THE PRODUCT APPROACH TO DATA QUALITY AND FITNESS FOR USE: A FRAMEWORK FOR ANALYSIS

(Research-in-Progress)

M. Pamela Neely
Rochester Institute of Technology
pneely@cob.rit.edu

Abstract: The value of management decisions, the security of our nation, and the very foundations of our business integrity are all dependent on the quality of data and information. However, the quality of the data and information is dependent on how that data or information will be used. This paper proposes a theory of data quality based on the five principles defined by J. M. Juran for product and service quality and extends Wang et al's 1995 framework for data quality research. It then examines the data and information quality literature from journals within the context of this framework.

Key Words: Data Quality, Information Quality, qualitative analysis, Endnotes library, fitness for use

INTRODUCTION

Data and information quality continue to be of concern to the developers and users of information systems. Awareness of the issue as an academic research area and practical industry consideration has been increasing. We now see job postings that include data quality characteristics, college courses focused on data quality and conferences with data and information quality as the central theme. It has been ten years since Wang et al [78] proposed a framework for data quality research based on a product manufacturing approach. In that paper, the literature at the time was identified and analyzed according to seven functions: management responsibilities, operation and assurance costs, research and development, production, distribution, personnel management, and the legal function. This approach to classifying the literature continues to be a valid theoretical foundation. However, increasingly we find that "quality is in the eye of the beholder." Data and information quality are multi-dimensional [16, 22, 26, 47, 64, 68, 74, 79, 85], and we need to consider the context of the data and information in order to identify the quality. In this paper a data quality framework is proposed that extends the five principles defined by J. M. Juran [29, 30] for product and service quality as well as the quality framework developed by Wang et al. Juran first coined the term "fitness for use" that is widely used in data quality literature [4, 64, 70].

This paper is organized as follows. We will first describe Juran's five principles underpinning fitness for use. We will then briefly discuss Wang et al's framework and explain how Juran's principles form the foundation for the new proposed data quality framework and extend the framework that Wang et al proposed in 1995. In the methodology section we will describe the collection of data for the literature review and highlight some interesting trends in publication outlets. Next, a subset of the data (e.g. journal articles) will be evaluated in terms of the proposed framework. Finally, areas for future research will be described based on the findings of the literature review.

FITNESS FOR USE

With regards to product and service quality, J.M. Juran sought a simple, short phrase to define "everything related to quality," including various features of products and services, the over-all value of their design, and the degree to which the deliverables conform to that design. The resulting term, "fitness for use," aims to encompass the innumerable factors that define "quality" and has been generally accepted in both academic and industrial settings. Of course, whether or not the term conforms perfectly to its intended design is a matter for debate; however, for two decades, fitness for use has been the de facto definition of quality [30].

Product and Service Features and Quality Characteristics

Human needs are extremely diverse, and this has led to a corresponding proliferation of product features and quality characteristics. This proliferation extends to multiple disciplines, as in the following examples:

- Technological: hardness, inductance, acidity, etc.
- Psychological: taste, beauty, status, etc.
- Time-oriented : reliability, maintainability, etc.
- Contractual: guarantee provisions, etc.
- Ethical: Courtesy of sales personnel, honesty of service shops, etc.

The concept of "quality characteristics" is as old as the human species (the entire biological world is responsive to the concept.) Moreover, there has been a long-range trend to quantify these relationships between the user and the quality. Technological characteristics, notably properties of materials, were extensively quantified beginning several centuries ago with accelerated growth of instrumentation. The twentieth and twenty first centuries have seen a similar movement to quantify the remaining types of characteristics.

Service industry quality characteristics, while including all of the above sub-species, are dominated by the psychological and ethical. In addition, the service industries generally regard promptness of service as a quality characteristic, whereas manufacturing industries generally do not. Instead, manufacturing companies regard promptness (i.e., timely delivery of products to customers in accordance with a promised date) as a parameter very different from "quality." The distinction is so sharp that there is a separate organization (Materials Management) to set standards for delivery time (schedules), measure performance, and stimulate compliance.

Juran developed a short list of inputs that companies, organizations, and individuals alike can use in determining a product or service's fitness for use. The questions or inputs for consideration are:

- The users of the product or service
- How the users will actually put the product or service to use
- The possibility and probability of any dangers to human safety
- The economic resources of both the producer and the user
- The user's specific determinants of a product or service that is fit for their use

Despite its brevity, this methodology comprises the highest level of the quality taxonomy. In other words, by asking such simple questions as these, a seemingly endless labyrinth of factors, characteristics, and qualities can be generated to describe the fitness for use of nearly everything for which quality is a factor. Once within this top taxonomic level, we can begin to explore how fitness for use relates to data and information quality.

When considering data and information quality rather than product or service quality we must keep in mind the distinct differences between data and other goods. A product or service is exhausted in its use. Data, on the other hand, is not depleted in its use. Multiple users can use the same data elements at the same time and the data is still available for subsequent users within a different context. This characteristic of data is important in the discussion of fitness for use. Let us turn now to the application of Juran's inputs in relation to product and service quality. The discussion of data and information quality fitness for use will be in a subsequent section.

