EVOLUTIONAL DATA QUALITY: A THEORY-SPECIFIC VIEW

(Research Paper)

Liping Liu
University of Akron
liu@acm.org

Lauren N. Chi
University of Akron
nan@uakron.edu

Abstract: A data evolution life cycle typically consists of four stages of data activities: collection, organization, presentation, and application. During the cycle, data evolve as per the needs and specifications of a theory. In this paper we propose a concept of theory-specific data quality (DQ) that stipulates DQ is defined and measured as the extent to which data meet the needs and specifications of a theory. Depending on when a theory is applied during data evolution, we define DQ respectively at collection, organization, presentation, and application levels. We derive measurement attributes based on a fishbone cause-effect diagram. We compare our models with existing ones and point our future research directions to further validate and apply the proposed measurement models.

Key Words: Data Quality, Data Evolution Life Cycles, Theory-Specific Data Quality, and Fishbone Diagrams

INTRODUCTION

Data quality (DQ) has become a critical concern in management information systems research [9]. Studies have shown that DQ is both a multi-dimensional and a hierarchical concept [24], and it falls into one of four general ways: as excellence, value, conformance to specifications, or meeting or exceeding consumer expectations [8]. Although research efforts on DQ presented in the existing literature have addressed a significant advance in its short history, a generally accepted DQ model has not appeared yet.

Existing studies have mostly used intuitive approaches to define DQ and derive its attributes. The resulting measurement models have shown their values for diagnosing practical DQ problems and crises [7, 19]. However, due to their exploratory nature, the models tend to be developed as a universal and exclusive checklist. They lack a conceptual base and theoretical justification. Their semantic validity has been debatable [3], needless to say their empirical validity such as the convergent and factorial validity of constructs [5, 14]. Although data may evolve in the process from being collected to being utilized, existing models are not precise regarding to which stage of data the models are applicable.

In this paper, we attempt to develop a theoretically sound definition of DQ. We argue that data are the reflection of real world objects through a theory that designates a set of models, methods, techniques, approaches, and heuristics used for data collection, organization, presentation, and application. We define DQ as the extent to which data meet the needs and specifications of the theory. Then by analyzing a typical data evolution process, we propose that DQ be defined as an evolutional construct that evolves from collection quality, to organization quality, to presentation quality, and finally to application

quality. In a sense, existing studies define DQ as application quality but include attributes of others such as collection quality and presentation quality. Finally, by conforming to the theory-specific view and analyzing most frequent DQ causes using a fishbone diagram, we theoretically derive the measurement attributes for the construct of evolutional DQ. Most of derived attributes are similar to those in existing studies. However, ours have clearer definitions, reflect unique quality problems or meet unique theory specifications, and are semantically distinct. In addition, our models are shown to be more efficient and control the confounding factors in measurement.

The rest of this paper is organized as follows. We first review existing studies about DQ and their limitations. Then we define data evolution life cycles, describe the theory-specificity of DQ, and present four evolutional definitions of DQ. We derive measurement attributes based on a fishbone diagram. Finally we conclude this study and point out future research directions.

LITERATURE REVIEW

Existing studies have reached an agreement that data quality (DQ) is a multi-dimensional construct. However, they have differences with respect to specific measurement attributes or dimensions [1, 21, 24, 25]. For example, Wang and Strong [25] proposed a measurement structure, which consists of four categories of attributes: intrinsic DQ, Contextual DQ, Representational DQ, and Accessibility DQ (see Table 1). On the other hand, Kahn et al. [8] proposed a different set of dimensions and grouped them into four different categories: sound information, useful information, dependable information, and usable information. The difference is attributable to many reasons. For example, there is an indication that the DQ measurement model changes as data application changes. Wang et al. [23] argued that different users had different quality standards, and even for a single user, different data applications required different DQ dimensions. Wand and Wang [22] also suggested that the notion of a data quality depended upon the use of data and what may be considered good data in one case might not be sufficient in another case.

Categories	**Data Quality Dimensions**
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Access, Security
Contextual DQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of Data
Presentational DQ	Interpretability, Ease of understanding, Concise Representation, Consistent representation

Table 1. The Measurement Model of Wang and Strong [25]

There have been three different approaches in which DQ attributes are derived or proposed. Among them, the intuitive approach is most prevalent whereas theoretical and empirical approaches are also documented. The intuitive approach identifies DQ attributes based on an expert's personal experience and intuitive understanding about what attributes are "important." Similarly, the empirical approach lets data consumers determine the characteristics to be used to assess whether data fit their tasks [25]. On the other hand, the theoretical approach emphasizes deriving attributes based on an established theory. Four theories have been proposed along this line of inquiries. Shannon and Weaver [17] proposed a mathematical theory of communication to provide a probabilistic treatment of noisy transmission. Information economics seeks to evaluate information in terms of its use. Operations research proposes an analogical reasoning to develop DQ attributes using the analogy between data and products [12, 21, 24]. Recently, Wand and Wang [22] proposed a theory based on ontological mappings.

