Information Processing & Management Vol. 30, No. 1, pp. 9–19, 1994
Printed in Great Britain.

0306-4573/94 $6.00 + .00
Copyright © 1993 Pergamon Press Ltd.

THE NOTION OF DATA AND ITS QUALITY DIMENSIONS

CHRISTOPHER FOX,[1] ANANY LEVITIN,[2] and THOMAS REDMAN[1]
[1]AT&T Bell Laboratories, Holmdel, NJ 07733, U.S.A.
[2]Villanova University, Villanova, PA 19085, U.S.A.

(*Received* 3 *January* 1992; *accepted in final form* 19 *October* 1992)

Abstract—The rapid proliferation of computer-based information systems is increasing the importance of data quality to both system makers and users. However, there is neither an established framework nor common terminology for investigating data quality. There is not even agreement on what the term "data" means. We lay a foundation for the study of data quality in this paper. In the first part of the paper we discuss five approaches to defining "data" in the literature. We then propose an approach especially conducive to discussing data quality. In the second part of the paper we discuss the most important dimensions of data quality: accuracy, completeness, consistency, and currentness. We define these four and several related dimensions and discuss them in detail. We close the paper by outlining several areas for further research on data quality.

1. INTRODUCTION

Nowadays, a great many human activities rely on data, often stored and manipulated in a computer. Despite the importance and ubiquity of data in modern society, there is a surprising lack of agreement about the meaning of the term "data." There are several explanations for this lack of agreement. First, the notion is deeper and more complex than it first appears. It is also fundamentally important in several branches of computer and information science whose theoreticians and practitioners use the term in their own ways. Finally, the variety of contexts where data appears makes generalization difficult. Nevertheless, we believe that "data" can be given a definition general enough to be useful in a wide range of disciplines, precise enough to capture the depth and complexity of the term, and useful in attacking practical data problems.

Our point of departure for studying "data" is concern with the very practical problem of data quality. Global competition, application of high technology, and continuous pressure for increased productivity have driven industry to improve its processes and products, reduce cycle times, eliminate waste and rework, and increase responsiveness to customers. Greatly improved data quality is essential for process and product design and improvement for many reasons, including the following:

1. Products and services are more and more dependent on computers, and hence data, than ever before. Progress in hardware and telecommunications, coupled with advances in software engineering, make feasible products and services that were pipe dreams only a few years ago. Such applications typically rely on huge volumes of rapidly changing data shared by many geographically dispersed users. Poor data quality in such systems renders them nearly useless. For example, telecommunications companies now offer dazzling products like cellular phones, super-reliable services, rapid service provisioning, instantly reconfigurable networks, and so on. These products and services depend on software driven by vast databases of facilities, circuits, customers, services, etc. These systems do not tolerate poor data quality like their hard-wired, slowly changing predecessors, but must have high-quality data if they are to work at all.

2. Once major process flaws have been corrected and process performance reaches high levels, further improvement may be limited by poor data quality, so data quality problems

Please address all correspondence to Thomas Redman, Room 2J-504, AT&T Bell Laboratories, 101 Crawford's Corner Road, Holmdel, NJ 07733, U.S.A.

come to the fore as other problems are resolved. For example, we have found organizations stymied in their business process improvement efforts by an inability to obtain high-quality data from their data suppliers.

3. The rapid proliferation of computer-based information systems has introduced an army of new and unsophisticated users to computers. Because they are often less well-trained, such users tend to feed systems erroneous data. Furthermore, inexperienced users are less able to recognize and deal with erroneous data than experienced users, and therefore can be victimized by it. These problems drive system and process data quality improvements to reduce error rates in data input, data output, and data processing, and to make processes more tolerant of data errors.

4. There is clear evidence in several publications (Wilson, 1992; Bendel, 1991; Laudon, 1986; Morey, 1982) that error rates in some existing databases are unacceptably high. For example, Laudon (1986) cites two FBI criminal history systems with record deficiency rates of 50% (for a computerized system) and 75% (for a manual system). Wilson (1992) reports a survey in which 70% of respondents claim to have had business disrupted by bad data.