The users of the product or service

Juran acknowledges that different users of the same product will have differing needs. Not only will they have differing needs, but that will bring different experiences to the use of the product. These experiences will cause them to interact with the product or service in a manner that is different from another user. "A technologically sophisticated user may be able to deal successfully with the nonconformance; a consumer may not. A nearby user may have easy access to field service; a distant or foreign user may lack such easy access" [28, page 428]

Experiences with a product may also change how a product or service is perceived and/or used over time. As a product is used, the experience that the consumer gains with the product can provide valuable feedback as to the quality of the product. "Other ultimate users are mostly employees of organizations that buy products to be used by employees. The products are goods and services of every imaginable sort: utility services, equipment, supplies. As repetitive users, the employees (like consumers) become experts in many aspects of product performance, environmental influences, and so on. Their expertise is obviously a valuable source of information about customer needs [30, page 52]"

The same product can be a raw material in one instance and a finished good in another. There will be different users of the product dependent on where in the process the good is. "Processor customers are also users. They employ our product in their processes. In their capacity as users their needs include worker safety, high productivity, low waste, and still other forms of internal goals. The processors then sell their products to their customers, whose needs may be quite different [30, page 48]"

To summarize, when a person evaluates the quality of a product or service they will draw upon their experiences with similar products or services. Some will look for corroborating evidence of quality, such as a Consumer Reports evaluation. They may ask friends for recommendations. They will determine how the product or service is going to be used to accomplish their goals. The users of a product or service will seek out all available evidence as to the quality of the product or service, and then put it into the context of their own experiences and needs.

How the users will actually put the product or service to use

The previous question (i.e. the users of the product or service) will determine who is using the product or service. This question addresses how the product or service will be used. Any given product or service can be used in multiple ways. "For many materials and standard products, the specifications are broad enough to cover a variety of possible uses, and it is not known at the time of manufacture the actual use to which the product will be put. For example, sheet steel may be cut up to serve as decorative plates or as structural members; a television receiver may be stationed at a comfortable range or at an extreme range; chemical intermediates may be employed in numerous formulas." [28, page 428]

As an illustration, Juran explains how a single purchase order may be used by the purchasing company and the supplier company. Although each company has the similar departments (Production, Quality Control, and Accounts Payable or Material Control), the purchase order is used differently by each company [30, page 7]:

USER DEPARTMENT	PURCHASE ORDER USED TO:
In Purchaser's Company	
Production	Confirm progress on purchase requisition; provide input for production scheduling
Quality Control	Provide quality standards for receiving inspection
Accounts Payable	Provide basis for verification of supplier's invoice
In Supplier's Company	
Production	Provide input to production scheduling, routing; trigger bill of materials explosion for component parts/material
Quality Control	Provide quality standards for product as produced
Material Control	Trigger inventory transfers to accomplish production of purchased goods

How a product is used can actually alter the way it is delivered. "A company in the health industry was processing seven hundred special orders annually, with delivery intervals averaging three months. Analysis showed that a relatively few part numbers accounted for 95 percent of the special orders. The remedy was to convert those frequent specials into the standard products. The delivery interval dropped dramatically – 85 percent of the orders were now delivered within two days. The number of special orders dropped from seven hundred to two hundred per year. All that was done at a substantial cost reduction. [30, page 122]."

Different users will use the same product in different ways, thus their use will determine the quality level of the product. A race car driver will require a fast engine and good cornering ability. The carpooler will look for seating for 8 and good gas mileage.

Even the same user may use a good or service in different ways at different times. As an example, consider the use of vacuum cleaners. At one time the vacuum cleaner may be used to vacuum the dust off the floor. At another time it might be used to blow the leaves off the lawn by reversing the hose so it becomes a blower, rather than a suction tool. When it is being used to vacuum the floor, the user will demand the ability to get into tight corners. The leaf blower may require the vacuum to be cordless. With both uses, the user will probably consider high power to be a necessity.

To recap, quality of a product or service is dependent on how it will be used. The same product or service can be used by different individuals (as in the case of the purchase order) or products and services of the same type may be used in different ways (as in the case of the sheet metal, car and vacuum cleaner.) Finally, the same product or service can change over time (special orders converting to standard products) which will affect the quality in the eye of the beholder.

The possibility and probability of any dangers to human safety

It should be obvious that a product or service that is harmful would not be of high quality. Most product recalls are based on the consideration of danger to the user. As Juran states, "where such risks are significant, all else is academic [28, page 428]"

When human safety is a concern, products and services must be tested in carefully controlled environments. "Toy designers use children as a source of inputs on customer needs. Some of the children are not yet able to speak, so it is necessary to create an environment (a playroom) that permits the behavior of children at play to answer such questions as: Is the toy safe? Can it be thrown about without breaking? Is it easy to handle [30, page 52]?"

Sometimes the product or service itself is dangerous. At other times, it is how the product or service is used that causes the unsafe condition. “Product features may pose direct threats to human health or safety or to the environment. Other threats may arise from user ignorance or misuse of the product. The aim of criticality analysis is to identify such threats so that steps can be taken to eliminate them [30, page 115].”

The economic resources of both the producer and the user

When considering product quality there may be a trade-off between quality and price. The car with anti-lock brakes and dual side air bags will be more expensive than a car without these safety features. It costs more to put these features into the car. Thus, the economic resources available to both the producer and the user will affect the perception of quality for the product. A small company manufacturing homemade potato chips in the kitchen of the local church will not have the same resources available for quality control as the producer of Lay’s potato chips, backed by the entire corporate organizational resources. Likewise, the new college graduate will not have the financial resources to purchase a Volvo, regardless of the safety features that have been built in. They will purchase the highest quality automobile obtainable for the resources available.

Because the economic resources of the user are different, their definition of quality will also change. Because of this, we actually have a market for “inferior” goods. “For some nonconformances, the economics of repair are so forbidding that the product must be used as is, although at a price discount. In some industries, e.g. textiles, the price structure formalizes this concept by use of a separate grade—‘seconds’” [28, page 428].

