Anchoring Data Quality Dimensions in Ontological Foundations

Yair Wand and Richard Y. Wang

POOR DATA QUALITY CAN HAVE A SEVERE IMPACT ON THE OVERALL EFFECTIVENESS of an organization. A leading computer industry information service firm indicated that it "expects most business process reengineering initiatives to fail through lack of attention to data quality." An industry executive report noted that more than 60% of surveyed firms (500 medium-size corporations with annual sales of more than $20 million) had problems with data quality. The *Wall Street Journal* also reported that, "Thanks to computers, huge databases brimming with information are at our fingertips, just waiting to be tapped. They can be mined to find sales prospects among existing customers; they can be analyzed to unearth costly corporate habits; they can be manipulated to divine future trends. Just one problem: Those huge databases may be full of junk. . . .In a world where people are moving to total quality management, one of the critical areas is data."

The quality of a product depends on the process by which the product is designed and produced. Likewise, the quality of data depends on the design and production processes involved in generating the data. To design for better quality, it is necessary first to understand what quality means and how it is measured.

Data quality, as presented in the literature, is a multidimensional concept. Frequently mentioned dimensions are accuracy, completeness, consistency, and timeliness. The choice of these dimensions is primarily based on intuitive understanding [4], industrial experience [10], or literature review [13]. However, a literature review [25] shows that there is no general agreement on data quality dimensions.

Consider accuracy which most data quality studies include as a key dimension. Although the term has an intuitive appeal, there is no commonly accepted definition of what it means exactly. For example, Kriebel [13] characterizes accuracy as "the correctness of the output information." Ballou & Pazer [4] describe accuracy as "the recorded value is in conformity with the actual value." Thus, it appears the term is viewed as equivalent to correctness. However, using one term to define the other does not serve the purpose of clearly defining either. In short, despite the frequent use of certain terms to indicate data quality, there does not exist a rigorously defined set of data quality dimensions.

To design information systems that deliver high-quality data, the notion of data quality must be well understood. An ontologically based approach to defining data may be the ticket of success in real-world systems.

Clearly, the notion of data or information quality depends on the actual use of data. What may be considered good data in one case (for a specific application or user) may not be sufficient in another case. For example, analysis of the financial position of a firm may require data in units of thousands of dollars, whereas auditing requires precision to the cent. This relativity of quality presents a problem. The quality of the data generated by an information system depends on the design of the system. Yet, the actual use of the data is outside of designer's control. Thus, it is important to provide a design-oriented definition of data quality that will reflect the intended use of the information.

The need to evaluate data in quantitative terms has long been recognized. Related work can be characterized as theory-based or design-oriented.[1]

Two theoretical approaches are particularly relevant to data characteristics: *communication theory* and *information economics.* In their mathematical theory of communication, Shannon and Weaver [17] provide a probabilistic treatment of noisy transmission. In a noisy channel the signal resulting from a given message and the message that originated a known signal are uncertain. The uncertainty of the originating message of a received signal is termed "equivocation." Communication theory only deals with the transmission of signals and uses the word "information" in the specific meaning of the freedom to choose messages [15, p. 109]. It does not relate to the use of the transmitted signals. In contrast, information economics [9, 15, 16] seeks to evaluate information in terms of its use. An information system is modeled as a mapping from events in the world to signals. Users take actions based on the signals provided by the system. The value of information is given in terms of the outcomes of user actions based on the information. This approach enables the comparison of various information systems in terms of their value for the users.

Both communication theory and information economics provide formal treatments. However, neither addresses the notion of data quality in the context of systems design. In contrast, design-oriented approaches to data quality intend to provide actual guidance to system designers [23]. These approaches study characteristics of data in information systems in terms of actual design and implementation concepts such as entities, attributes and values. Such approaches can be termed "data-centric" as they focus on the structure and values of the data in a system. Although pragmatic, they have two main shortcomings. First, they do not derive data quality dimensions from fundamental principles. Second, since these approaches rely on specific data design concepts, they implicitly assume the detailed design should be known before data quality needs can be specified. Thus, they do not support early specification of data quality requirements.

This article analyzes data quality in terms that are not data-centric yet are oriented towards system-design. Specifically, we suggest rigorous definitions of data quality dimensions by anchoring them in ontological foundations; and we show how such dimensions can, in principle, provide guidance to systems designers on data quality issues.

We base our approach on the notion the role of an

[1] Accounting literature has always extensively addressed issues such as reliability, relevence, timeliness, and accuracy [9, 11, 19]. However, the primary interest there is in the accounting function such as auditing rather than information systems design.

information system is to provide a representation of an application domain (also termed the real-world system) as perceived by the user. Representation deficiencies are defined in terms of the difference between the view of the real-world system as inferred from the information system and the view that is obtained by directly observing the real-world system. From various types of representation deficiencies, we derive a set of data quality dimensions. Thus, in our approach users' views serve as a standard against which data quality is defined.

