated in Figure 8.

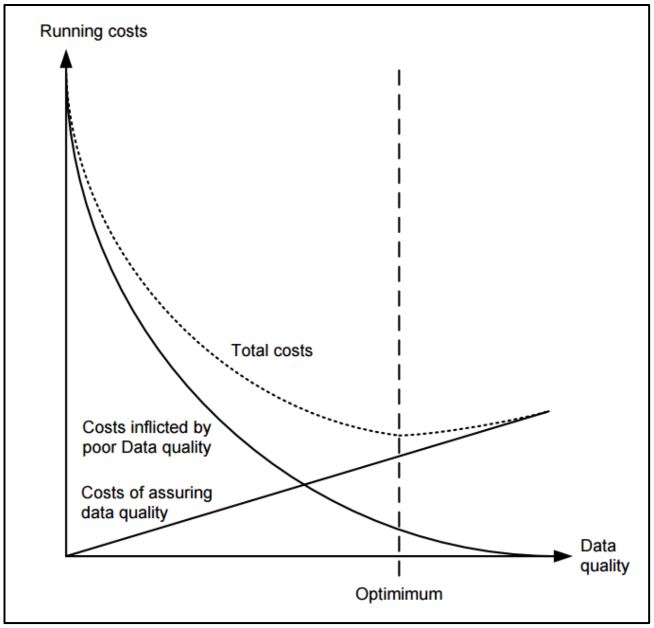

Figure 8. “Total costs incurred by data quality on the company”

(Haug, Zachariassen, & Van Liempd, 2011, p. 179)

Another article associated with the category “data maintenance” is Li and Joshi (2012), where the authors studied the data cleaning aspect of data integration. According to the authors, data integration involves consolidating enterprise-wide transactional data to populate data warehouses, which involves gathering, understanding and cleaning the data coming from different parts of the organization to a certain data quality level before it is saved it in the data warehouses. Nonetheless, the authors described the process of achieving an enterprise-wide single version of the truth as “extremely challenging”, and that a large portion of the extant literature is focused on data quality issues, but there are not many available techniques to help practitioners determine the cost and benefits of performing “data-cleansing approaches” (p. 363). In addition, the authors explained that developing “coherent data cleansing strategies” (p. 361) to improve data quality is very challenging because it requires deep analysis of interacting factors, and it is difficult to justify the “financial and human capital cost” (p. 361) involved in obtaining high data quality. Seeking to alleviate this situation, the author developed an approach that uses discrete-event simulation as a decision tool for making data-cleansing decisions. According to the authors, this can help practitioners systematically determine the tradeoffs between resource costs and “performance outcomes” (p. 361) of different data cleansing approaches. More specifically, the study compares two alternative data cleansing approaches –“Test-Re-Test (TRT) approach” and “Data Profiling (DP) approach”–via a discrete-event modeling decision tool called “ProModel” (simulation software) (Li & Joshi, 2012, p. 372). This study helps understand how financial resources, human capital, and personnel productivity interact in a company’s attempt to attain a desired level of data quality in its data warehouses. According to Li and Joshi (2012), this research contributes to the literature by studying the complex interactions between organizational resources and other factors through discrete-event simulation.

Another article in the “data maintenance” category is Jones-Farmer, Ezell, and Hazen (2014), where the authors provided an overview of the methods for gathering and storing data in large organizations. An illustration of this process can be seen in Figure 9. The authors performed a review of the relevant data quality literature, from the perspective of process improvement, and of the literature on “statistical process control (SPC)” (p. 29) monitoring methods applied to data quality problems. According to the authors, the current academic literature does not provide applicable methods for “measuring, monitoring and improving” data quality, and that right now, there is only “rudimentary” use of SPC methods to monitor and improve data quality. Additionally, the authors explained that data quality dimensions should be operationally defined in order to make them useful for monitoring procedures.