Despite the importance of data quality, it has received little attention. Papers by Hoare (1975) and Brodie (1980) and a monograph by Naus (1975) were among the few early publications devoted exclusively to the subject. More recent contributions include a book by Tasker (1989), a collection of papers edited by Liepins and Uppuluri (1990), and surveys by O'Neill and Vizine-Goetz (1988) and Sy and Robbin (1990) on the quality of online databases, and U.S. government statistical data, respectively. Few educational or research institutions, and still fewer business organizations, have committed resources to studying and improving data quality. No common framework for studying data quality problems, nor an agreed-on terminology for discussing data quality, has emerged from the modest efforts to date. For example, Hoare (1975) uses the term "data reliability" as a synonym for "data quality," whereas Brodie (1980) states that "data quality has three components: data reliability, logical (or semantic) integrity, and physical integrity." AT&T Bell Laboratories is among the few organizations working in the area of data quality. The Information Quality Management and Technology group at Bell Labs (of which we are a part) studies data quality, and develops methods, techniques, and technologies for its assessment, control, and improvement. This paper reports some of the results of our foundational work on data and its quality dimensions. The paper is limited to discussion of data quality—information quality, and the relationships among data, information, and knowledge are beyond the scope of this paper. See Teskey (1989) or Fox *et al.* (1988) for discussions of these broader topics consistent with our approach. See Redman (1992) for a fuller treatment of data quality in the context of process improvement.

This paper has three major goals: first, we present an approach to the notion of data useful in understanding data quality issues. This approach has much in common with others already shown useful in data processing and computer science. Second, we identify and discuss dimensions of quality for data values, currently the focus of most data quality improvement efforts. Third, we list several related issues not treated in detail in this paper that must be addressed in continuing the work reported here. The final section summarizes the paper.

2. WHAT IS DATA?

In this section, we discuss several definitions of "data." Our main goal in discussing these definitions is not to uncover their weaknesses, but to use their strengths to develop an approach useful for advancing data quality. In particular, we seek an approach that meets two sets of criteria: first, that it lead to a definition that is linguistically adequate (Linguistic Criteria); second, that it provide a grounding for techniques to improve data quality (Usefulness Criteria).

Our Linguistic Criteria are the following:

- clear and simple: the approach should lead to a clear and simple definition of "data"; this is a requirement of any good definition;

- not mention information: the approach should not use the concept of information so that it avoids circular definitions; and
- agree with common usage: the definition should agree with our everyday use of the word "data," so we can claim to be defining "data" and not some other term.

Our Usefulness Criteria are the following:

- both conceptual and representational: the approach should reflect both the conceptual and representational facets of data, since both are important for data quality;
- widely applicable: the approach should apply to a wide range of cases, especially those involving computerized collections of data (databases), since those are the primary targets of data quality improvement; and
- quality dimensions: the approach should suggest dimensions important for data quality, leading to deeper understanding of data quality problems.

Most who have written about the notion of data have avoided giving a definition of the term and have used it informally, often as a synonym for "information." Nevertheless, dozens of definitions have been published. They generally follow one or a combination of the few approaches outlined below.

Some authors (Blumenthal, 1969; Fry & Sibley, 1976) simply follow the Latin origins of the term by defining "data" as a set of facts. Though these writers do not define "fact," usually a fact is understood to be a state of affairs, or a way that the world actually is. Facts are also the standard for truth and falsehood (a claim is true if and only if what is claimed is a fact). There is no question that data is collected and used in the hope that it accurately represents reality (we will discuss this point in more detail below). However, as the standard for truth, a fact cannot be false. This implies that false data cannot exist, though they certainly seem to. From the usefulness perspective, this definition has the drawback that it says nothing about the representational aspects of data. How is a set of facts to be symbolized and recorded? This definition does not help tackle such questions. We conclude that though data may be about facts, it is not a collection of facts.