To base data quality concepts on the role of an information system as a representation, we need to define what is directly observed in the real-world system, and how an information system acts as a representation of the real-world system. The subject of ontology encompasses what is in the world.[2]

Foundations of the Data Quality Model

We begin by making a distinction between the external and internal views of an information system [20]. The external view is concerned with the use and effect of an information system. It addresses the purpose and justification of the system and its deployment in the organization. In the external view, an information system is considered "given," that is, a black box with the functionality necessary to represent the real-world system.

In contrast, the internal view addresses the construction and operation necessary to attain the required functionality, given a set of requirements which reflect the external view. System construction includes design and implementation. System operation includes activities involved in producing the data such as data capture, data entry, data maintenance, and data delivery. For simplicity, we assume perfect implementation because, for our purposes, a faulty implementation is equivalent to a faulty design with a perfect implementation. Thus, our analysis concentrates on the internal view, and is oriented towards system design and data production. This has two important implications. First, since the internal view is use-independent, it supports a set of definitions of dimensions of data quality that are comparable across applications. Hence, these dimensions can be viewed as being intrinsic to the data. Second, this view can, in principle, be used to guide the design of an information system with certain data quality objectives.

The distinction between the external and internal views should not be interpreted as a sequential systems development process. Rather, it intends to establish the designer, having no control of users requirements, should take the requirements as given at any time during development. It is possible, that system designers and users will cooperate in an iterative design process as needed.

Fundamental Principles

Our model is based on four assumptions. The first establishes the purpose of an information system:

1. The Representation Assumption: An information system is a representation of a real-world system as perceived by users.

The view of an information system as a representation is not new. For example, Kent [12] states that "an information system . . . .is a model of a small, finite subset of the real world." This view might seem somewhat restricted because it does not relate to the social and organizational aspects of information systems (e.g., [14]). Moreover, it assumes that all relevant knowledge about the real-world system should be represented in the information system. This contradicts the notion that some events cannot be anticipated, yet information about them might be very useful [7]. However, for our objective of defining data quality dimensions for design purposes, it is sufficient to consider the information system in its role as a representation of known aspects of the real-world.

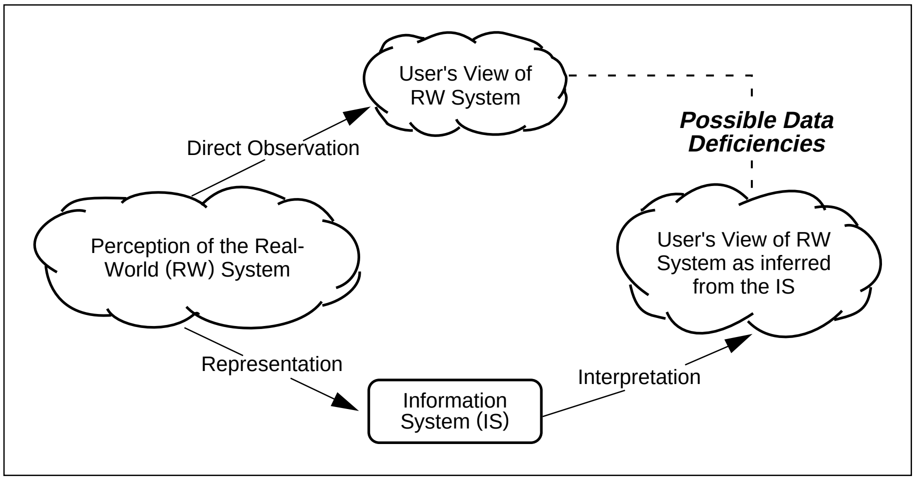

User's View of RW System
Possible Data Deficiencies
Direct Observation
Perception of the Real-World (RW) System
User's View of RW System as inferred from the IS
Representation
Interpretation
Information System (IS)

Figure 1. Possible data dificiencies in the data quality model

The development and use of an information system involve two transformations: the representation transformation (*rep* for short) and the interpretation transformation (*int*) [21]. The representation transformation deals with creating a representation of a view of the real-world system. This includes creating the information system and populating it with data. The interpretation transformation is the use of the information system to infer a view of the represented real-world system.

Information systems users may not be those involved

[2] Angeles, in *Dictionary of Philosophy*, defines ontology as "That branch of philosophy which deals with the order and structure of reality in the broadest sense possible."

in defining the requirements for the information system. Hence, to assure that the interpretation transformation will be able to reproduce the original view of the real-world system we introduce:

2. The Interpretation Assumption: An information system is built for use by the user whose view of the real-world system is captured in the design of the system.