The intuitive and empirical approaches have been important in the early stages of DQ research. For example, using an empirical approach Wang and Strong [25] derived seventeen dimensions of DQ,

which Lee et al. [9] further grouped into four categories: Intrinsic DQ, Contextual DQ, Presentational DQ, and Accessibility DQ (see Table 1). These dimensions have been adopted in many later studies and applications. However, the intuitive and empirical approaches have a common drawback. That is, their derived measurement models including attributes and factor structures that are often limited by a researcher's personal experience. They lack theoretical underpinnings on how a certain attribute is derived and defined. They lack theoretical justifications on why and how attributes are grouped into certain first- or second-order constructs and whether each construct is a formative or reflective factor of DQ [4]. For example, Wang and Strong [25] suggested DQ categories from the aspects of data storage (accessibility and presentational DQ) and data application (contextual DQ). Later Strong et al. [20] suggested four different categories: intrinsic, accessibility, contextual, and presentational. However, there was no theoretical basis that justifies why they classified those dimensions. It is also not clear why there exist just the four categories, no more and no fewer.

As a result, the intuitive and empirical approaches often create divergent and confusing definitions of basic DQ attributes. For example, existing studies have shown that "completeness" is an important attribute of DQ. However, its definition varies from study to study. Pipino et al. [11] and Kahn et al. [8] defined "completeness" as "the extent to which information is not missing and is of sufficient breadth and depth for the task at hand." In contrast, Ballou and Pazer [1] defined it as "no missing value." They also create divergent and conflicting factorial structures of common DQ constructs. At the super-attribute level, Wang et al. [24] intuitively proposed four categories according to their importance to users. Later, Wang and Strong [25] proposed totally different four categories by surveying data users. At the sub-attribute levels, the attribute "completeness" has been classified as a sub-attribute for believability [24], for contextual quality [25], and for integrity [3]. Similarly, Wang et al. [24] classified "believable" as one of the four categories of DQ whereas Wang and Strong [25] counted "believability" as one sub-attribute of intrinsic quality.

In contrast, the theoretical approach overcomes many of the above difficulties. For example, based on ontological mapping, Wand and Wang [22] argued that data should be in an exhaustive mapping with the real world. I.e., a real-world state can be mapped into more than one state in an information system but a state in an information system cannot represent two or more states in the real world. Therefore, data are deemed incomplete if there is no state in the information system corresponds to a real-world state. Similarly, data are deemed ambiguous if a state in the information system corresponds to two or more real-world states; data are meaningless if a state in the information system does not correspond to any real-world states. Attributes derived as such have crystal-clear definitions and theoretically sound justifications.

Although the theoretical approach is superior to intuitive and empirical counterparts, existing theoretical approaches are limited in their ability to derive a full-fledged DQ measurement model. For example, using ontological mappings Wand and Wang [22] derived four DQ attributes: complete, unambiguous, meaningful, and correct. However, these attributes are only a small sample of the attributes in assessing intrinsic DQ. The ontological mapping approach, consequently, leaves many other important attributes unspecified. Similarly, using the analogy between data and products Wang et al. [24] were able to derive more attributes. However, the analogical approach is still limited because data are after all different from products. From being collected, through being organized and presented, to being applied, data are subject to many different sources of errors from products in a manufacturing process. Most importantly, product quality may be measured against a set of industrial specifications whereas such specifications often do not exist when measuring DQ. Realizing the limitation of the analogical approach, recently Kahn et al. [8] defined DQ from both product and service quality perspectives to emphasize the dimensions related to the service delivery process as well as addressing the intangible measures of ease of manipulation, security, and added value of the information to consumers.

EVOLUTIONAL DATA QUALITY

Data are the reflection of business objects and processes. From being observed to being utilized, data typically evolve through a sequence of stages consisting of data collection, organization, presentation, and application. First, data are captured through observing real world processes, measuring real world objects, tangible and intangible as well, and perceiving real world stimulus. Then data are organized and stored in simple file-based data stores or sophisticated databases. Then data are processed, re-interpreted, summarized, formatted, and presented in certain views in the parlance of database management. Finally, data are utilized to achieve a certain application purpose, which in turn directs further data capturing. We call such a sequence of data evolution the *data evolution life cycle* (DELC) as shown in Figure 1.