Consumers must look at the total cost of ownership (TCO) when considering their economic resources. “The national economy would benefit enormously if the ‘life cycle cost’ concepts were made effective. Under that concept, product design is aimed at minimizing the ‘cost of ownership,’ which is the sum of the purchase price plus the subsequent costs of operation and maintenance. However, many product designs are aimed at minimizing the original purchase price. That makes it easier for the supplier to sell the product, but it is no bargain for the client [30, page 109].”

Consumers may be willing to pay a premium price if their use demands a different quality than another user, and the economic resources are available. “The Power Tool Case: A maker of power tools succeeded in improving their reliability to a level well beyond that of competing tools. A team was then sent to secure field data from users on the costs of using those high-reliability tools versus those of competitors. Based on those data the differences in reliability were converted into differences in operating costs. The cost data were then propagandized and it became feasible to secure a premium price[30, page 126].”

In summary, economic resources of both the producer and the user will vary over time and the quality provided and demanded will also vary over time. Ultimately, we need to address this variation when defining the quality of the product. “Economic Analysis: The goal in economic terms is to minimize the combined costs of the customers and suppliers. Arriving at the optimum obviously requires that we determine [30, page 158]:
-What are the alternative ways for meeting (or revising) customers’ needs
-What are the associated costs, for customers and for suppliers”

The user’s specific determinants of a product or service that is fit for their use

When considering the quality of a product or service the user will consider not only the use of the product, but also what characteristics are necessary to meet that use. Referring to the car example used earlier, if safety defines a quality car, then important characteristics would include anti-lock brakes and dual side air bags. If the individual is more concerned with the image that the car will give him or her, then the style and color will be more important characteristics as to the quality of the car.

As illustrated in the following example, the characteristics required by the user may or may not be something that the producer will be able to evaluate. "…these may differ significantly from those available to the manufacturer. For example, a manufacturer of abrasive cloth used a laboratory test to judge the ability of the cloth to polish metal; a major client evaluated the cost per 1000 pieces polished [28, page 428]"

Even if the characteristics can be evaluated, they need to be evaluated within the overall context of use. "A chemicals manufacturer asked its clients to rank the company relative to its competitors on various aspects of performance: product innovation, quality, promptness of delivery, technical assistance, and so on. The company was quite pleased to learn that it was ranked first, second, or third in virtually all aspects of performance. Then someone noted that the study was biased—it included no non-clients. So a supplemental research was conducted with special attention to former clients: Why had they stopped buying? This time the research findings were less than pleasing [30, page 115]."

Finally, what the producer considers quality may be very different from what the user considers quality. If a user doesn't value a given characteristic, then it won't factor into the quality equation. Per Juran's Qantas Airways example: "the results contained surprises as well as confirmations. For example, the company managers had given high ranking to on-time departures and arrivals. It came as a surprise that those needs were not given high priority by the passengers surveyed [30, page 115]." Many of the passengers said that baggage arriving at their destination was a much higher priority and helped define a quality experience.

DATA QUALITY FRAMEWORK

The information product approach used by Wang is ideally suited to extending his approach to include Juran's fitness for use principles. Wang distinguishes different roles in the information supply chain---suppliers, producers (called "manufacturers"), consumers, and managers [80]. Although the perspective of the user is first and foremost the foundation for fitness for use, it is important for the suppliers, producers and managers to consider how the data and information will be used. Recalling the toy designer scenario from earlier, the producers of the toys must be concerned with safety issues if the toys are ever going to be used at all. If we assume that the information supply chain and the product supply chain are analogous, then we should be able to extend Juran's five principles into the data and information quality arena. Users of data and information should follow the same rigorous approach when considering the quality of their data and information as do users of products or services. We now turn to a discussion of the framework described in Wang's 1995 article and show how Juran's five principles can be integrated with it in terms of data and information quality to create a new framework.

Wang's Framework for Data Quality Research

Wang et al [78] described a framework for data quality research that had seven elements: management responsibilities, operation and assurance costs, research and development, production, distribution, personnel management and legal function. Very briefly, *management responsibilities* pertain to policies, requirements, and the data quality system. *Operation and assurance costs* were broken down into three areas: information systems, database, and accounting. *Research and development* was also focused on three areas: analysis and design of the quality aspects of data products (also semantics), incorporating data quality into the design of an information system, and R & D pertaining to dimensions and measurement. *Production* relates to raw data and the correctness of process vs. the correctness of the data. In addition, we have added data tags to this category. *Distribution* is moving the data through the system and deals with metadata and documentation. With the increasing use of data warehouses over the last ten years, we have added integration and tools to this category. *Personnel management* is concerned with training, qualifications, and motivation. Finally, *legal function* deals with safety aspects as well as product liability.

Juran's Five Principles

The above framework examines data throughout its lifecycle, as well as the policies, personnel and legal aspects of data. Let us now discuss Juran's fitness for use principles and then tie the two frameworks together.

The users of the product or service (Who?)

Who will be using the data or information? From the perspective of data production and distribution, it is important to identify who these users are going to be. Are they primary users examining financial statements from an accounting system [32], or secondary users, making management decisions based on output from a data warehouse [5, 13, 67, 82]? As indicated earlier, the very term "fitness for use" implies that the users are a key component in defining the quality of the data and information. Data consumers should talk to individuals who have worked with the data, consider where the data is coming from, and how it may have been transformed in moving from one database to another. Data consumers should then put this into the context of their own experiences and needs.

How the users will actually put the product or service to use (How?)