For the information system to function properly, both the representation and interpretation transformations need to be performed flawlessly. This is the basis for our definition of data deficiency (Figure 1).

THE QUALITY of data depends on the DESIGN AND PRODUCTION PROCESSES involved in GENERATING THE DATA. To design for better quality, it is necessary first to understand WHAT QUALITY MEANS and HOW IT IS MEASURED.

Definition 1. A data deficiency is an inconformity between the view of the real-world system that can be inferred from a representing information system and the view that can be obtained by directly observing the real-world system.

The interpretation transformation can be decomposed into two processes. First, the information system creates a perceptible representation (most commonly, but not solely, a visual display). Then, the user should be able to perform the required inference about the real-world system. (A 'user' can, in principle, be a human-being or a machine.) However, the user's ability is beyond the control of the system designer, and therefore, beyond the scope of our model. Hence, we separate interface-related issues from our model:

3. The Inference Assumption: The information system can create a perceptible representation from which the user can infer a view of the real-world system as represented in the information system.

Finally, we confine our model to system design and data production aspects by excluding issues related to use and value of the data:

4. The Internal View Assumption: Issues related to the external view such as why the data are needed and how they are used are not part of the model.

This assumption does not imply that use and value are unimportant, but rather that data quality in our model is specified with respect to a given set of requirements that, we assume, capture the true intentions of the users.

Two notes are in order. First, according to our assumptions data quality dimensions are relative to user requirements. Second, although we assume that user requirements are given, this still allows possible requirements changes during the design process.

Since models of the world are the domain of ontology, we base our analysis on ontological constructs. The fundamental ontological concepts that we use and their application to information systems have been addressed in detail elsewhere ([6, 20, 22]). Here we summarize only the main concepts needed for our analysis.

Ontological Concepts

The world is made of things that possess properties. A thing can be a composite— made of other things. Properties are represented as attributes which are characteristics assigned to things by humans, depending on purpose and experience. The values of the attributes at any given time comprise the state of the thing.[3]

The knowledge of a thing is captured in terms of its states. Not all combinations of values of attributes are possible. There are laws that limit the allowed states to the lawful state space. An information system is also a thing. To be a good representation of a real-world system, the lawful states of the information system should reflect the lawful states of the real-world system.

Postulate 1. Things are modeled in terms of their states and laws.

[3] The state concept is also used in a semiotic approach to information systems [18].

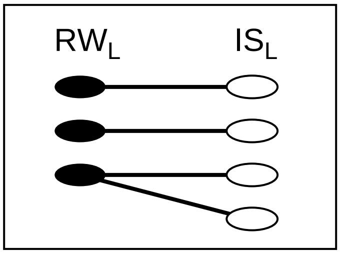


Figure 2. Proper representation

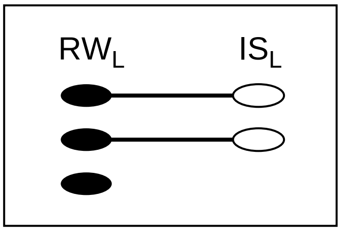


Figure 3. Incomplete representation

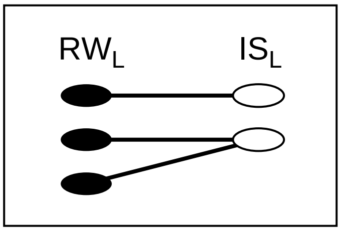


Figure 4. Ambiguous representation

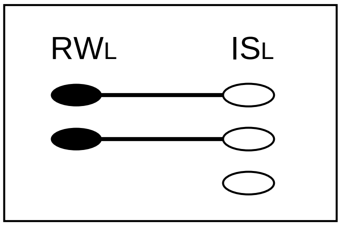


Figure 5. Meaningless state

Postulate 2. The real-world system is a thing, described in terms of its states and laws.

Postulate 3. An information system is a thing, described in terms of its states and laws.

Usually, a system is not viewed just as a whole, but its components are also of interest, hence:

Postulate 4: A system can be described as a composite made of other things.

Each of the components of a system is a thing, again modeled in terms of its states and laws. There is a connection between the states of components and the states of the whole system:

Postulate 5. Let the components of a system with a state space S be $\{X_1,...,X_N\}$ with state spaces $\{S_1,...,S_N\}$ respectively. There exists an exhaustive and one to many mapping: $S \rightarrow S_1x...xS_N$ (every element in S has at least one counterpart in $S_1x...xS_N$).