Another approach defines "data" by indicating ways it can be obtained. Davis & Rush (1979, p. 193) define data as follows: "The simplest way of defining data is to say that it is the result of measurement or observation." Yovits (1981), in his article for the *Encyclopedia of Computer Science*, combines etymology with the same approach: "Data are facts or are believed to be or are said to be facts which result from the observation of physical phenomena." Measurement and observation *are* two important sources of data. (There is no need whatsoever to restrict ourselves to physical phenomena, though.) However, there are many common examples of data not obtained from measurement or observation. For example, someone's name, social security number, phone number, and so on, were assigned to that person, not observed or measured. The way data is obtained is of profound importance, but it does not follow that the *means* to get data define what data *is*.

Many information systems specialists define data as "the raw material from which information is developed" (Dorn, 1981). In other words, data is the input to some process whose "refined" output is information. As noted above, we do not think it is a good idea to define "data" in terms of "information." Further, this definition does not get us far—we are left wondering where data leaves off and information begins. This approach also requires bringing in some process that refines data into information, an unnecessary and difficult complication. Finally, this approach fails to meet our requirements that "data" be defined in a way that brings in its representational aspect, and that helps shed light on data quality problems. We thus leave information out of our discussion of data.

Another popular alternative is exemplified by Burch *et al.* (1983, p. 4): "Data are language, mathematical, and other symbolic surrogates which are agreed to represent people, objects, events, and concepts . . . The sound of a train whistle and a customer order are two examples of data." The strength of this approach is that it brings in the representational aspect of data; its principal weakness is that is disconnects data from its referent and restricts it to the representational level of symbols. This cannot be right, because the same

data can be represented in many different ways, which is not possible if data is just symbols. Even proponents of the "data are symbols" school, such as Langefors and Samuelson (1976, p. 110), admit "it *is* common to consider data as items that can be talked about without necessarily considering details of their stored format." Clear separation between the conceptual and representation aspects of data, called the *data independence objective* by Codd (1981), was a paramount breakthrough in data modeling. We conclude that data cannot be explained as symbols alone.

From the usefulness perspective, we do not find it fruitful to consider the sound of a train whistle as a datum even if the sound can be treated as a symbol of the train. Further, a customer order, with no context provided, should be regarded not as data but as an object, or a relationship between objects, about which we have data. This point is discussed further below.

The last example brings us to the approach developed by the database research community, though its elements can be found in such early publications as Mealy (1967). Before proceeding, two remarks are in order. First, though terms like "object," "entity," "relationship," "attribute," and so on, seem to be clear, they are full of ambiguities and unexpected traps (see Kent, 1978, for an unsurpassed discussion of the subject). Second, the many data models suggested over the years differ in the way the basic notions of object, entity, attribute, and so on, are defined, structured, and used. Therefore, our discussion considers only the most general ideas applicable in most of the important models.

Though we are usually interested in data about real world objects (physical or abstract) and relationships among objects, we work with an abstract *view*, or *model* of the world. A view yields *entities* that serve as model representatives of their real-world counterparts. Some specialists advocate a distinction between entities standing for objects and those standing for relationships among objects; the weight of opinion deems this unnecessary.

In the classical database approach, an *entity* is an element of an *entity class* (also called a *type*) defined by a set of *attributes*. For example, an entity class `EMPLOYEE` can be defined by the attributes `Employee-Number`, `Name`, `Data-of-Birth`, `Address`, `Department`, `Salary` and `Dependents`. Each attribute has a *domain*, a set of values permissible for the attribute in the model. There are no restrictions on the elements in a domain: they may be entities (e.g., departments) or sets of entities (e.g., dependents). If an attribute's values are a measurable quantity, units of measurement must be specified with its domain. Finally, for entities having no value for a particular attribute, a special element, called the *null value*, is added to the domain. (We discuss null values in more detail below.) Following the same logic, another special element can signify an illegitimate value.

Within the framework outlined above, Tsichritzis and Lochovsky (1982, p. 7) define a *datum* (or *data item*) as a triple $\langle e, a, v \rangle$ where the value v is selected from the domain of the attribute a to represent the attribute's value for the entity e. Then *data* can be defined as any collection of data items.