How will the data or information be used? Will it be used to make financial predictions or strategic plans? Will it be used to monitor how the business is doing in relation to goals? Unlike physical products, where a given product or service can be used by only one person at a time, data can, and is, used concurrently and in succession. Thus, we might see several different users of the data at its source with differing needs. And to further complicate matters, we can migrate the data to a secondary source, such as a data warehouse, and have a whole new set of users. One user will use the data from a general ledger accounting system to audit the organization and determine the ability of the financial statements to accurately reflect the position of the company. Another might use the same data, in conjunction with historical data and market analysis data to make strategic decisions regarding the direction in which the company should head. Yet another user of the same data may be responsible for creating budgets which will then be used to gauge how well the company is doing in relation to its goals.

How the data will be used is very closely tied to the quality of the data. Ultimately, we can not develop measures of data quality unless we know how the data will be used. If the data is used in a summary format it may not need to be as accurate as if it will be used in detail. Fir example, at the detail level of ethnicity it may be vital to have the exact country of origin. On the other hand, if this data rolls up into a summary based on ethnic groups then a record with country of origin as Brazil will roll up the same as if the country of origin was Argentina. Thus, development of metrics or measurements should be closely tied to how the data will be used (e.g. [8, 31, 43, 73]). Likewise, the dimension of the data to be considered by the user will have a significant bearing on how the data will be used [6, 48, 68, 79].

The possibility and probability of any dangers to human safety

Although the concept of human safety is not obvious in regards to the quality of data, it nevertheless exists. Medical data, in particular, can cause great harm to an individual if it is inaccurate, incomplete, inaccessible or insecure. Similarly, in our post 9/11 world, issues of quality play a significant role in homeland security [18, 34]. Related to the legal function in Wang's framework, product safety and reliability can have a considerable influence on the quality of the data or information.

The economic resources of both the producer and the user

Historically, the producers of data have not been rewarded for high quality [64]. Traditionally, the users of the data have worked closely with the producers of the data (many times they were the same person) and the quality of the data has not been an issue. As data is removed from the producer and taken out of the context in which it was originally collected we find that the quality of the data changes. This is what we see in the case of the data warehouse. The producers of the data have limited economic resources and

are not concerned with the quality of the data downstream. The users of the data, now removed from the producers, will need to determine the quality of the data for their needs. If the data is not of sufficient quality they will need to expend additional resources. These resources are scarce, so a suitable method for determining where the greatest benefit in relation to cost is concerned will help to allocate the resources most efficiently [4].

The economic resources of the producer and user will determine whether the quality of the data is "good enough." Even with unlimited resources, it would be impossible to achieve a quality level of 100%, within the constraints of fitness for use. Thus, it would be useful to have methods for allocating economic resources to achieve the quality level that meets the needs of the users.

The user's specific determinants of a product or service that is fit for their use (What?)
What are the characteristics that define quality for a given user? These characteristics can be defined by data quality dimensions [68, 70, 74]. Thus, if the user needs secure data that would be a different characteristic than if they needed accessible data. If the data is to be used to make long range plans for the company then accuracy may be more important than timeliness. On the other hand, if a company is bidding on a contract and the deadline was yesterday, it doesn't matter how accurate the numbers are; the bid will not be accepted because it is too late.

Putting it all together

The data quality framework described by Wang et al gives us a look at the research on data and information quality across the life cycle of the data. If we now include Juran's five principles we can build a matrix that allows us to categorize research. As can be seen in Table 1, if we lay out Wang's criteria along the vertical and Juran's principles along the horizontal, we can categorize research on two dimensions- the point in the life cycle, and how the user will define the quality. Some of the cell intersections are obvious- the legal function deals with product liabilities and thus with human safety. The specific determinants of data quality (what?) align themselves with research and development on the dimensions of data quality. Other intersections may be open to more interpretation- integration (distribution) of data for a data warehouse will inevitably involve the principles of who and how, and frequently what. Production, which includes data tags, will frequently relate to how the data is used. And operation and assurance costs generally will deal with economic resources.

	Who?	**How?**	**Human Safety**	**Economic Resources**	**What?**
Management Responsibilities					
O & A Costs- Information Systems					
O & A Costs- Database					
O & A Costs- Accounting					
R & D- Analysis & Design					
R & D- DQ in IS Design					
R & D- Dimensions & Measurement					
Production					
Distribution					
Personnel Management					
Legal Function					

Table 1- A framework for data quality literature

Next follows a discussion of the data collection methods for this study, as well as some general findings related to data quality literature in general. An analysis of the 76 journal articles that were part of the data will show where the concentration of research lies, and where further research is needed.

METHOD

Shortly after publication of the Wang et al article, the first Conference on Information Quality was held at the Massachusetts Institute of Technology (MIT) in 1996. There were 19 academic and practitioner papers presented at that conference. Since that time, the conference has grown each year and has attained an international following as well as a good blend of academic and industry papers. From 1996 through 2004 a total of 253 papers have been presented at the conference, now known as the International Conference on Information Quality (ICIQ). (See Table 2 for the breakdown by year).

Year	Number of Papers Published	Year	Number of Papers Published
1996	19	2001	39
1997	21	2002	36
1998	20	2003	35
1999	18	2004	33
2000	33	Total	253

Table 2- Count of Papers presented at ICIQ

In addition to an increase in the number of papers presented each year, the overall number of papers submitted has grown, such that the acceptance rate of the conference has decreased and the conference has become more competitive. This conference has served as an outlet for much of the emerging research and has fostered a community of data and information quality researchers.