Next we formalize the notion of an information system as a representation of a real-world system:

Definition 2. An information system is said to be a representation of a real-world system if observing the state of the information system at a given time enables the inference of a state of the real-world system (at the same or another time).

We now define the link between the data-oriented and ontological views of information systems:

Postulate 6. The data stored in an information system at a certain time represent the state of the information system at that time.

This is consistent with the well-accepted database concepts. For example,

> "The data in the database at a particular moment in time is called a database state. In a given data base state. . . .Every time we insert or delete a record, or change the value of a data item, we change one state of the database into another state [8]."

Finally, we note two implications of our model. First, the notion of state does not rely on precise values, hence, the model can accommodate qualitative ("soft") data. Second, we assume the granularity of states reflects exactly the user's needs. States that are equivalent from the user's point view are combined into one state (e.g., [15]).[4]

Deriving Data Quality Dimensions

We begin by identifying the criteria for a real-world system to be properly represented by an information system. Based on this, we identify possible representation deficiencies that can occur during system design and data production. These deficiencies are used to define intrinsic data quality dimensions.

Let RW_L denote the lawful state space of a real-world system, and IS_L that of an information system representing this real-world system. Recall the representation and interpretation transformations. These transformations imply that two mappings must exist: A mapping from RW_L to IS_L, Rep: $RW_L \rightarrow IS_{L,}$ and a mapping from IS_L back to RW_L, Int: $IS_L \rightarrow RW_L$.

For a real-world system to be properly represented, two conditions must hold (Figure 2). First, every lawful state of the real-world system should be mapped to at least one lawful state of the information system (a real-world state can be mapped into multiple information system states). Second, it should be possible, in principle, to map an information system state back to the "correct" real-world state.

[4] Combining states whose difference is significant reflects the concept of materiality in accounting where insignificant events should not be recorded (e.g., [3, p. 70] who compare this practice to the legal notion *minimis non curat lex*, or trivial matters will not be considered).

Definition 3. A real-world system is said to be properly represented if: (1) there exists an exhaustive mapping, Rep: $RW_L \rightarrow IS_L$, and (2) no two states in RW_L are mapped into the same state in IS_L (the inverse mapping is a function).

Our analysis of data deficiencies is based on deviations from the conditions of Definition 3. We distinguish deviations due to system design flaws from those due to data production (system operation) flaws.

Definition 3 treats states in RW_L and IS_L as a whole, similar to considering the total data in a database at a particular moment as the database state. In practice, it is common to decompose the model of the real-world, i.e., view it as an aggregate of things and to decompose the information system to represent these components. By Postulate 5, RW_L can be viewed as a subset of the outer product of the components' state spaces. Correspondingly, a database state can be viewed at the global or at the component (e.g. entity, or object) level. Unless explicitly mentioned, our analysis applies to both the global and the decomposed views. However, it will be shown that the decomposition of RW_L and IS_L can generate special cases of representation deficiencies.

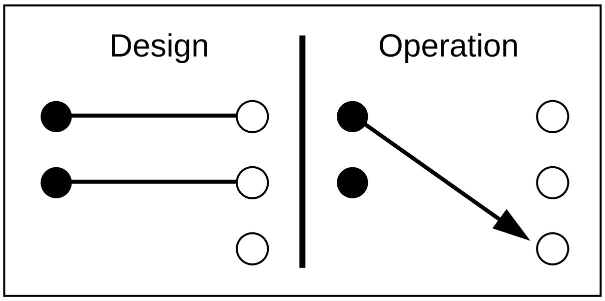
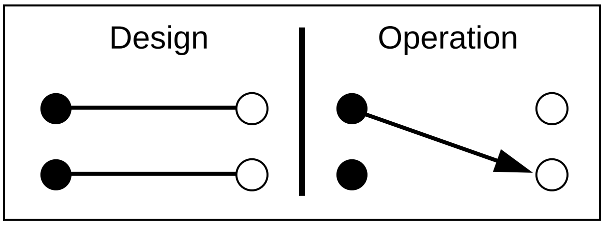


Figure 6. Garbling (map to a meaningless state)

Figure 7. Garbling (map to a wrong state)



Design Deficiencies

Based on our proper representation definition, we identify three generic categories of design deficiencies: *incomplete representation, ambiguous representation,* and *meaningless states.*

Incomplete representation: For an information system to properly represent a real-world system, the mapping from RW_L to IS_L must be exhaustive (i.e., each of the states in RW_L is mapped to IS_L). If the mapping is not exhaustive, there will be lawful states of the real-world system that cannot be represented by the information system (Figure 3). We term this "incompleteness." An example is a customer information system design which does not allow a non-U.S. address (a lawful state of the real-world system) to be recorded.