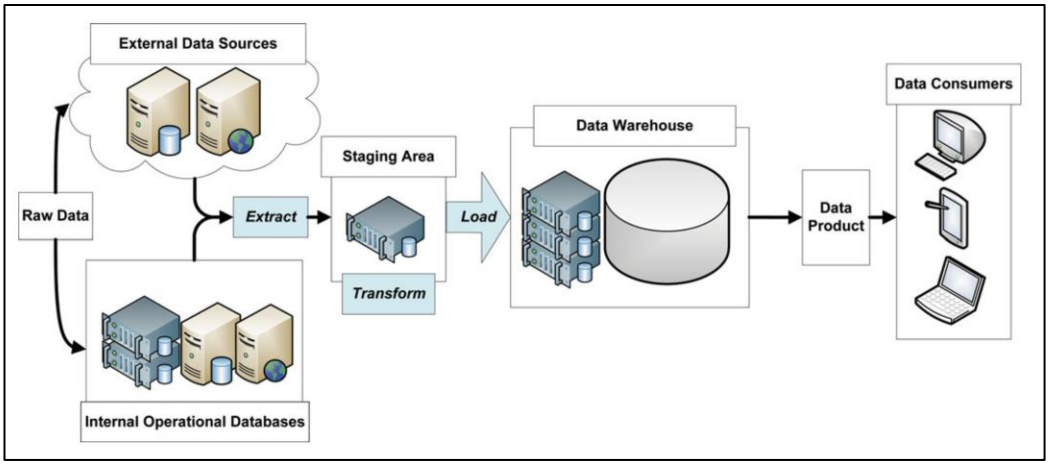

Figure 9. “An example data warehouse/data production process”

(Jones-Farmer, Ezell, & Hazen, 2014, p. 30)

The authors proposed the use of control charts as a method to monitor and improve the quality of data used by organizations to make business decisions. They found similarities between automobile production and data production processes in the way that the final product/result may be measured and monitored, and provided an example of how to measure data quality in a data production process using a real aircraft maintenance database. Lastly, the authors highlight future research opportunities in control charting for data quality.

3.4 Data Flow and Data Maintenance

From the set of twenty two relevant papers, three papers fell under the categories referred to as “data flow” and “data maintenance”. Figure 11 describes how these papers were grouped based on the topics they discussed and the research methods used. The papers in this grouping are concerned with the quality of data as it is shared within different areas of an organization or across organizational boundaries, and with activities related to cleaning data before it can be integrated into enterprise-wide data warehouses.

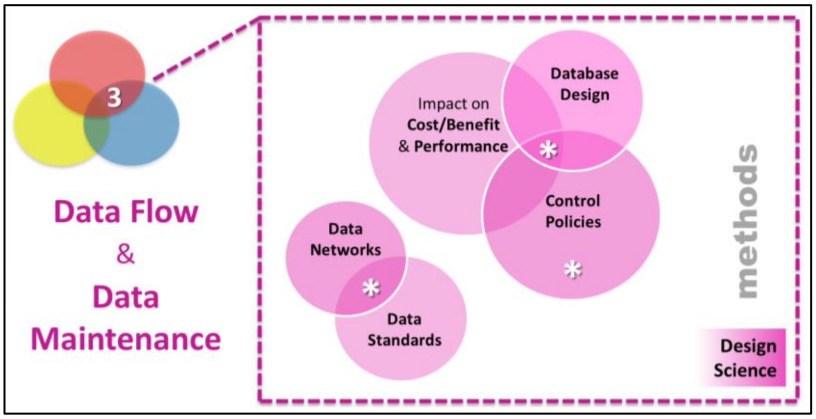

Figure 10. Data Flow and Data Maintenance

Cai and Shankaranarayanan (2007) explained how the sharing and exchanging of data between organizations is a normal part of business relationships –"business operations require routine exchange of data" (p. 255). For this reason, the authors explained how this situation creates the need for a method that can help organizations evaluate the data exchanged coming from other business partners in such a way that they can understand the quality of the data received. The authors stated that data networks for inter-organizational data exchanges are characterized by "multiple, independent data sources from which this data is extracted, and multiple, independent data repositories in which the data is captured/stored" (p. 255). The authors also explained that the main barrier of the data quality management problem in inter-organizational data networks is the lack of a standard for managing data quality, and explained how a solution to successfully managing the data quality across multiple organizations are the adoption of data quality standards, and the commitment to manage the quality of data internally. The authors also highlighted data quality management requirements for inter-organizational data networks, and described the metadata specifications for the exchange of data. The authors proposed “DQ 9000” as a data quality standard based on the “information product approach”, and developed an “XML-based data quality specification” for “Information Product metadata (IP-XML)” as a cost-efficient solution to deal with the data quality management issue in inter-organizational situations (Cai & Shankaranarayanan, 2007, p. 256).