As recognized with the growth of the ICIQ conference, research on data quality (DQ) and information quality (IQ) issues continues to grow. Using the primary criteria identified in Wang et al's paper that the research is motivated by a data quality issue, an Endnotes database was compiled containing 370 references. These references come from journals, conference proceedings and Ph.D. dissertations. Journal articles have been found in Communications of the ACM, IEEE Transactions on Knowledge and Data Engineering, Information Resources Management Journal, Information Systems Management, Journal of Data Warehousing, Journal of Database Management, Journal of Management Information Systems, MIS Quarterly, and The DATA BASE for Advances in Information Systems, among others. The majority of the conference proceedings can be attributed to ICIQ, AMCIS and ICIS. Schools represented by students completing dissertations in data and information quality include University at Albany, MIT, Arizona State, University of Michigan, University of South Florida, and University of Texas at Dallas, among others. See Table 3 for a full breakdown.

Reference Source	Count	% of Total
Journal Articles	76	21%
Conference Proceedings	284	77%
Ph.D. Dissertations	10	3%

Table 3- Breakdown of Endnotes References by Source

Gathering the data

In order to classify the DQ and IQ literature it was necessary to create a database of the conference proceedings, Ph.D. dissertations and journal articles. We began by creating an Endnotes library of all of the relevant DQ and IQ references. Endnotes is a bibliographic software designed to capture the data

related to references in a database and then interface with a word processor to format the citations and a bibliography. Additionally, the Endnotes library is searchable, allowing key word searches to facilitate categorization of references into one or more research buckets.

All of the ICIQ conference proceedings from 1996 through 2004 were entered. We then added conference proceedings from AMCIS and ICIS, as well as five articles from other conferences that were frequently cited in the DQ and IQ literature.

In addition to the conference proceedings, we searched the Proquest- ABI/Inform and EBSCO databases for peer reviewed journal articles that contained the key words "data quality" or "information quality". The focus was on articles written since 1995 in order to avoid duplication of the references in the Wang et al article. Extending the journal list through examination of article references resulted in additional articles.

Finally, the Dissertation Abstracts (via FirstSearch) database was searched for Ph.D. dissertations pertaining to data or information quality. Again, only dissertations that were motivated by a data quality issue were considered. Those that dealt with data quality as a by-product of the research (e.g. Methods for improving microwave radiometer calibration and data quality for geophysical applications) were eliminated.

The Endnotes library is available on request from the author.

Each of the 76 journal articles was coded for inclusion in the framework described above. This coding was done by the researcher. It is expected that two additional coders will be employed in the near future to provide inter-rater reliability.

RESULTS

Table 4 shows the results of the categorization of 76 journal articles related to data and information quality. Articles were coded with one or more criteria from the Wang et al framework and one or more of Juran's five principles, thus an article may be in multiple cells. Coding was based on abstracts as well as article scans. Coding was aided with the use of the data visualization tool, RefViz.

	Who?	**How?**	**Human Safety**	**Economic Resources**	**What?**
Management Responsibilities	[9], [18], [37], [36], [44], [46], [45], [50], [51], [53], [63], [64], [66], [69],	[44], [46], [51], [66], [69]	[18]		[44], [43], [46], [45], [53], [66], [68], [69]
O & A Costs- Information Systems				[8]	[8]
O & A Costs- Database	[53]			[7]	
O & A Costs- Accounting	[11], [32]				[11], [53]
R & D- Analysis & Design	[19], [36], [83]	[19], [38]			[19], [61], [68], [83]
R & D- DQ in IS Design	[21], [55], [59], [65], [74], [80]	[21], [59], [80]			[59], [80]

R & D- Dimensions & Measurement	[3], [6], [11], [35], [57], [70], [79]	[35], [70], [79]		[8], [35]	[2], [3], [6], [8], [11], [43], [48], [57], [60], [61], [70], [79]
Production	[3], [12], [15], [18], [27], [44], [46], [45], [49], [51], [62], [66], [76], [80], [77], [81], [84]	[12], [25], [27], [44], [46], [49], [51], [66], [76], [80], [77]	[15], [18]	[1], [7]	[3], [24], [25], [31], [44], [46], [45], [66], [75], [80], [84]
Distribution	[6], [5], [10], [13], [15], [17], [18], [21], [27], [33], [40], [39], [42], [44], [46], [52], [53], [54], [58], [62], [66], [67], [71], [76], [80], [79], [82].	[5], [14], [17], [21], [27], [40], [39], [44], [46], [54], [56], [58], [66], [67], [76], [80], [79], [82]	[15], [18], [20], [34], [58]	[1], [34], [67], [72]	[6], [5], [10], [40], [39], [42], [44], [46], [53], [54], [66], [67], [71], [73], [80], [79], [82]
Personnel Management	[15], [12]	[12]	[15]		
Legal Function			[15], [18], [20], [34], [58]		

Table 4- Analysis of Data Quality Journal Articles 1996- 2005

As can be seen, significant research has been completed regarding the production and distribution of data and information. Distribution in particular has seen a significant amount of research related to data warehousing. The question of who will use the data and what determinants are important (particularly with respect to quality dimensions) has been well researched. Much of the R & D pertaining to measurement and dimensions is related to dimensions, rather than measurement. More work still needs to be done with regards to measurement or metrics. However, given the fitness for use questions, measurements must be in the context of the use and there is still some question as to how those measurements should be defined. Significant holes reside where management responsibilities meet economic resources and where personnel management and the legal function intersect with economic resources and what determinants are important to users. Little work has been done across the board in terms of operation and assurance costs.

DISCUSSION

The field of data and information quality has matured a great deal in the last ten years and articles are being published in top journals. It is a field of importance to practitioners, and this may be a part of the reason that much of the literature is skewed towards the production and distribution of data. This is the phase of the life cycle that supports management decision making and justifies the investment in information systems.

Due to the many factors involved in developing and deploying information systems, return on investment (ROI) is difficult to calculate. ROI is considered in Table 4 in the economic resources cells. Since data are an integral part of information systems, it is equally difficult to calculate costs associated with quality. When you factor in the need to assign costs based on how the data is used, the difficulty increases significantly. Thus, we find that there has not been a great deal of research in this area. If we could define measurements or metrics of quality then we would be able to get a step closer to calculating a return on investment.