This definition has several attractive features. It recognizes that any data is based on a model of a part of the world (see Fig. 1). Consequently, full assessment of data quality must include assessment of the model supporting the data. (Unfortunately, however, in practice quality assessment is often limited to data values—see Loebl, 1990, for several

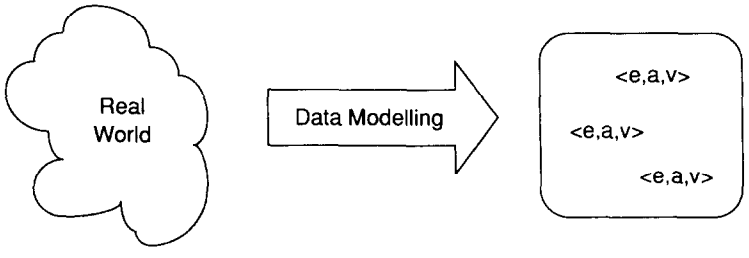

Fig. 1. The notion of data as a result of modelling.

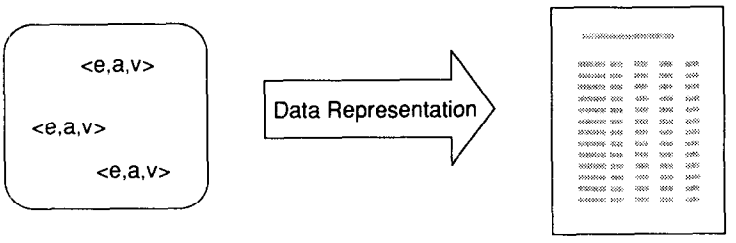

Fig. 2. Data representation.

interesting examples pertinent to this point.) Furthermore, by emphasizing that a data model is composed of entity classes, this definition reflects the repetitive nature of most data, which typically is values for several attributes of many entities of the same class. Finally, the discipline imposed by a model's entity classes and attributes (whose domains are predefined) suggests that data is a formally organized collection.

However, this definition does not go far enough; it fails to emphasize the distinction between the conceptual and representational aspects of data, as our usefulness criteria require. To correct this deficiency, we define a *data representation* as a set of rules for recording triples on some medium, and a *data recording* as a physical instance standing for a set of data items according to a data representation (see Fig. 2). For the value portion of the data, representation is usually done via a *format** and the recording via a *symbol*. For example, values of the attribute `Date-of-Birth` can be recorded in either the American format *mm/dd/yy* or the European format *dd/mm/yy*, among several natural choices.

Note that this approach does not confuse data with the facts that the data models. It does not make any requirements about how data is obtained. It avoids any mention of information, so it can be used to analyze "information" without any danger of circular definitions. Since the data and its representation are separate in this account, the same data can be represented in many ways, data represented in a certain way can be recorded many times, and data can exist without being represented or recorded. (Of course, usually it is both symbolized and recorded.) Thus this approach to the notion of data avoids the pitfalls discovered in the other definitions considered above.

This definition of data as a collection of triples $\langle e, a, v \rangle$, along with the definitions of data representation and data recording, meets our criteria for an adequate account of data. A particularly attractive feature of this approach for our purposes is that it leads to three sets of quality issues: those related to the quality of the model or view, those related to the quality of data values themselves, and those related to the quality of data representation and recording. Therefore we adopt these definitions for data, data representation, and data recording. Table 1 summarizes the definitions we have reviewed and whether they succeed or fail to meet our linguistic and usefulness criteria.

3. QUALITY DIMENSIONS OF DATA VALUES

In this section, we define and discuss the dimensions of data most pertinent to the quality of data *values*. Quality dimensions of models and data representations are beyond the scope of this paper, but are discussed elsewhere (Redman, 1992, Chapter 3).

We group dimensions of data quality involving data values into four categories: accuracy, currentness, completeness, and consistency. As mentioned above, data quality assessment should not be restricted to concerns about data values only, though these are the dimensions most often associated with data quality.