Even and Shankaranarayanan (2009) focused on the “cost-benefit” (p. 127) tradeoffs associated with managing data quality. According to the authors, the objective of this paper was to highlight the economic factors that impact data management decisions. The authors proposed a “sequential data processing model” (p. 129), illustrated in Figure 11, that models the cost-utility tradeoffs (at each possible stage) of a process that manufactures data products. The authors were only interested in fixing data quality errors, and were not concerned with what caused them.

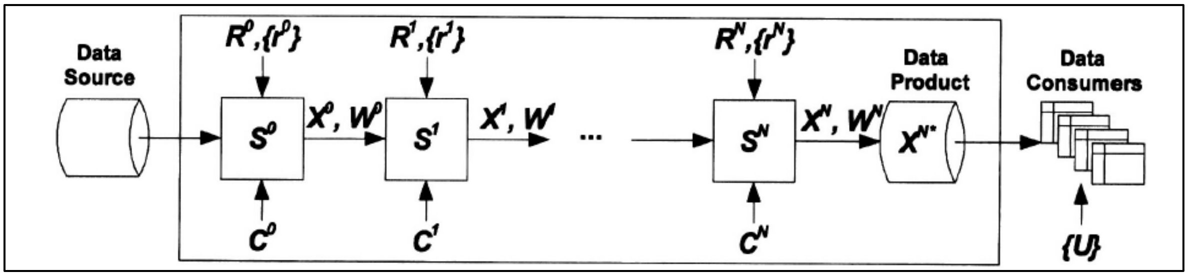

Figure 11. (Even & Shankaranarayanan, 2009, p. 129)

This model assumes that data quality errors –“random data quality defects” –occur at the different processing stages, and uses a “metadata vector” to represent the characteristics of the dataset at each processing stage. This model seeks to maximize the overall benefit from selecting the optimal set of “transformations” (or error correction techniques) at each stage from a “feasible set of transformations”. It models the transformations that the data set undergoes at each processing stage as changes to the “metadata vector” associated with the input dataset that enters a specific stage, and to the “metadata vector” associated with the dataset that leaves that stage. A few examples of the type of transformations that the research refers to are: “manual intervention”, “automated processes”, and leaving the data “as is”. According to this model, the total variable cost fixing data quality errors is affected by the transformation chosen, because different transformations have different cost. The model assumes that data quality defects at a certain stage occur or are identified after the dataset has been processes at that stage, and before it is transferred to the next stage. To illustrate the model, the study develops a data quality management optimal policy for online error correction in a data manufacturing system. The results of the study indicate that “decisions that consider economic tradeoffs can be very different compared with decisions that are driven by technical and functional requirements only”.

4 Repairable Systems Models

The relevant set of research papers described in the Literature Review portion of this thesis served to learn about different ways to address data quality issues; however, the opportunity to relate classical maintenance models to data maintenance efforts is investigated. Just as data quality problems can cause a company to incur additional and significant expenses, Mehdi, Sai Hong, Ismail, and Mohammadreza (2011) described how in a production facility, the equipment maintenance optimization process is also of critical importance because of the potential costs caused by a system failure event during times of operation. For this reason, the next step of this research consisted of referencing additional research papers and other academic material on the topic of equipment maintenance optimization models with the objective of identifying the basic concepts that could be adapted to data maintenance activities. For the remainder to this paper, the words system, machine and equipment will be used interchangeably.

Maintenance actions are value added activities performed because they serve to prolong the lifetime, reduce the probability of an unpredictable failure, or increase the availability of equipment or repairable systems that can fail after an unpredictable time (Sharma, Yadava, Deshmukh, 2011) (Springer, 2013). Yadava and Deshmukh (2011) and Mehdi, Sai Hong, Ismail, and Mohammadreza (2011) explained that there are usually three types of maintenance actions called “preventative maintenance”, “corrective maintenance”, and “predictive maintenance”. “Preventative maintenance” refers to planned and periodic activities performed on an operating system based on a specific schedule with the objective of keeping equipment in working conditions. “Corrective maintenance” refers to activities performed as a result of an equipment break down, and intends to return the equipment to a specified working condition. “Predictive maintenance” refers to maintenance activities that occur based on the condition diagnosis of “modern measurement and signal processing methods”. According to the authors Meselhy and ElMaraghy (2009), the structured combination of these maintenance actions forms the “maintenance strategy”.