It is also time to move beyond the definition of data quality dimensions and determine how these dimensions define the quality of data in terms of the user. How do we assign an economic cost of data quality if we don't know how the user will use the data or what specific requirements the user has for the data? And how do we develop policies for data quality if we don't incorporate secondary users into the equation?

Data quality in terms of human safety has belatedly been recognized, primarily due to 9/11 and medical malpractice suits. Given the relatively short time that this has been acknowledged, there are a surprising number of articles addressing it. However, the cost of human safety (operation and assurance costs) still needs to be investigated.

Finally, more research in terms of improving the analysis and design of information systems to include quality constructs is needed. In particular, we need to address the economic resources associated with providing poor quality data for secondary use.

LIMITATIONS AND FUTURE RESEARCH

As noted earlier, the classification of journal articles must be supported by additional researchers. Ideally, the definitions for classification will be more finely tuned, thus enabling the research to be replicated. In addition, all of the available references on data quality should be classified, including conference proceedings.

Other research "buckets" could also be examined to highlight areas where research is heavy or light. We could examine research in terms of domain, industry affiliation, functional area, organizational level, application areas, DQ/IQ paradigms or research methodologies. For instance, many frameworks are proposed [5, 6, 13, 23, 27, 39, 41, 51, 52, 61, 66, 68, 76, 79, 83], but only a small subset of the research could be described as empirical [9, 12, 21, 24, 40, 64, 73, 74, 79, 82]. It is time to support the frameworks with empirical data.

CONCLUSION

When "A Framework for Analysis of Data Quality Research" was written in 1995 the field was still in its infancy in terms of research. We now see recognition of data and information quality as a critical player in the information systems field. Conferences are devoted to the topic and special journal issues are called. As a field of research, it is beginning to mature. The approach described in the Wang et al article continues to be a solid foundation for classifying the literature. Additional insights can be gained if we overlay Juran's principles defining fitness for use. Quality continues to "be in the eye of the beholder" and accepting this premise will help us to further the research in the field.

REFERENCES

1. Aiken, P.H., *Reverse Engineering of Data.* IBM Systems Journal, 1998. **37**(2): p. 246-259.
2. Audini, B., A. Pearce, and P. Lelliott, *Accuracy, completeness and relevance of Department of Health returns on provision of mental health residential accommodation: A data quality audit.* Journal of Mental Health, 2000. **9**(4): p. 365-370.
3. Ballou, D., et al., *Modeling Information Manufacturing Systems to Determine Information Product Quality.* Management Science, 1998. **44**(4): p. 462-484.
4. Ballou, D.P. and G.K. Tayi, *Methodology for Allocating Resources for Data Quality Enhancement.* Communications of the ACM, 1989. **32**(3): p. 320-329.
5. Ballou, D.P. and G.K. Tayi, *Enhancing data quality in data warehouse environments.* Communications of the ACM, 1999. **42**(1): p. 73-80.
6. Ballou, D.P. and H.L. Pazer, *Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts.* IEEE Transactions on Knowledge & Data Engineering, 2003. **15**(1): p. 240-243.
7. Becker, S., *A practical perspective on data quality issues.* Journal of Database Management, 1998. **9**(1): p. 35-37.
8. Birman, S., *Control the Data, Control the Costs.* Quality, 2003. **42**(1): p. 50-56.
9. Bowen, P.L., D.A. Fuhrer, and F.M. Guess, *Continuously Improving Data Quality in Persistent Databases.* Data Quality, 1998. **4**(1): p. 19.
10. Bricker, J. and A. Maydanchik, *Data quality assurance: Plan redesign affords an opportunity to consider an automated approach to cleansing employee records.* Compensation & Benefits Management, 1999. **15**(4): p. 49-55.
11. Cappiello, C., C. Francalanci, and B. Pernici, *Time-Related Factors of Data Quality in Multichannel Information Systems.* Journal of Management Information Systems, 2003. **20**(3): p. 71-91.
12. Chengalur-Smith, I., D.P. Ballou, and H.L. Pazer, *The Impact of Data Quality Information on Decision Making: An Exploratory Analysis.* IEEE Transactors on Knowledge and Data Engineering, 1999. **11**(6): p. 853-864.
13. Chiang, R.H.L., E.-P. Lim, and V.C. Storey, *A Framework for Acquiring Domain Semantics and Knowledge for Database Integration.* The DATA BASE for Advances in Information Systems, 2000. **31**(2): p. 46-64.
14. Curtis, M.B. and K.D. Joshi, *Lessons learned From the Implementation of a Data Warehouse.* Journal of Data Warehousing, 1998. **3**(2): p. 12-18.
15. Davidson, B., Y. Lee, and R. Wang, *Developing data production maps: meeting patient discharge data submission requirements.* Int. J. Healthcare Technology and Management, 2004. **6**(2): p. 223-240.
16. Davis, G.B. and M.H. Olson, *Management Information Systems: Conceptual Foundations, Structure and Development*. 1985, New York, NY: McGraw Hill Book COmpany.
17. Embury, S.M., et al., *Adapting integrity enforcement techniques for data reconciliation.* Information Systems, 2001. **26**(8): p. 657-689.
18. English, L., *Information quality: Critical ingredient for national security.* JOURNAL OF DATABASE MANAGEMENT, 2005. **16**(1): p. 18-32.
19. Fan, W., et al., *Discovering and reconciling value conflicts for numerical data integration.* Information Systems, 2001. **26**(8): p. 635-656.
20. Fisher, C.W. and B.R. Kingma, *Criticality of data quality as exemplified in two disasters.* Information & Management, 2001. **39**(2): p. 109.
21. Fisher, C.W., I. Chengalur-Smith, and D.P. Ballou, *The Impact of Experience and Time on the Use of Data Quality Information in Decision Making.* Information Systems Research, 2003. **14**(2): p. 170-189.
22. Fox, C., A. Levitin, and T. Redman, *The Notion of Data and Its Quality Dimensions.* Informaiton Processing & Management, 1993. **30**: p. 9-19.
23. Fox, C., A. Levitin, and T. Redman, *The Notion of Data and Its Quality Dimensions.* Information Processing & Management, 1993. **30**: p. 9-19.
24. Heerwegh, D. and G. Loosveldt, *An evaluation of the effect of response formats on data quality in Web surveys.* Social Science Computer Review, 2002. **20**(4): p. 471-484.
25. Henderson, I. and D. Murray, *Prioritising and deploying data quality improvement activity.* Journal of Database Marketing & Customer Strategy Management, 2005. **12**(2): p. 113-119.
26. Huh, Y.U., et al., *Data Quality.* Informaiton and Software Technology, 1990. **32**: p. 559-565.
27. Jang, Y., A.T. Ishii, and R.Y. Wang, *A Qualitative Approach to Automatic Data Quality Judgment.* Journal of Organizational Computing & Electronic Commerce, 1995. **5**(2): p. 101-122.
28. Juran, J.M., *Quality Planning and Analysis: From Product Development Through Use*. 1980, New York: McGraw-Hill. 629.
29. Juran, J.M., *Juran's Quality Control Handbook*. 4 ed, ed. F.M. Gryna. 1988, New York: Mc-Graw-Hill, Inc.
30. Juran, J.M., *Juran on Planning for Quality*. 1988, New York: The Free Press.