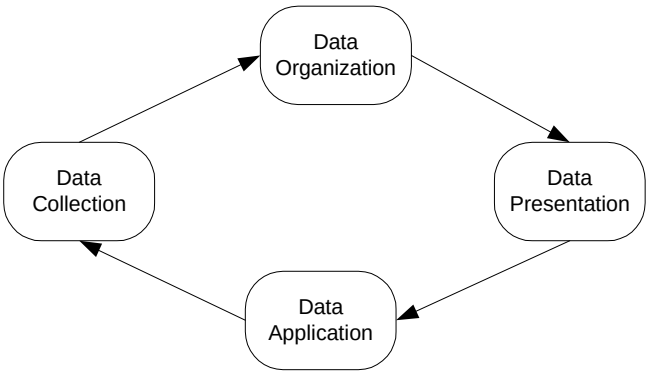

Figure 1. Data Evolution Life Cycle

Existing studies often consider data as a type of products and use the analogy between data and products to develop measurement models of DQ [12, 21, 24]. However, we believe that the analogy is flawed at best. Unlike physical products that can exist independently, data can only exist and have meaning through a theory. First, when data are captured, we use a method such as randomization or factorial design to guide the collection process. Second, when data are stored, we use a framework such as a relational or object-oriented model to guide the data organization. Third, when data are presented, we choose a view model, a layout model, a style model, and a language to summarize and format data and add syntactic and semantic derivatives. Finally, when data are utilized, we employ a technique, approach, method, or model such as regression analysis or linear programming to guide the data application. For simplicity, we use "theory" as a general designation for any technique, method, approach, or model that is employed during a DELC. Accordingly, theories are used to create, organize, interpret, and apply data whereas data serve the purpose of theories and reflect real world objects through theories. Due to the attachment of data to theories, we believe that, when defining DQ, we need to consider how data meet the specifications or serve the purposes of a theory. We call such a concept of quality *theory-specific*.

As data evolve through the stages of a DELC, they are typically under a sequence of transformations and exist independently as different morphons. They first appear as captured data, then as organized data, then as presented data, and finally as utilized data. The transformations generally cause the utilized data different from what was presented, presented data different from what was stored, and organized data different from what was captured. Therefore, data at different stages of a DELC, called *morphons*, are not the same objects of which quality is to be measured. In addition, each transformation introduces independent errors to the different stages of a DELC. Examples of such errors include measurement errors during data collection, data entry errors during data organization, and interpretation biases for data presentation. Consequently, the quality of data as different morphons in a DELC is not the same. In other words, the quality of captured data may not be the same as that of organized data, of presented data, and of utilized data. Finally, as we have argued, DQ is theory-specific. Typically, different theories are applied at different stages of a DELC. Since different theories have different specifications and purposes, the criteria to meet them will be different and the definitions of DQ for different morphons must be dif-

ferent. For example, to judge whether data meet the specifications of a relational model for data organization, we typically use the criteria to ensure that the data are efficiently stored and can be efficiently retrieved. However, when considering whether data are good for a statistical analysis, we pay more attention to their relevance and whether they are normally distributed.

Based on these arguments, we propose a concept of evolutional DQ that consists of the following three components. First, the concept emphasizes the use of different definitions to measure the quality of data at the different stages of a DELC. Instead of a single universal concept of DQ, it promotes four hierarchical views of DQ from the bottom to the top as collection quality, organization quality, presentation quality, and application quality, which are respectively applied to measure the quality of collected data, of stored data, of presented data, and of utilized data. Second, the concept suggests the evolutionary nature of the four view of DQ, i.e., the quality of data at earlier stages of a DELC positively contributes to that of data at later stages. In details, collection quality is a component for organization quality. Both collection and organization qualities are components for presentation quality. Finally, collection, organization, and presentation qualities all contribute to the measurement of application quality. Third, the concept suggests a monotonically increasing order of specificity of four views of DQ; application quality is more specific than presentation quality, which is more specific than organization quality, which in turn is more specific than collection quality. This order of specification implies that the definition of DQ at a lower level has a broader view and is less restrictive than that at a higher level (see Figure 2). This makes sense first from the perspective of the mapping cardinality of data transformations in a DELC; one set of collected data may be stored as many sets of organized data, each of which may be further formatted or interpreted as many sets of presented data, and each of which in turn can be further applied for many tasks. Therefore, one measure of DQ at a lower level is useful to measure the quality of many sets of data at a higher level. The order of specification also makes sense from the notion that DQ is theory-specific; the higher a level is, the more theories have to be satisfied in order to define the DQ at the level. For example, to define application quality, we consider not only how data serve the purpose of an application theory but also how they meet the specifications of theories for data presentation, organization, and collection. In contrast, to define collection quality, we consider the specification of the theory for data collection only.

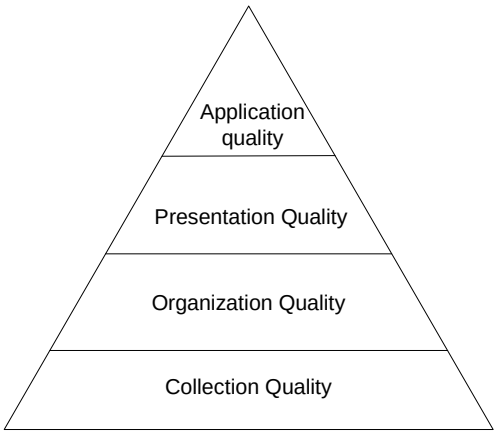

Figure 2: Evolutional Data Quality

In sum, the notion of evolutional quality conceptualizes four evolutional definitions of DQ for four evolutional data morphons in a DELC. It states two evolutions of DQ during the data evolution. First, the measurement model is becoming increasingly restrictive by having more and more attributes. Second, the quality measure is accumulative, i.e., the quality of early stage data contributes to the quality of later stage morphons. The two evolutions jointly imply that the measurement model is cumulative. In other words, while having additional attributes, a measurement model accumulates the attributes that are used to measure the quality of data at earlier stages. For example, the measurement model for organized

data includes the attributes that reflect data organization process and the specification of a data organization theory. In addition, it includes the attributes to measure the quality of collected data.