There exist different maintenance optimization models to approach maintenance problems (Springer, 2013; Yadava, Deshmukh, 2011). For the purposes of this study, various basic maintenance optimization models were reviewed, including two types of basic replacement models known as the “age replacement policy”, and the “block replacement policy”. These models consider a technical system with lifetime described by a positive random variable T and a distribution F that is immediately replaced upon failure by an equivalent system, and incurs a cost of $c > 0$ for each replacement and a “penalty cost” of $k > 0$ for each system failure (Springer, 2013). The “age replacement policy” considers a fixed system replacement age, specified by a positive constant, and seeks to minimize the cost per unit time over successive system lifetimes. The average cost after n cycles is calculated by getting the sum of all successive cycle costs divided by the operating time of successive system lifetimes. The model also considers the total cost per unit by dividing the sum of all successive cycle costs by the time passed up until time t (or the current time). According to this policy, the system is replaced at failures, or at the fixed system replacement age, depending on which one occurs first. In contrast, the "block replacement policy" involves preventative replacements performed at fixed time periods –times which the policy seeks to determine in advance –without regard for the age of the system components, and at the cost of c. The system is also replaced at system failures that occur within the fixed time periods at a cost of $c + k$. According to the authors, the block replacement policy is easier to administer because the system / system component replacements occurs at scheduled times. However, the "age replacement policy" is considered to be more flexible because it takes into account the age of the system. A more detailed description of both policies is contained in Springer (2013).

Maintenance optimization problems rely on the underlying deterioration process, and the failure rate behavior of the equipment or system under study. According to the authors, the performance of a system can be quantitatively analyzed through reliability. Cassady and Nachlas (2008) defined reliability as "the probability that a device properly performs its intended function over a specified period of time". According to Springer (2013), a reliability study can have multiple objectives that may require different approaches and reliability modeling methods. The author described that models can seek to minimize the equipment maintenance cost subject to reliability requirements, maximize the equipment reliability subject to maintenance cost constraints, or determine the best time to execute maintenance activities. In addition, the author described that maintenance optimization problems must also consider the cost per unit of performing preventative and corrective maintenance, the cost of repairable system replacement, and the cost per unit of downtime in case of system failure. Additionally, it explained that in order to start the modeling process, the study must first consider certain characteristics of the situation in question. For instance, the study must decide on how to represent the system and its components. Also, the study must consider if "preventative maintenance" or "corrective maintenance" actions will be performed on the system. In addition, the study must consider if the overall state of the system or of each one of the system components will be continuously monitored, or if they will only be known after inspection at discrete points in time. In the same manner, the study must determine what information will be available about the system and its components that could help to reveal evidence about its underlying condition. Along the same lines, it is also important to reflect on how to model the deterioration process of the system and its components, evaluate if the state of the system at any given point in time can be described by a binary or multi-state model, and, decide on how to model the times for maintenance actions. Lastly, the study must think about what the state of the system will be after a repair activity, and what the available amount of resources will be for performing those repairs (Springer, 2013).

Cassady, Mohammed, Scnneider, and Pohl (2005) stated that the academic literature has vastly studied the use of mathematical modeling for evaluating improving, and optimizing the performance of repairable systems. Cassady et al. (2005) also explained that, with regards to the effect of repair/maintenance actions on repairable systems, these studies have made assumptions which can be classified into "minimal repair", "perfect repair", and "general repair" or "imperfect repair", and that the majority of models assume "perfect repair" and "minimal repair". After a "perfect repair", a repairable system can be treated as "good as new", or the repair can be considered a complete replacement of the system. In practice, according to the authors, a “perfect repair” can be assumed for a structurally simple system, or for a system consisting of just one component. After a “minimal repair”, equipment can be considered to be “bad as old”, or as having the same age it had before the failure. Both perfect and minimal repair can be considered extreme maintenance actions. However, according to Cassady et al. (2005), mathematical maintenance models can also assume general or imperfect repairs, which is a concept that assumes that a repair is somewhere between the two extremes of “perfect repair” and “minimal repair”. Part of the imperfect/general repair models are the “virtual age” models. Virtual age is a concept introduced by Kijima (1989) that represents the age of the equipment, different to the actual total operating time of the equipment, and is a function of the total number of repairs and the operation time of a system.