31. Kahn, B., D. Strong, and R. Wang, *Information quality benchmarks: Product and service performance.* Communications of the ACM, 2002. **45**(4).
32. Kaplan, D., et al., *Assessing Data Quality in Information.* Communications of the ACM, 1998. **41**(2): p. 72-78.
33. Kelly, V.E., C.P. Thomas, and H. Wang, *Managing Data-Based Systems Across Releases Using Historical Data Dictionaries.* Bell Labs Technical Journal, 2000. **5**(2): p. 121-133.
34. Kim, W., *On US homeland security and database technology.* JOURNAL OF DATABASE MANAGEMENT, 2005. **16**(1): p. 1-17.
35. Klein, B.D., *How Do Actuaries Use Data Containing Errors? Models of Error Detection and Error Correction.* Information Resources Management Journal, 1997. **10**(4): p. 27-36.
36. Klein, B.D., D.L. Goodhue, and G.B. Davis, *Can Humans Detect Errors in Data? Impact of Base Rates, Incentives, and Goals.* MIS Quarterly, 1997. **21**(2): p. 169-194.
37. Klein, B.D., *Detection of Data Errors in the Practice of Inventory Management.* The Journal of Computer Information Systems, 1999. **40**(2): p. 34-40.
38. Klein, B.D. and D.F. Rossin, *Data quality in neural network models: Effect of error rate and magnitude of error on predictive accuracy.* Omega, 1999. **27**(5): p. 569.
39. Klein, B.D., *User perceptions of data quality: Internet and traditional text sources.* The Journal of Computer Information Systems, 2001. **41**(4): p. 9-15.
40. Klein, B.D., *Internet data quality: Perceptions of graduate and undergraduate business students.* Journal of Business and Management, 2002. **8**(4): p. 425-432.
41. Lam, M. and R.K.H. Ching, *Information Integration in Multidimensional Databases: A Case Study.* Information Systems Management, 1998. **15**(4): p. 36-45.
42. Lee, T.Y. and Y. Yang, *Constraint-based wrapper specification and verification for cooperative information systems.* Information Systems, 2004. **29**(7): p. 617-636.
43. Lee, Y., et al., *AIMQ: a methodology for information quality assessment.* Information and Management, 2002. **40**(2): p. 133-146.
44. Lee, Y., *Crafting Rules: Context-Reflective Data Quality Problem Solving.* Journal of Management Information Systems, 2003. **20**(3): p. 93-119.
45. Lee, Y.W. and D.M. Strong, *Knowing-Why About Data Processes and Data Quality.* Journal of Management Information Systems, 2003. **20**(3): p. 13-39.
46. Lee, Y.W., et al., *Process-Embedded Data Integrity.* Journal of Database Management, 2004. **15**(1): p. 87-104.
47. Levitan, A.V. and T.C. Redman, *Data as a Resource: Properties, Implications, and Prescriptions.* Sloan Management Review, 1998. **40**(1): p. 89-100.
48. Levitin, A. and T. Redman, *Quality dimensions of a conceptual view.* Information Processing & Management, 1994. **31**(1): p. 81-88.
49. Levitin, A.V. and T.C. Redman, *A model of the data (life) cycles with application to quality.* Information and Software Technology, 1993. **35**(4): p. 217-224.
50. Levitin, A.V. and T.C. Redman, *Data as a resource: Properties, implications, and prescriptions.* MIT Sloan Management Review, 1998. **40**(1): p. 89-102.
51. Loosveldt, G., A. Carton, and J. Billiet, *Assessment of survey data quality: a pragmatic approach focused on interviewer tasks.* International Journal of Market Research, 2004. **46**(1): p. 65-83.
52. Low, W.L., M.L. Lee, and T.W. Ling, *A knowledge-based approach for duplicate elimination in data cleaning.* Information Systems, 2001. **26**(8): p. 585-606.
53. Lyon, J., *Customer Data Quality: Building the Foundation for a One-to-One Customer Relationship.* Journal of Data Warehousing, 1998. **3**(2): p. 38-47.
54. Madnick, S., R. Wang, and R. Xiang Xian, *The Design and Implementation of a Corporate Householding Knowledge Processor to Improve Data Quality.* Journal of Management Information Systems, 2003. **20**(3): p. 41-70.
55. Marsh, R., *Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management.* Journal of Database Marketing & Customer Strategy Management., 2005. **12**(2): p. 105-112.
56. Naumann, F., J.-C. Freytag, and U. Leser, *Completeness of integrated information sources.* Information Systems, 2004. **29**(7): p. 583-615.
57. Nord, G., J. Nord, and H. Xu, *An investigation of the impact of organization size on data quality issues.* JOURNAL OF DATABASE MANAGEMENT, 2005.
58. Orr, K., *Data Quality and Systems Theory.* Communications of the ACM, 1998. **41**(2): p. 66-71.
59. Pierce, E.M., *ASSESSING DATA QUALITY WITH CONTROL MATRICES.* CACM, 2004. **47**(2): p. 82-86.
60. Pipino, L.L., Y.W. Lee, and R.Y. Wang, *Data Quality Assessment.* Communications of the ACM, 2002. **45**(4): p. 211-218.
61. Price, R. and G. Shanks, *A semiotic information quality framework: development and comparative analysis.* JOURNAL OF INFORMATION TECHNOLOGY, 2005. **20**(2): p. 88-102.