THE MEASUREMENT OF EVOLUTIONAL DATA QUALITY

By viewing DQ to be both theory-specific and evolutional, we can develop two complementary approaches to derive attributes and develop semantically valid measurement models for collection, organization, presentation, and application quality. The first approach is based on the analysis of the sources of data errors during data evolution. The second is based on the analysis of how data at each stage of a DELC meet the specifications of a theory. In a sense, the first approach derives attributes measuring the quality of a data evolution process such as collection and organization. The second approach derives the attributes measuring the quality of data as information goods after each evolution process.

To implement the first approach, we use a fishbone diagram (see Figure 3), an analysis tool invented by Japanese quality control statistician Kaoru Ishikawa, to systematically examine the causes that contribute to DQ. Note that the design of a fishbone diagram looks much like the skeleton of a fish. The head of fish shows the problem to be studied. Each bone of the fish labels a cause that leads to the problem. In addition, the analysis tool suggests that we look for causes from typical categories signified as the 4 M's — Methods, Machines, Materials, Manpower, the 4 P's — Place, Procedure, People, Policies, and the 4 S's — Surroundings, Suppliers, Systems, Skills. Of course, we may use one of the four categories suggested, combine them in any fashion, or make up our own. By adapting the tool to the analysis of poor DQ, Figure 3 shows typical process errors along the way of data evolution.

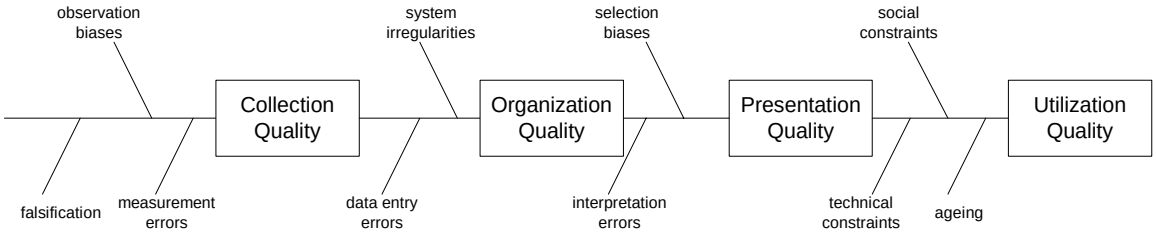

Figure 3. A Fishbone Diagram for Evolutional Data Quality

To construct the fishbone diagram in Figure 3, we surveyed the existing literature on DQ and ran three sessions of braining storming among researchers and DQ practitioners. We initially produced a long list of root causes to poor quality. However, to concisely present the list, we grouped similar and related causes into parsimonious ones. We also deleted some root causes that are not frequently encountered and whose effects are insignificant to DQ. In the following subsections, we will explain the listed causes from stage to stage, derive corresponding measurement attributes, and develop measurement models.

Collection Quality

During data collection, the most frequent quality problems include observation biases, measurement errors, and intentional falsification. When selecting a sample to observe, it is imperative that the sample is representative of the population under study. Otherwise, the collected data will lack objectivity. A typical example of observation biases is phone interviews with home residents in weekdays; the survey is biased because it does not reflect the opinions of those who work. Measurement errors are not avoidable since any instrument has limited capacity [6]. They can also occur due to human mistakes. In either case, they reduce the accuracy of collected data. Falsification refers to the behavior of making up false data. It includes the creation of data that do not correspond to real world objects and intentional distortion of obser-

vations. Falsification reduces the credibility of observed data and has negative impacts on data consistency, i.e., different values of the same object are logically and intuitively consistent, and existence, i.e., each datum describes an object that exists.

After data being collected, they must be compared with the specifications of a theory that is used to guide the data collection. The two most fundamental requirements of any collection theory are completeness and clarity. Completeness implies that all values that are supposed to be collected as per the collection theory should be collected. Clarity means that data contain no fuzzy and ambiguous observations. Besides completeness and clarity, a specific collection theory may have its own additional requirements. For example, when using a balanced factorial design, the number of observations under each condition must be the same. Otherwise, DQ suffers. By combining the attributes obtained from both root causes and theory specifications, we propose a measurement model for collection quality as shown in Figure 4.