*Formally, a *format* for an attribute a can be defined as a function from the attribute's domain to a set of symbolic representations S (Redman, 1992, p. 58).

Table 1. Assessment of approaches to defining "data" according to the paper's six criteria (S = success; F = failure)

	Linguistic criteria			Usefulness criteria		
Approach	Clear & simple	Not mention information	Agree with common usage	Both conceptual & representational	Widely applicable	Quality dimensions
Set of facts	S	S	F	F	F	F
How obtained	F	S	F	F	S	F
Raw material of information	F	F	S	F	S	F
Symbols	S	S	F	F	S	F
Collection of triples	S	S	S	F	S	S
Representable triples	S	S	S	S	S	S

3.1 *Accuracy, precision, and reliability*

Accuracy of a datum refers to the degree of closeness of its value v to some value v' in the attribute domain considered correct for the entity e and the attribute a. (Sometimes v' is referred to as the standard.) If the datum's value v is the same as the correct value v', the datum is said to be *accurate* or *correct*.

For example, consider an `EMPLOYEE` entity (identified by the `Employee-Number` 314159) and the attribute `Year-of-Birth`. If the value of `Year-of-Birth` for employee 314159 is the year the employee was born, the datum is correct.

Accuracy may be expressed as a measure of inaccuracy. We can quantify inaccuracy for a single datum by computing the magnitude of the difference between the correct value v' and the value v of the attribute in the triple. For the example just considered, inaccuracy might be gauged by the difference between the year of birth v' and the value v for the `Year-of-Birth`. The accuracy of an entire database can be measured by finding the fraction of incorrect triples in the database.

The notions of accuracy and correctness are not as trivial as this simple example might suggest. The principal difficulty lies in determining the correct value. For one thing, it may not be uniquely defined (for example, some foreign names have alternative spellings). Sometimes, the correct value is undefined. This may happen when the entity component e of the triple is mistaken. Thus, for the above example, there may be no correct value for `Year-of-Birth` when it is impossible to determine the employee identified in the triple.

There are also difficulties in quantifying inaccuracy, even when the correct value is known. First, the attribute in question may resist quantification. Consider, for example, `Gender` or `Department` of an `EMPLOYEE`. Or consider names. Should inaccuracy be quantified by counting the correct letters? Or correct letters in correct positions? A more sophisticated method such as the Dice's coefficient (O'Neill & Vizine-Goetz, 1988, p. 140) could be used as well. Even the inaccuracy of numerical values may be tricky to quantify. We mentioned that the absolute value of the difference between an actual and stated year of birth can serve as a natural measure of `Year-of-Birth` inaccuracy. However, according to this measure, the error is the same for the value 1970 when the correct value is 1979 as it is for the value 1960 when the correct value is 1969. But employing the first person may violate a juvenile labor law—a fact that might be worth reflecting with a more sophisticated inaccuracy measure. To summarize, quantification of data inaccuracy is a nontrivial task even when it is possible. In particular, an inaccuracy measure may depend not only on the entity and attribute in question, but also on the application.

Precision refers to the measurement or classification detail used in specifying an attribute's domain. For example, measuring heights in inches is more precise than doing so in feet; similarly, having twenty possible values in the domain for the `Color` attribute may be more precise than having just seven (assuming both sets of colors are uniformly distributed across the color spectrum). Thus, precision depends on the domain structure and not on a particular datum; precision is associated with the model, not the datum. This

distinction clearly separates precision from accuracy. Also note that precision is not a characteristic applicable only to quantitative data.

Reliability is a term used in the literature in several ways. For example, Hoare (1975) uses the term as a synonym for overall quality, Chapple (1976) as a synonym for accuracy. Brodie (1980) considers data reliability to be one of three principal components of data quality (with logical and physical integrity being the other two), and gives two definitions of the term. His first definition is the following:

> Data reliability is a (statistical) measure of the extent to which a database can be expected to exhibit the externally observable structural properties specified for a database.