Ambiguous representation: For a proper representation no two states of the real-world should be mapped into the same state of the information system. If several states in RW_L are mapped into the same state in IS_L, there is insufficient information to infer which state in RW_L is represented. We term this situation "ambiguity" (Figure 4). A typical case of ambiguity is when there is insufficient number of digits to represent some states of the real-world system. This is usually viewed as a precision problem. However, we consider it a special case of ambiguity which is more general as it relates to any type of data, not just to numeric values. For example, a system design may allow only for one telephone number, without indicating whether it is the office or home telephone.

Meaningless states: It is not required that the mapping from RW_L to IS_L be exhaustive with respect to IS_L. However, when this situation exists, there are lawful states in IS_L that can not be mapped back to a state in RW_L (Figure 5). Such states are termed meaningless states. An information system design with meaningless states can still represent a real-world system properly. However, it is not a good design as it allows, in principle, meaningless data. For such meaningless data to materialize, some operational failure will have to occur.

We have identified two main design deficiencies, corresponding to the two conditions of proper representation (Definition 3): The representation mapping (from RW_L to IS_L) is not exhaustive, and the representation mapping is many-to-one. We also identified a potential deficiency when the mapping does not exhaust IS_L. Consider the fourth case: the representation mapping is one-to-many. Having multiple representations of a real-world situation may or may not be detrimental, depending on the user's cognitive style, and on the purpose of the system. These issues are not subject to designer's decisions and therefore we do not consider this a design deficiency.

Operation Deficiencies: Garbling

At operation time, a state in RW_L might be mapped to a *wrong* state in IS_L. We refer to this as garbling, and distinguish between two cases: If there exist meaningless states of the information system, the mapping might be to a meaningless state, and the mapping might be to a meaningful, but incorrect information system state. In the first case the user will not be able to map back to a real-world state (Figure 6). In the second case the user will be able to infer back, but to an incorrect state of the real-world (Figure 7). Typically, garbling occurs due to incorrect human actions during system operation (e.g., erroneous data entry, or failure to record changes in the real world).

Note, our analysis of design and operational flaws does not encompass the case where the user perceives a "wrong" state of the real world (either by error or due to malicious intent). This is because the information system is only required to enable mapping into perceived states, not "real" states.

Table 1. Intrinsic data quality dimensions

D.Q. Dimension	Nature of Associated Deficiency	Source of Deficiency
Complete	Improper representation: missing IS states	Design failure (Figure 3)
Unambiguous	Improper representation: multiple RW states mapped to the same IS state	Design failure (Figure 4)
Meaningful	Meaningless IS state and Garbling (map to a meaningless state)	Design failure (Figure 5) and Operation failure (Figure 6)
Correct	Garbling (map to a wrong state)	Operation failure (Figure 7)

Decomposition-Related Deficiencies

When an information system is decomposed, it is possible that each component of the system will act as a proper representation of a component of the real-world system, yet the joint representation be deficient. We identify three cases: The joint state of the information system represents a lawful but incorrect state of the real world (incorrect state); the joint state of the information system does not represent a lawful state of the real world (meaningless state); and the information system state corresponds to two or more states of the real world (ambiguity).

Table 3. Data quality dimensions as related to the internal or external views

	Dimensions
Internal View (design, operation)	***Data-related*** accuracy, reliability, timeliness, completeness, currency, consistency, precision ***System-related*** reliability
External View (use, value)	***Data-related*** timeliness, relevance, content, importance, sufficiency, usableness, usefulness, clarity, conciseness, freedom from bias, informativeness, level of detail, quantitativeness, scope, interpretability, understandability ***System-related*** timeliness, flexibility, format, efficiency

We use an example to demonstrate how these deficiencies can occur. Consider a risk management information system used by an investment firm that operates in several stock markets. The system is made of components (subsystems) each reflecting the firm's position in a market.

First, suppose that transactions occurred in two markets, but that one transaction was not reported by the time the other transaction was reported. Then both components would be in lawful states, and the joint state would be lawful but incorrect (garbling to a wrong state).

Second, assume there exists a supervisory mechanism which prevents traders in all circumstances from exceeding a maximum allowed exposure. Suppose that a trader invested in one market and sold in another without exceeding the total allowed exposure. Assume also the component for the market in which the trader purchased was updated before the component for the market in which the trader sold. There will be a period in which the global information system state might show a total balance higher than the allowed exposure, namely, an unlawful state of the trading system (garbling to a meaningless state).

Table 2. Notable data quality dimensions

Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Third, assume exchange rates for various countries are stored in all the subsystems. If the exchange rate for a certain currency is updated in one subsystem but not in another subsystem, then different values could be inferred (ambiguity).