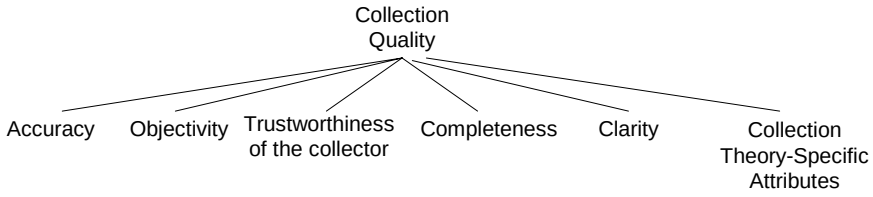

Figure 4. The Measurement of Collection Quality

Organization Quality

During data organization, the most frequent quality problems include data entry errors and system irregularities. The four basic data manipulations are creation, retrieval, updating, and deletion. Data entry errors occur when creating new records. The magnitude of entry errors is determined by the extent to which the data entry clerk is reliable. Thus, we use reliability of data entry to account for this aspect of DQ. System irregularities occur during data maintenance such as updating and deletion. Typical irregularities include lack of concurrency control so that two or more users can simultaneously update the same data, lack of transaction control or atomicity so that a transaction can be partially finished, and lack of a mechanism to automatically update the same data in multiple places. All these irregularities create one common problem — data inconsistency, i.e., different data in a database are not logically compatible. For examples, the ending balance of an account does not match all the transactions on the account; the same information about one customer appears differently in different places of the database. Thus, we can use consistency to account for the impact of system irregularity on DQ.

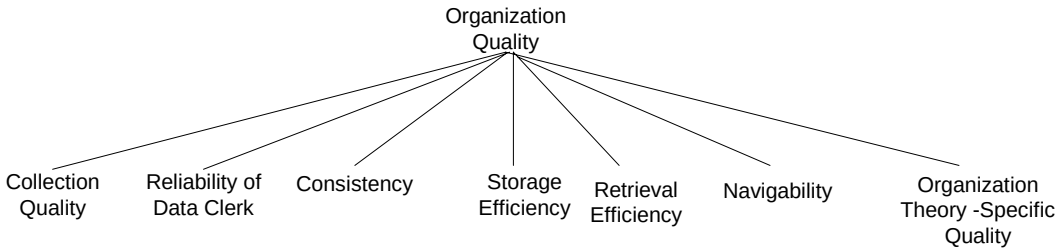

Figure 5. The Measurement of Organization Quality

After data being organized, they must be compared with the specifications of a theory such as the relational model, the hierarchical model, the network model, and the object-oriented model, which is used to guide the organization. The three most fundamental requirements of any data organization theory are storage efficiency — it takes less space to store the data, retrieval efficiency — it is fast to find desired information, and navigability — one can navigate around the related information [10]. Besides these three criteria, an individual organization theory may have its own specific ones. For example, when using the relational model, one cares about missing values (completeness) and redundancy. However, when XML is

used, completeness is not required. Similarly, redundancy is not an issue when the network model is used. By pooling the attributes obtained so far, we propose a measurement model for organization quality as shown in Figure 5. Note that, because DQ is evolutional, we include Collection Quality as a component for Organization Quality.

Presentation Quality

During data presentation, the most frequent causes to DQ problems are selection biases and interpretation errors. Selection biases prevent one from presenting data objectively and neutrally. They include hiding data that have conflicts of interest and highlighting data that favor certain opinions. Interpretation errors occur when original data have ambiguity in meaning, when they are difficult to be understood, or when there are language or tool deficiencies. For example, in a distributed heterogeneous system, the same field may have different definitions, formats, and values in different subsystems. In order to synthesize the subsystems and present their data in a unified view, interpretations and reinterpretations of certain data are needed and errors are not avoidable. Interpretation errors influence the faithfulness of a presentation to its original source.

Besides neutrality and faithfulness, the presented data must be measured against the requirements of a theory that guides the data presentation. The most fundamental specifications of any presentation theory include interpretability — data must have clear meaning, formality — data must be presented concisely and consistently, and semantic stability — the same data have same meaning across time and space. In addition, individual theories may have additional theory-specific criteria. For example, when using web pages to present data, being able to navigate around the data through hyperlinks is very crucial. In sum, we propose a measurement model for presentation quality as shown in Figure 6.

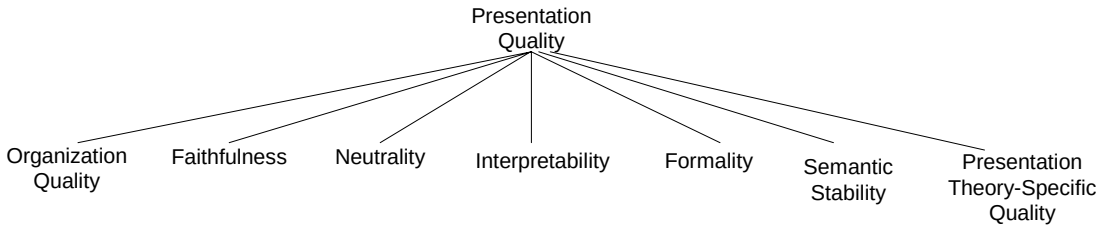

Figure 6. The Measurement of Presentation Quality

Application Quality

From presented data to utilized data, the most frequent quality problems are technical and social constraints that prevent data from being efficiently and effectively utilized. Strong et al. [19] listed four out of ten potholes that are related to such constraints. Regarding technical constraints, they noted that, due to lack of appropriate technology, it might be difficult to handle (index and analyze) non-numerical information presented in charts, images, and audio and video recordings. Also due to the limitation of computer and network resources, it is difficult to access sufficient data for information intensive tasks such as data mining. Regarding social constraints, Strong et al. [19] noted that easy access to information might conflict with requirements for security, privacy, and confidentiality. For example, patient medical records must be kept confidential and thus their usefulness for a research task may diminish. To account for technical and social constraints, we introduce three attributes to measure the quality of utilized data: ease of manipulation — how easy data can be indexed and analyzed, privacy — the extent to which a task has permissions to access the data, and security — the extent to which a task has secured access to the data. Besides technical and social constraints, information ageing is another significant cause to poor DQ; as

time passes and organizational environment changes, data may become outdated even though the task is still the same. We use Timeliness as an attribute to account for this aspect of DQ.