The Kijima models consider a system that at any point in time can either be functioning or failed, but assume that at time $t = 0$ the equipment is functioning (Cassady et al., 2005). The models also use the variable X_n to represent the duration of time of equipment functioning between the $(n-1)^{th}$ repair completion and the n^{th} system failure. In addition, the models use the variable V_n to represent the “virtual age” of the system at the time of completion of the n^{th} repair, and the variable V_{n-1} to represent the accumulated virtual age of the system after the $(n-1)^{th}$ repair. The Kijima models assume that age is accumulated through each time period of equipment functioning, X_1, X_2…, X_n, as illustrated in Figure 12. In practical terms, the accumulated age can represent any type of wear out or damages incurred during the past period of operation. The Kijima models can be classified in two depending on how the repairs affect the virtual age of the system under study. These two classifications consist of the Kijima Type I model and the Kijima Type II model.

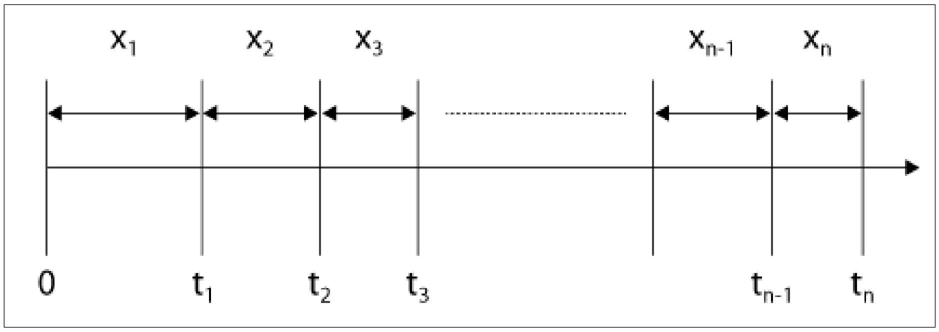

Figure 12. ReliaWiki.org (46)

The Kijima Model I assumes that the n^{th} repair of a system can remove a portion of the system's accumulated age during the past period of operation, between the $(n-1)^{th}$ and n^{th} failures, but is incapable of eliminating the damage accumulated up until $(n-1)^{th}$ failure. The first Kijima model is described by the author as

$$V_n = V_{n-1} + \alpha \cdot X_n$$

Equation 3. (Cassady et al., 2005, p. 565)

where α, is a constant value such that $0 \leq \alpha \leq 1$, and where $(1-\alpha)$ represents the "degree of restoration" after a repair, or in other words, the portion of the accumulated age that has been removed from the virtual age of the system after the completion of the n^{th} repair. Figure 13, a graphic provided by Cassady et al. (2005), describes the effects of the repairs on the virtual age of a machine that undergoes "imperfect repairs" after system failures under the Kijima Type I model.

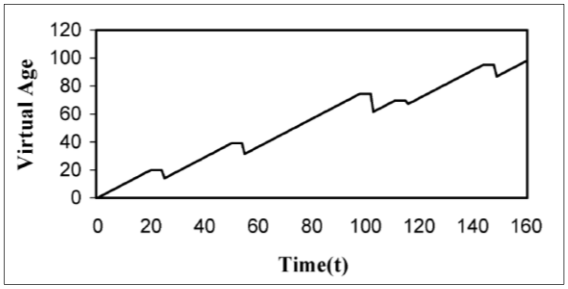

Figure 13. (Cassady et al., 2005, p. 565)

To further describe the Kijima Type I model, the authors use $F_1(x)$ to denote the life distribution of a new machine or the cumulative distribution function of X_1. The authors also use $F_n(x \mid y)$ to describe the conditional cumulative distribution function of X_n. It is represented as

$$F_n(x \mid y) = \frac{F_1(x+y)-F_1(y)}{1-F_1(y)},$$

Equation 4. (Cassady et al., 2005, p. 565)

and through it, the Kijima Type I model describes how the duration of equipment functioning after completing the last repair depends on the virtual age of the equipment at the beginning of the interval of operation. This model is described more in detail in the research paper by Cassady, Mohammed, Scnneider, Pohl (2005) (45). This model was also expanded by modeling the virtual age of the equipment after the n^{th} repair as

$$V_n = V_{n-1} + A_n \cdot X_n,$$

Equation 5. (Cassady et al., 2005, p. 566)

where A_n can take any value in the set $\{A_1, A_2, \dots\}$ representing independent random variables over the interval of real numbers $[0,1]$. In contrast to the Kijima Type I model, the Kijima Type II model assumes that the n^{th} repair will remove a portion of the total accumulated system age, or system wear out or damages incurred over time until the system failure. This model was presented by the authors as

$$V_n = A_n \cdot (V_{n-1} + X_n).$$

Equation 6. (Cassady et al., 2005, p. 566)

These two models assume that the maintenance activities are executed in a negligible amount of time, and that the repairs do not change the system’s underlying failure distribution (Guo, Liao, Zhao, and Mettas, 2007). The Kijima models were of interest to this research project focused on data quality because they allow for the system under study to operate outside of the system statuses caused by perfect and minimal repairs, given that it is very difficult to operate at a "good as new" level unless all system/system components are replaced every single time.