Though Brodie provides a brief example, neither his definition nor his example is clear. In particular, the term encompasses format correctness, integrity, and value correctness in some unspecified statistical sense. Brodie's second definition follows Parnas (1977) and defines reliability as "a measure of robustness (e.g., the absence of system failures)." But this definition encompasses the entire system rather than data, and hence goes beyond our topic.

The ANSI/ASQC *Standard A3-1978* (1978) defines (numerical) reliability as: "The probability that an item will perform a required function under stated conditions for a stated period of time." If one interprets "a required function" of a datum as veridical representation of a property of an entity by the value of its attribute, then the definition equates reliability with probability of correctness. (The influence of time is discussed below.)

3.2 *Currentness, age, and timeliness*

Our definition of data treats it as a static snapshot of (part of) the world. But most objects change with time. This has prompted proponents of the infological school (Langefors & Sundgren, 1975) to require each datum to have a time indicator. Despite the importance of time, we agree with Tsichritzis and Lochovsky (1982, p. 7) that is should not be required by the definition of "data." Though there are problems modeling time (Bolour *et al.*, 1982; Jardine & Matzov, 1988), it can be treated as an attribute. However, we do recognize the special relationship between change over time and data quality. The following definitions address this issue.

A datum is said to be *current* or *up-to-date* at time t if it is correct at time t. A datum is *out-of-date* at time t if it is incorrect at t but was correct at some moment preceding t. Currentness for a datum may be expressed as a measure of how far *out-of-date* the datum's value is. Currentness for an entire database can be measured by determining the fraction of out-of-date triples in the database. Thus, according to our definitions, being up-to-date is simply being correct at present, and being out-of-date is a special case of being inaccurate—an inaccuracy caused by a change over time.

Consider, as an example, the annual `Salary` of an employee that can only change at the beginning of a calendar year. Let us assume that the modelled employee's salary was \$38,000 in 1989, \$39,000 in 1990 and \$40,000 in 1991. If at some time during 1991 a datum shows the employee `Salary` as \$40,000, it is up-to-date. If it shows \$38,000, it is two years out-of-date. Finally, if the datum shows the `Salary` for 1991 as \$42,000—a figure that was never correct—the datum is not out-of-date, but simply incorrect.

This is an example of an attribute whose value is supposed to be updated periodically. The definitions also apply to attributes whose values may change but are not required to do so, such as an employee's `Address` or `Name`. The notion of being up-to-date does not apply to permanent properties, like data of birth or blood type. It may be difficult, however, to determine if a property is permanent. Whereas a birth date is permanent, country of birth is not (consider those born in the Soviet Union). A person's name is changeable as long as the person is alive, but becomes permanent after death. Who imagined until recently that a person's gender is changeable? And should we expect the same of blood type?

There are two other time-related terms often used in discussing data and information: "age" and "timeliness." Davis and Olson (1985, p. 223) define "age" as a function of the processing delay necessary to generate and deliver information, and the reporting interval used in the system. In a more thorough analysis, Kleijnen (1980, Chapter 6) considers data in the context of decision making, and analyzes timeliness as the availability of information for decision making. He discusses the importance of update processing delay, update interval, retrieval delay, and decision delay for various kinds of decisions.

3.3 *Completeness and duplication*

Completeness is the degree to which a data collection has values for all attributes of all entities that are supposed to have values. There are two aspects to the problem of data completeness. First is the possibility of missing values in the triples present. Then completeness for a single existing datum may be expressed as a binary measure (Yes or No) of whether it has a value if required (as discussed further below). This can be extended into a measure of completeness for an entire data collection by measuring the fraction of triples with missing values.

When a triple is missing its value component, a special element of an attribute's domain is assigned for it. As we mentioned above, such an element is called "null." Its meaning can be different, depending on whether the attribute is mandatory, optional, or inapplicable.