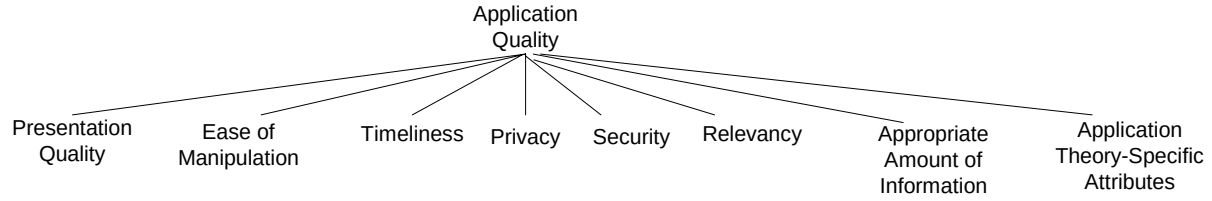

Figure 7. The Measurement of Application Quality

Different theories of application have different requirements on DQ. For example a linear programming model demands data robustness, i.e., a slight perturbation of data does not cause the change of the optimal solution. On the other hand, a regression analysis requires data to be normally distributed. Nevertheless, the following two criteria are fundamental to all theories: relevance and appropriate amount of information. Relevance refers to how data fit the needs of a task whereas appropriate amount of information stipulates that data are sufficient and necessary for the task, no more and no less. The latter also dictates that data be complete for the task. In sum, we propose a measurement model of application quality as shown in Figure 7.

Data Quality Evaluation

Unlike existing studies, we suggest using a different model to evaluate the quality of data at a different stage of a DELC. To evaluate data that are not organized and presented and do not serve a particular purpose, we should apply the model for collection quality. To evaluate stored data such as a database or a data warehouse, we should apply the model for organization quality. To evaluate presented data such as web sites or account statements, we should apply the model for presentation quality. Table 2 lists examples of data to which different DQ models may be applied. When selecting a right model for DQ evaluation, one should bear in mind that collected, organized, and presented data might not serve any particular application purpose. If there is no particular purpose, the task-related attributes are irrelevant. Similarly, if they are not presented, the attributes related to presentation are not applicable.

Definition	**Applicable Examples**
Collection Quality	Raw data, Surveys, Observations, Recordings,
Organization Quality	Data Files, Database, Data Warehouses
Presentation Quality	Web pages, Financial Reports, Account Statements
Application Quality	Research Data, Medical Diagnose Data

Table 2. Applicable Examples of Evolutional Data Quality

As we defined, four measurements of data quality are evolutional; when evaluating data at a later stage of a DELC, the quality measurement at the previous stage is considered. As far as exactly how it is considered, our conception does not provide a readily available answer. For example, when evaluating application quality, Figure 7 shows that we should consider presentation quality. However, do we consider presentation quality as one attribute of application quality or do we consider it as a multi-dimensional construct, which is further measured by the attributes in Figure 6? If we choose the latter, then how do we handle organization quality and collection quality? We believe the answer to these questions depends on the traceability of DQ evidence [1,16]. If there exists sufficient evidence from which one can evaluate the quality of data at an earlier stage, one tends to evaluate individual attributes for the earlier-stage quality measurement rather than the measurement as a combined value. However, when such

evidence is not available, using the combined value is reasonable. Our case here is analogous to the performance evaluation of computer products. When measuring his or her satisfaction with a computer, a user may consider its fitness for use and its ease of use, and then factor the quality of the computer as one combined value into the consideration. However, if the user has detailed knowledge about the quality of components being used in the computer, he or she may use that knowledge and assess the satisfaction by evaluating the qualities of critical component.

CONCLUSIONS AND DISCUSSIONS

In this paper we reviewed and summarized the existing conceptual studies on DQ and pointed out the divergence of the current research trends. To add to further dialogues, we proposed a notion of theory-specific DQ. Our conception is mostly influenced by the philosophy that there is no absolute truth (objective data) existing outside a theory. Our view has been fully or partially shared by many scholars in computer sciences, statistics, and other fields. For example, Shafer [15] used the term frame of discernment instead of sample space to emphasize the epistemic nature that a sample space is deliberately constructed according to our knowledge and opinion. Similarly, Deming [6] refused to the existence of so-called exact values or true values and argued that any data are the result of applying a given procedure.