This analysis provides some insight into how these three cases of deficiencies might happen. In particular, it demonstrates that because of the timing of updating the state of an information system, decomposition-related inconsistencies may occur even when all components operate properly.

Defining Intrinsic Data Quality Dimensions

Based on the analysis of the representation mapping from states of the real-world system to states of the information system ($RW_L \rightarrow IS_L$) we identified four potential representation deficiencies. As a consequence of these deficiencies, information system states can be incomplete, ambiguous, meaningless, or incorrect.[5] According to our assumptions, data represent the information system state (Postulate 6). Accordingly, we propose a set of four intrinsic data quality dimensions as shown in Table 1.

Analysis of Dimensions

We first summarize (Table 2) the most often cited[6] data quality dimensions based on a comprehensive literature review [25].

These dimensions can be categorized based on the definitions of internal and external views (Table 3). Since we exclude interface issues from our model, we include them in the external view. Table 3 also indicates whether a dimension is related to the data or to the system. Note, timeliness appears as related to both the internal and external views. Furthermore, timeliness and reliability appear to be both data and system-related.

Here we analyze those data quality dimensions from the literature that we identified as the internal view.

Accuracy and Precision: As indicated in the introduction, there is no exact definition for accuracy. In terms of our model we propose that inaccuracy implies that information system represents a real-world state different from the one that should have been represented. Therefore, inaccuracy can be interpreted as a result of garbled mapping into a wrong state of the information system.

Moreover, inaccuracy can be related to other data deficiencies identified in our model. First, ambiguity can lead to inference of the wrong state of the real-world system. Lack of precision is a case which is typically viewed as inaccuracy, but is ambiguity in our model. Second, incompleteness may cause choice of a wrong information system state during data production, resulting in incorrectness.

Note that inaccuracy refers to cases where it is possible to infer a valid state of the real world, but not the correct one. This is different from the case of meaningless states where no valid state of the real world can be inferred.

Reliability: Reliability has been linked to probability of preventing errors or failures [11], to consistency and dependability of the output information [13], and to how well data ranks on accepted characteristics [1]. In addition, reliability has been interpreted as a measure of agreement between expectations and capability [5], and as how data conforms with user requirements or reality [1]. It is clear there is no generally accepted notion of reliability and that it might be related either to characteristics of the data or of the system. However, one interpretation—that reliability indicates whether the data can be counted on to convey the right information—can be viewed as correctness of data in our analysis.

Timeliness and Currency: Timeliness has been defined in terms of whether the data is out of date [4] and availability of output on time [13]. A closely related concept is currency which is interpreted as the time a data item was stored [24].

Timeliness is affected by three factors: How fast the information system state is updated after the real-world system changes (system currency); the rate of change of the real-world system (volatility); and the time the data is actually used. While the first aspect is affected by the design of the information system, the second and third are not subject to any design decision.

In our model, timelines refers only to the delay between a change of the real-world state and the resulting modification of the information system state. Lack of timeliness may lead to a state of the information system that reflects a past state of the real world. Whether this matters or not, depends on the use of the data and is therefore in the external view. However, there is one effect of timeliness which can lead to data deficiencies independent of the use of the data, and is in the designer's domain. As our analysis of decomposition shows, wrong states, meaningless states, or ambiguous states may occur when the components operate properly, but are not updated at the same time.

Completeness: Generally, the literature views a set of data as complete if all necessary values are included: "All values for a certain variable are recorded" [4]. In our analysis, completeness is the ability of an information system to represent every meaningful state of the represented real world system. Thus, it is not tied to data-related concepts such as attributes, variables, or values. A state-based definition to completeness provides a more general view than a definition based on data, in particular, it applies to data combinations rather than just to null values. Also, it enables data items to be mandatory or optional depending on the values of other data items.

Consistency: In the literature, consistency refers to several aspects of data. In particular, to values of data, to the representation of data, and to physical representation of data. Details of internal representation or physical appearance of data are not part of our model. Hence, we only relate consistency to the values of data. Clearly, a data value can only be expected to be the same for the same situation. In terms of our model, different values can only occur if there is more than one state of the information system matching a state

[5] Note, we do not view inconsistency as a separate dimension, as it is manifested as ambiguity, lack of meaning, or incorrectness resulting from decomposition.

[6] Each appearance in a published article is counted as one citation. Thus, the result is biased in favor of the dimensions used by authors who have published extensively and authors whose articles have been quoted by others. However, as as indicator of the notable data quality dimensions, the result provides a reasonable basis for further discussion.