In (Cassady et al. (2005), the authors studied the impact of the general/imperfect repairs on the availability of a repairable system using the Kijima Type I model that is required to operate twenty four hours a day, seven days week for a certain useful life value represented by the constant L. The authors considered the system as a single unit, and assumed that no preventative maintenance activities were to be performed on the system. In addition, they assumed that at any point in time, the system could only have one of two possible states –"functioning or failed" (p. 565) , and that at the time of deployment, it was fully functional. The state of the equipment is described by $Y(t)$, where $Y(t) = 0$ means that the system has failed at time t, $Y(t) = 1$ means that the system is functioning at time, and when $t = L$, the system has reached the end of its useful life. Following the structure presented by the Kijima Type I model, the authors developed a model that considers similar variables, such as X_n representing the time between the last repair activity and the system failure. To model the cumulative distribution of X_n, the authors used a Weibul distribution with shape and scale parameters, β and η, respectively. In the same manner, the authors also represented the cumulative distribution function for X_1 as

$$F_1(x) = 1 - exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right]$$

Equation 7. (Cassady et al., 2005, p. 566)

and a conditional cumulative distribution function for X_n as

$$F_n\left(x \,\middle|\, y\right) = 1 - exp\left[-\left(\frac{x+y}{\eta}\right)^{\beta} + \left(\frac{y}{\eta}\right)^{\beta}\right].$$

Equation 8. (Cassady et al., 2005, p. 566)

In addition, the authors considered the time required to complete a repair activity to be constant value represented by t_r. Considering all the above variables, the authors defined the equipment availability function as

$$A(t) = Pr[Y(t) = 1],$$

Equation 9. (Cassady et al., 2005, p. 566)

were $A(t)$ is the probability that the equipment is functioning at time t. The authors also performed additional studies to develop a cost based "near-optimal" (p. 568) replacement time for an equipment that considers the equipment acquisition cost, represented as c_a; the cost per unit downtime, represented as c_d; the replacement time for the equipment, represented as τ; and the average availability over the first τ time units of the equipment, represented as $A_{avg}(\tau)$. The authors determined the "near-optimal" or recommended replacement time by finding the minimum average cost per unit time of equipment functioning through the equation

$$AvgCost(\tau) = \frac{c_a}{\tau} + c_d \cdot \left[1 - A_{avg}(\tau)\right],$$

Equation 10. (Cassady et al., 2005, p. 568)

together with the use of a standard mathematical software package. The authors stated that there is an opportunity for future research to expand this model to consider the availability of a system when maintenance activities are performed at the component level.

The research performed by Cassady et al. (2005) on the impact of imperfect maintenance on the availability of repairable equipment, together with the concepts introduced by the Kijima models, provide a useful modeling framework for modeling maintenance efforts. As a start, consider a system consisting of an information instance. In a similar manner as the Kijima models, maintenance actions could be performed only at system failures, that is, only through corrective maintenance activities when the information instance becomes outdated, and not through preventative maintenance actions. In addition, if the study performed by Cassady et al. (2005) was expanded to consider maintenance activities at the component level of a system, the study could be used to evaluate the system of an information instance formed by a number of data elements, were the accumulated “virtual age” of a system could be reduced by performing repairs at the component level (in other words, on certain data elements). Additionally, the cost per unit downtime, c_d, could be translated into the cost of having obsolete information based on outdated data elements, and the cost of equipment acquisition, c_d, could be translated into the cost of collecting the initial set of data elements that form the information instance. Along the same line, the replacement time for the system can be considered to be the proper time for updating the data elements forming the instance of information.

Conclusion

As mentioned in earlier portions of this thesis, the quality of data is an important issue today, and given what has been learned from the academic literature on data quality and on equipment maintenance models, there is an opportunity to expand and adapt the concepts used in equipment maintenance processes to prolong the useful life of repairable systems to the research on data maintenance mathematical models.