We defined the concept of data evolution life cycle and explained that data typically evolve through four states. The conception has two implications. First, we need to use four different measurement models to assess the quality of data in each of the four states. Therefore, we proposed the notion of evolution data quality. Second, by using a fishbone diagram to illustrate the most frequent quality problems during data evolution, we were able to derive the attributes for the four measurement models. Comparing with intuitive, empirical, and case-based approaches of deriving attributes, our approach builds on one conceptual foundation and is more theoretically sound. Comparing with the ontological approach [22] ours is more systematic and capable of deriving all DQ attributes. Comparing with the analogical approach [12, 22, 24], our approach does not necessarily require viewing data as products and data evolution as product manufacturing. Instead, we consider data as theory-bounded and data exist and evolve as per the needs of theories.

Table 3 summarizes all attributes we listed in Figures 4-7 with their definitions. As we can see, most of them are similar to those proposed by existing studies. For example, the attributes for collection quality are very similar to those of intrinsic quality [25], of believability [24], and of integrity [3]. The attributes for presentation quality are similar to those for presentational DQ [25]. Similarly, application quality seems to be similar to the contextual DQ [25]. However, there are some differences. First, our attributes are clearly defined and semantically distinct; each pertains to a distinct aspect of DQ at a particular stage of a DELC. For example, we use trustworthiness to account for intentional falsification. It measures the credibility of the agent who collects the data. In contrast, attributes in existing models tend to be cross-stage and reflect overlapping root causes. For example, objectivity in Table 1 reflects not only biased sample selections but also biased data presentation; believability may inflate DQ by double-counting dimensions such as accuracy, objectivity, and reputation. In addition, attributes like integrity, credibility, or reputation overlap in meaning with each other and with other attributes such as accuracy and objectivity. That is why some feel the relationships between these attributes are debatable [3]. Similarly, attributes like existence [3] or meaninglessness [21] can be fully reflected by objectivity and trustworthiness. Including them again into a measurement model will double-count certain dimensions of DQ.

Second, we do not believe in the classification of intrinsic vs. extrinsic DQ. As we argued, DQ is theory-specific and thus is extrinsic by nature. For example, to be objective according to one theory say, a factorial design, may create an observation bias according to another theory say, a randomized design. Our view also underlines different definitions of some attributes from those in existing studies. For example, the literature defines accuracy as the extent to which data conforms to the true values. The problem is from where on the earth one may know the truth. If the true values are never known, the accuracy can

never be assessed. In contrast, our definition is that accuracy refers to the extent to which collected data are free of measurement errors. Defined as such, accuracy may be measured based on variability [6].

Third, unlike existing studies that advocates a universal and exclusive checklist, we suggest to apply relevant attributes to each state of data in a DELC. For example, to evaluate the quality of a web site, attributes such as relevance are irrelevant because the web site serves many viewers, who may bear a different application purpose in mind when surfing the site. By forcing a judgment on relevance, one confounds the purpose-specific variation into the measurement of DQ. Similarly, when assessing the quality of data stored in a database, attributes such as formality becomes irrelevant. Assessing formality without respect to a specific presentation will confound the variation of presentation into DQ.

Quality	Attributes	Definition
Collection Quality	Accuracy	The extent to which collected data are free of measurement errors.
	Objectivity	The extent to which the sample selected for observation is representative of a population.
	Trustworthiness of the collector	The extent to which the collector has integrity of not committing falsification.
	Completeness	All values that are supposed to be collected as per a collection theory are collected.
	Clarity	The extent to which data contain no fuzzy and ambiguous observations.
Organization Quality	Reliability of Data Clerks	The extent to which data entry clerks are able to avoid mistakes.
	Consistency	Different data in a database are logically compatible.
	Storage Efficiency	It takes less space to store data.
	Retrieval Efficiency	It is fast to find desired information.
	Navigability	One can navigate around the related information.
Presentation Quality	Semantic Stability	The same data have same meaning across time and space.
	Faithfulness	The extent to which the presented data are identical to the origin in meaning and precision.
	Neutrality	Data selected for presentation are not in favor of any particular opinion or purpose.
	Interpretability	Data have clear meaning.
	Formality	Data are presented concisely and consistently
Application Quality	Ease of Manipulation	The extent to which data can be processed easily (e.g., indexed and analyzed).
	Timeliness	The extent to which data are sufficiently up-to-date for a task.
	Privacy	The extent to which a task has permissions to access the data.
	Security	The extent to which a task has secured access to the data.
	Relevancy	The extent to which data are applicable and useful for a specific theory.
	Appropriate Amount of Data	The extent to which the volume of information is appropriate for a specific theory.