62. Rahm, E. and H.H. Do, *Data Cleaning: Problems and Current Approaches.* Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 1999.
63. Redman, T.C., *Improve data quality for competitive advantage.* Sloan Management Review, 1995: p. 99-107.
64. Redman, T.C., *The Impact of Poor Data Quality on the Typical Enterprise.* Communications of the ACM, 1998. **41**(2): p. 79-82.
65. Scannapieco, M., et al., *The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems.* Information Systems, 2004. **29**(7): p. 551-582.
66. Shankaranarayan, G., M. Ziad, and R.Y. Wang, *Managing Data Quality in Dynamic Decision Environments: An Information Product Approach.* Journal of Database Management, 2003. **14**(4): p. 14-33.
67. Shankaranarayanan, G. and A. Even, *MANAGING METADATA IN DATA WAREHOUSES: PITFALLS AND POSSIBILITIES.* Communications of the Association for Information Systems (CAIS), 2004. **14**(2004): p. 247-274.
68. Shanks, G. and P. Darke, *A Framework for Understanding Data Quality.* Journal of Data Warehousing, 1998. **3**(3): p. 46-51.
69. Strong, D.M., Y. Lee, and R. Wang, *10 Potholes in the Road to Information Quality.* IEEE Computer, 1997. **30**(8): p. 38-46.
70. Strong, D.M., Y.W. Lee, and R.L. Wang, *Data Quality in Context.* Communications of the ACM, 1997. **40**(5): p. 103-110.
71. Tejada, S., C.A. Knoblock, and S. Minton, *Learning object identification rules for information integration.* Information Systems, 2001. **26**(8): p. 607-633.
72. Vassiliadis, P., et al., *Arktos: towards the modeling, design, control and execution of ETL processes.* Information Systems, 2001. **26**(8): p. 537-561.
73. Wallace, W.A., *Assessing the quality of data used for benchmarking and decision-making.* The Journal of Government Financial Management, 2002. **51**(3): p. 16-21.
74. Wand, Y. and R.Y. Wang, *Anchoring Data Quality Dimensions in Ontological Foundations.* Communciations of the ACM, 1996. **39**(11): p. 86-95.
75. Wand, Y., V.C. Storey, and R. Weber, *An Ontological Analysis of the Relationship Construct in Cenceptual Modeling.* ACM Transactions on Database Systems, 1999. **24**(4): p. 494-528.
76. Wang, R., et al., *Manage Your Information as a Product.* Sloan Management Review, 1998. **39**(4): p. 95-105.
77. Wang, R.Y., M.P. Reddy, and H.B. Kon, *Toward quality data: An attribute-based approach.* Decision Support Systems, 1995. **13**(3/4): p. 349-362.
78. Wang, R.Y., V.C. Storey, and C.P. Firth, *A Framework for Analysis of Data Quality Research.* IEEE Transactions on Knowledge and Data Engineering, 1995. **7**(4): p. 623-641.
79. Wang, R.Y. and D.M. Strong, *Beyond Accuracy: What Data Quality Means to Data Consumers.* Journal of Management Information Systems (JMIS), 1996. **12**(4): p. 5-34.
80. Wang, R.Y., *A product Perspective on Total Data Quality Management.* Communications of the ACM, 1998. **41**(2): p. 58-65.
81. Winkler, W.E., *Methods for evaluating and creating data quality.* Information Systems, 2004. **29**(7): p. 531-550.
82. Wixom, B.H. and H.J. Watson, *An Empirical Investigation of the Factors Affecting Data Warehousing Success.* MIS Quarterly, 2001. **25**(1): p. 17-38.
83. Xu, H., et al., *Data quality issues in implementing an ERP.* Industrial Management + Data Systems, 2002. **102**(1/2): p. 47-58.
84. Zeffane, R. and B. Check, *Does User Involvement During Information Systems Development Improve Data Quality?* Human Systems Management, 1998. **17**(2): p. 115-121.
85. Zmud, R.W., *An Empiricaql Investigation of the Dimensionality of the COncept of Information.* Decision Sciences, 1978. **9**: p. 187-195.