Table 4. Generic data quality problems

D.Q. Dimension	Mapping Problem	Observed Data Problem
Complete	Certain RW states cannot be represented	Loss of information about the application domain
Unambiguous	A certain IS state can be mapped back into several RW states	Insufficient information: the data can be interpreted in more than one way
Meaningful	It is not possible to map the IS state back to a meaningful RW state	It is not possible to interpret the data in a meaningful way
Correct	The IS state may be mapped back into a meaningful state, but the wrong one	The data derived from the IS do not conform to those used to create these data

of the real system. In this sense, inconsistency would mean that the representation mapping is one to many. As indicated, in our analysis this is not considered a deficiency.

Implications to Information Systems Design

We identified four intrinsic dimensions of data quality. Accordingly, we identify four generic types of data problems that can be observed in using an information system (Table 4): Loss of information, insufficient information (ambiguity), meaningless data and incorrect data.

The generic data quality dimensions were derived by analyzing possible failures of the representation transformation: ($RW_L \rightarrow IS_L$). Based on this analysis, we can identify the types of design actions that can be used to avoid or correct these problems (Table 5).

The first two deficiencies require modifications to the lawful state space of the information system or to the mapping into this space. Such decisions are, in principle, under designer's control. In contrast, meaningless and incorrect data result from operational failures (usually due to human actions). However, meaningless data can only occur when there exist meaningless states of the information system. The designer can reduce such states through the application of information system controls such as integrity constraints [5].

Table 5. Data deficiencies repairs

Observed Data Problem	Reason(s) for Deficiency	Repair
Loss of information	Missing lawful ststes of the information system	Modify *IS_L* to allow for missing cases
Insufficient information (ambiguous data)	Several states of the real world mapped into same state of information system	Change the mapping *$RW_L \rightarrow IS_L$* This may require adding states to *IS_L*
Meaningless data	(1) These are information system states that do not match real-world, and (2) Garbling	Reduce *IS_L* to include only meaningful states This can be done by adding integrity constraints
Incorrect data	Garbling	Design to reduce garbling This might be done by adding some controls

The situation is more complicated for incorrect data, as they result from incorrect mapping into meaningless information system states. However, automated mechanisms may still be used to reduce this problem. Assume the state space of the information system was increased by adding a large number of meaningless states. Then the probability that incorrect operation will result in a meaningless state rather than a meaningful state would increase. Meaningless states can be controlled by integrity constraints. Thus, some garbling might be prevented by artificially increasing the possible state space of the information system and adding controls. This approach is usually implemented by increasing the possible state space of the information system without increasing the lawful state space. Specific examples are the addition of a check digit to identification codes, and the use of control totals for transaction batches.

Concluding Remarks

Despite extensive discussion in the data quality literature, there is no consensus on what constitutes a good set of data quality dimensions and on an appropriate definition for each dimension. Even a relatively obvious dimension, such as accuracy, does not have a well established definition.

We have analyzed data quality based on inconformities between two views of the real world system: The view obtained by direct observation and the view inferred from the information system. The analysis generated four intrinsic (system-oriented) data quality dimensions. These dimensions specify whether data is complete, unambiguous, meaningful, and correct. Each of these dimensions is well-defined in terms of a specific deficiency in the mapping from real-world system states to information system states. Therefore, they can be used to reason about data quality. Such reasoning can be done for the purpose of improving data quality. Conversely, there are cases when low quality data can be advantageous. An example would be preventing an adversary from knowing true real-world states in national defense or commercial competition situations. Knowing the sources of the generic deficiencies can help plan for intentionally low quality data.

There are several directions in which this work can be extended. First, the current model provides rigorous definitions for reasoning about data quality, but does not provide concrete guidelines for systems designers.

Research needs to be conducted on how to operationalize these formally derived dimensions in terms usable in systems design practice. Second, the data quality dimensions can be used to develop data quality audit guidelines and procedures. Third, data quality metrics can be developed for use in specification and audit of information systems. Fourth, the dimensions identified can be used to guide data collection in field studies of data quality problems and practices. Also, they can be used to compare the outcomes of different studies.

Beyond these, we believe a rigorously defined set of data quality dimensions has a value in itself, in providing a common set of terms, and thus supporting the development of a cumulative body of work in the data quality area.