Table 3. A Summary of Attributes for Evolution Data Quality

There are a few future research directions we may take to advance the research on DQ in general and the concept of theory-specific DQ in particular. First, even though we deliberately developed our measurement attributes to ensure the semantic and discriminant validities, there is still a need to empiri-

cally validate such properties. Two constructs (attributes) that we thought are distinct might well be the reflection of one fundamental construct. It is also possible that a certain attribute is a multidimensional construct so that it needs to be divided into two or more sub constructs. In any case, by following the definitions as in Table 3, we can develop measurement indicators and use statistic tools such as exploratory or confirmatory factor analysis to assess the efficacy of our measurement models. Second, based on the fishbone cause-effect diagram, we deliberately developed our measurement models so that DQ may be assessed using the evidential reasoning approach (for a review see [18]). For example, we may assess the trustworthiness of a data collector, the reliability of a data entry clerk, relevance, and clarity using subjective beliefs. We may evaluate the accuracy and semantic stability using sample statistics. We may also measure attributes like completeness and storage efficiency using direct observations. By pooling all the evidence together, we can then apply Dempster's rule of combination [15] to obtain an overall assessment of DQ. Finally, we may alternatively regard the problem of assessing DQ as a multi-criterion decision-making problem and use the analytical hierarchical process [13] to actually derive the relative weights of the attributes in the measurement models. With these weights, we can either use a measurement model to assess the quality of certain data. We can also empirically examine inter-person, inter-task, and inter-stage differences of the weights in order to understand what factors influence DQ judgments and the selection of measurement attributes.

REFERENCE

[1] Ballou, D. P. and Pazer, H. L, Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science* 31 (2) 1985, pp.150-162.

[2] Ballou, D. P., Wang, R. Y., Pazer, H. and Tayi, G.K., Modeling Information Manufacturing Systems to Determine Information Product Quality, *Management Science*, 44 (4) 1998, pp.462-484.

[3] Bovee, M., Srivastava, R. P., and Mak, B., A conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality, *Technical report. Forthcoming*, (2002).

[4] Chin, W.W., Issues and opinions on structural equation modeling. *MIS Quarterly*, 1998.

[5] Churchill, G. A, A Paradigm for Developing Better Measures of Marketing Constructs, *Journal of Marketing Research* 16 (1979), pp.64-73.

[6] Deming, W. E., *Out of The Crisis*, Massachusetts Institute of Technology Center for Advanced Engineering Study, 1986.

[7] Fisher, C. W. and Kingma, B. R., Criticality of Data Quality as exemplified in two disasters, *Information & Management* 39 (2001), pp.109-116.

[8] Kahn, B. K., Strong, D.M., and Wang, R.Y., Information Quality Benchmarks: Product and Service Performance, *Communications of the ACM*, 45 (4) 2002, pp. 184-193 .

[9] Lee, Y. W., Strong, D.M.,, Kahn, B.K., and Wang, R. Y., AIMQ: A Methodology for Information Quality Assessment, *Technique Report*, forthcoming, 2002.

[10] National Research Council, *Funding a Revolution: Government Support for Computing Research*. National Academy Press, Washington D.C., 1999.

[11] Pipino, L.L., Lee, Y. W., and Wang, R. Y., Data Quality Assessment, *Communications of the ACM* 45 (4) 2002, pp. 211-218.

[12] Redman, T. C., *Data Quality: Management and Technology*. New York: Bantarn Books, 1992.

[13] Saaty, T.J., *The Analytical Hierarchy Process*, McGraw-Hill, 1980.

[14] Segars, A., Assessing the Unidimensionality of Measurement: a Paradigm and Illustration Within the Context of Information Systems Research, *Omega*, 25 (1) 1997, pp.107-121.

[15] Shafer, G., *A Mathematical Theory of Evidence*. Princeton: Princeton University Press, 1976.

[16] Shankaranarayanan, G., Wang, R. Y., and Ziad, M., IP-MAP: Representing the Manufacture of an Information Product, *Proceedings of the 2000 Conference on Information Quality*, 2000.

[17] Shannon, C.E., and Weaver, W. *The Mathematical Theory of Com-munication.* University of Illinois Press, Urbana, Ill. 1949.

[18] Srivastava, R. P., and Liu, L., Applications of Belief Functions in Business Decisions: A Review, to appear in *Information Systems Frontier*.

[19] Strong, D.M., Lee, Y. W., and Wang, R. Y., 10 Potholes in the Road to Information Quality, *Computer*, August 1997, pp.38-46.

[20] Strong, D.M., Lee, Y. W., and Wang, R. Y., Data Quality in Context, *Communications of the ACM*. (40) 3, 1997. pp.103-110.

[21] Te'eni, D., Behavioral Aspects of Data Production and their Impact on Data Quality, *Journal of Database Management*, vol. 4 (2) 1993, pp.30-38.

[22] Wand, Y., and Wang, R.Y., Anchoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM*, 39 (11) 1996, pp.86-95.

[23] Wang, R. Y., Kon, H.B., and Madnick S. E., Data Quality Requirements Analysis and Modeling, In *Proceedings of the 9th International Conference on Data Engineerin*g, Vienna: IEEE Computer Society Press, 1993.

[24] Wang, R.Y., Reddy, M.P., and Kon, H. B., Toward quality data: An attribute-based approach, *Decision Support Systems* 13, 1995, pp. 349-372.

[25] Wang, R.Y. and Strong, D. M. Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information System*, 12 (4) 1996, pp.5-28.