References

Bai, X. (2012). A Mathematical Framework for Data Quality Management in Enterprise Systems. *Informs Journal on Computing, 24*(4), 648-664. Retrieved from http://dx.doi.org/10.1287/ijoc.1110.0475

Bai, X., Krishnan, R., & Wang, H. J. (2013). On Risk Management with Information Flows in Business Process. *Information Systems Research*, 731-749. Retrieved from http://dx.doi.org/10.1287/isre.1120.0450

Bai, X., Nunez, M., & Kalagnanam, J. R. (2012). Managing Data Quality Risk in Accounting Information Systems. *Information Systems Research, 23*(2), 453-473. Retrieved from http://dx.doi.org/10.1287/isre.1110.0371

Cai, Y., & Shankaranarayanan, G. (2007). Managing data quality in inter-organisational data networks. *Int. J. Information Quality, 1*(3), 254-271.

Cassady, R. C., & Nachlas, J. A. (2008). *Probability Models in Operations Research.* CRC Press.

Cassady, R. C., Iyoob, I., Schneider, K., & Pohl, E. A. (2005, December). A Generic Model of Equipment Availability Under Imperfect Maintenance. *IEEE Transactions on Reliability, 54*(4), 564-571.

Chen, K., Chen, H., Conway, N., Hellerstein, J. M., & Parikh, T. S. (2011, August). Usher: Improving data quality with dynamic forms. *IEEE Transactions on Knowledge and Data Engineering, 23*(8), 1138-1153.

Chiang, F., & Miller, R. J. (2008). Discovering Data Quality Rules. *Proceedings of the VLDB Endowment, 1*(1), 1166-1177.

Computer Sciences Corporation. (2012, 11 20). *Big data universe beginning to explode*. Retrieved 2015, from CSC: http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode

Data Revolution Group. (n.d.). *What is the 'data revolution'?* Retrieved 11 20, 2015, from Data Revolution Group: http://www.undatarevolution.org/data-revolution/

Even, A., & Shankaranarayanan, G. (2007). Utility-driven configuration of data quality in data repositories. *Int. J. Information Quality, 1*(1), 22-40.

Even, A., & Shankaranarayanan, G. (2009). Utility Cost Perspectives in Data Quality Management. *The Journal of Computer Information Systems*, 127-135.

Haug, A., Zachariassen, F., & Van Liempd, D. (2011, January). The costs of poor data quality. *Journal of Industrial Engineering and Management, 4*(2), 168-193.

Jones-Farmer, A., Ezell, J. D., & Hazen, T. B. (2014). Applying Control Chart Methods to Enhance Data Quality. *Technometrics, 56*(1), 29-41. doi:10.1080/00401706.2013.804437

Krishnan, R., Peters, J., Padman, R., & Kaplan, D. (2005). On Data Reliability Assessment in Accounting Information Systems. *Informs*, *16*(3), 307-328.

Li, Y., & Joshi, K. (2012). Data Cleansing Decisions: Insights from Discrete-Event Simulations of Firm Resources and Data Quality. *Journal of Organizational Computing and Electronic Commerce, 22*(4), 361-393.

Madnick, S. E., & Wang, R. Y. (2009, June). Overview and Framework for Data and Information Quality Research. *ACM Journal of Data and Information Quality, 1*(1), 1-22.

Nicolaou, A. I., Ibrahim, M., & Van Heck, E. (2012). Information quality, trust, and risk perceptions in electronic data exchanges. *Elsevier*, 986-996.

ReliaSoft Corporation. (n.d.). *Imperfect Repairs*. Retrieved 11 20, 2015, from ReliaWiki.org: http://www.reliawiki.org/index.php/Imperfect_Repairs

Storey, V. C., Dewan, R. M., & Freimer, M. (2012). Data quality: Setting organizational policies. *Elsevier*, 434-442.

Terje, A., & Jensen, U. (2013). *Stochastic Models in Reliability*. New York, NY, USA: Springer. Retrieved from http://0-www.ebrary.com.library.uark.edu

Vasili, M., Hong, T. S., Ismail, N., & Vasili, M. (2011). Maintenance optimization models: a review and analysis. *Proceedings of the 2011 International Conference on Industrial Engineering and Operations Management*, 1131-1138.