References

1. Agmon, N., and Ahituv, N. Assessing Data Reliability in an Information Systems. *J. of Manage. Info. Syst. 4*, 2 (1987), pp. 34–44.
2. Angeles, P.A. *Dictionary of Philosophy*. Harper Perennial, New York, 1981.
3. Anthony, R.N., and Reece, J.S. *Accounting: Text and Cases.* Richard D. Irwin, Homewood, Ill. 1979.
4. Ballou, D.P., and Pazer, H.L. Modeling data and process quality in multi-input, multi-output information systems. *Manage. Sci. 31,* 2 (1985), pp. 150–162.
5. Brodie, M.L. Data quality in information systems. *Info. Manage.* (1980), pp. 245–258.
6. Bunge, M. Ontology I: The furniture of the world. *Treaties on Basic Philosophy,* Vol. 3-4. Reidel Publishing, Boston, Mass. 1977, 1979.
7. Ciborra, C., Migliarese, P,. and Romano, P.A. Methodological inquiry of organizational noise in sociotechnical systems. *Human Relations 37,* 80 (1984), pp. 565–588.
8. Elmasri, R., and Navathe, S. *Fundamentals of Database Systems.* Benjamin/Cummings, Reading, Mass., 1994.
9. Feltham, G. The value of information. *Account. Rev. 43,* 4 (1968), pp. 684–696.
10. Firth, C.P., and Wang, R.Y. *Data Quality Systems: Evaluation and Implementation.* Cambridge Market Intelligence, London. 1996.
11. Hansen, J. V. Audit considerations in distributed processing systems. *Commun. ACM 26,* 5 (1983), pp. 562–569.
12. Kent, W. *Data and Reality.* North Holland, New York. 1978.
13. Kriebel, C.H. Evaluating the quality of information systems. *Design and Implementation of Computer Based Information Systems.* N. Szysperski and E. Grochla, Ed. Sijhtoff & Noordhoff, Germantown. 1979.
14. Land, F. The Information Systems Domain. *Information Systems Research — Issues, Methods and Practical Guidelines.* R. Galliers, Ed. Blackwell Scientific Publications, Oxford, England. 1992.
15. Marschak, J., and Miyasawa, K. Economic comparability of information systems. *Int. Econ. Rev. 9,* 2 (1968), pp. 137–174.
16. Marschak, J., and Radner, R. *Economic Theory of Teams.* Yale University Press, New Haven, Conn. 1972.
17. Shannon, C.E., and Weaver, W. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, Ill. 1949.
18. Stamper, R. *Critical Issues in Information Systems Research.* R.J. Boland and R.A. Hirschheim, Ed. John Wiley, New York, 1987.
19. Sterling, R.R. *Toward a Science of Accounting.* Scholars Book, Houston, Tex. 1979.
20. Wand, Y., and Weber, R. On the deep structure of information systems. *J. Info. Syst.* (1995), pp. 203–223.
21. Wand, Y., and Weber, R. On the ontological expressiveness of information systems analysis and design grammars. *J. Info. Syst. 3,* 3 (1993), pp. 217–237.
22. Wand, Y., and Weber, R. An Ontological Model of an Information System. *IEEE Trans. Soft. Eng.. 16,* 11 (1990). pp. 1282–1292.
23. Wang, R.Y., Kon, H.B., and Madnick, S.E. Data quality requirements analysis and modeling. In *Proceedings of the the 9th International Conference on Data Engineering.* (Vienna, Austria, 1993), pp. 670–677. 1993
24. Wang, R.Y., Reddy, M. P., and Kon, H.B. Toward quality data: An attribute-based approach. *Decision Support Syst.* (1995) pp. 349–372.
25. Wang, R.Y., Storey, V.C., and Firth, C.P. A framework for analysis of data quality research. *IEEE Trans. on Knowl. Data Eng. 7,* 4 (1995), pp. 623–640.

YAIR WAND (yair.wand@ubc.edu) is Professor and MIS Division Head at the University of British Columbia, Faculty of Commerce and Business Administration, in Vancouver. He has served on *Communications'* Editorial Board.

RICHARD Y. WANG (rwang@mit.edu)is Co-Director for Total Data Quality Management and associate professor at MIT Sloan School of Management. He is widely recognized as a leading authority in data quality research and practice.

This work was supported in part by grants from The Natural Sciences and Engineering Research Council of Canada, MIT's Total Data Quality Management Research Program, MIT's International Financial Services Research Center, and the U.S. Naval Command, Control, and Ocean Surveillance Center.

Nominations for ACM Eastern Regional Representative Election

Officers of permanent SIGs are elected by plurality vote of the SIG members for two-year terms beginning July 1 in odd-numbered years. At the end of the term, the SIG Board may exercise an option at the request of the SIG Officers to extend the terms for an additional two years. In accordance with this policy, the SIG Board voted on June 13 to approve the request to extend the terms of officers for the following groups: ARCH, COMM, CPR, IR, OPS, SAC, and UCCS. Elections will be held for any of these SIGs if a petition is brought forward by one percent of the members of the SIG. A petition to initiate an election must be received at ACM Headquarters Office of Policy and Administration by November 30. Questions about the petition process should be directed to Pat Ryan, Deputy Executive Director of Operations, at ACM Headquarters (+1-212-869-7440).