If an attribute is mandatory, a non-null value is expected. Here the null value is interpreted as "value unknown," the classical interpretation of database theory. Now consider optional attributes, like `Residence-Telephone-Number`. The null value can mean three things:

1. the person has a telephone, but the number is unknown;
2. the person does not have a telephone; or
3. it is not known whether the person has a telephone.

In 1, we assert that the attribute applies to a given entity, but we lack knowledge of its value; in 2, we claim the attribute does not apply; in 3, we declare a lack of knowledge about whether the attribute applies.

Finally, for an inapplicable attribute, for example, `Name-of-Spouse` for a single person, the null value signifies the attribute's inapplicability. Note than the null value is the only correct choice. Therefore the presence of null values of this kind in a database contributes to its completeness.

To summarize, we have discussed five different situations requiring a null value. We can combine two optional attribute cases with corresponding mandatory and inapplicable ones to get the following classification of null values (compare with Lee, 1988):

- an unknown value of an applicable attribute;
- a nonapplicable attribute; and
- an attribute of unknown applicability.

Theoretically, it would be easy to distinguish among these null value cases by having three special elements in an attribute's domain. However, introduction of even one null value causes serious problems for standard database operations (see Date, 1986, Chapter 15; Imielinski & Lipski, 1984).

The second kind of incompleteness stems from the possibility that a data collection can be incomplete because some triples are missing entirely. This "missing triple incompleteness" can be measured by finding the fraction of triples missing from the data collection. On the other hand, a data collection may also contain extra triples, possibly with distinct values for the same attribute of the same entity. It may even contain triples totally irrelevant to the data model. This is known as the *duplicate records* problem.

3.4 *Consistency and integrity*

Date (1983, p. 35) says "consistency" should be used when two or more values in a database are required to agree in some way (that is, they must be consistent). This interpretation, though intuitively appealing, is too narrow because it fails to include some types of constraints. For example, it does not cover the requirement that a value of a particular attribute cannot be null. Therefore, more generally, data is said to be *consistent* with respect to a set of data model constraints if it satisfies all the constraints in the set (Elmasri & Navathe, 1989, p. 597). A possible measure of the consistency of an individual datum is a binary indication (Yes or No) of whether the datum satisfies all constraints. This can be extended into a measure for an entire data collection by determining the fraction of inconsistent triples.

In any good model, correct data must be consistent, but the converse is not true. Consistency is necessary but not sufficient for correctness. Nevertheless, it is useful to single out consistency as a special dimension, separate from accuracy. The main reason for doing so is the practical utility of checking constraints without referring to real-world objects.

The term "integrity of data" is used in the literature in several ways, from accuracy or correctness of data to security and concurrency control in database management systems (Date, 1986, p. 444). It is also used as a synonym for consistency. A detailed exploration of this topic can be found in Brodie (1978), Tsichritzis and Lochovsky (1982), Date (1983), and references therefrom.

We summarize our discussion of quality dimensions for data values in Table 2.

4. RELATED ISSUES AND FURTHER WORK

There are many issues pertinent to data quality not discussed in this paper. In this section, we outline some of the most important of them. This paper has discussed neither dimensions of data quality related to a conceptual view (i.e., those related to a model's entity classes and attributes) nor quality dimensions of data representation and recording. Obviously, if a view is poorly defined (e.g., important attributes are omitted), then even accurate, current, complete, and consistent data values may not be useful. Similarly, if the values are correct, but represented incomprehensibly, the data may not be useful either. Quality dimensions of a conceptual view are discussed by Redman (1992), Fidel (1987), and Flavin (1981). Quality dimensions of data representation can be found in Redman (1992).

Another important topic not investigated here is measurement of data quality. Data quality measurement involves three interrelated questions:

- What should be measured?
- How should measurements be made?
- To what degree should quality be measured on an application-specific basis?

These questions require application of the dimensions discussed here to processes that create, assemble, move, and manage data (Redman, 1992). A key point of the quality paradigm (Juran, 1964) is that efforts to control and improve quality must focus on individual processes. Thus, the dimensions discussed in the paper will typically have to be customized and extended for use in controlling and improving particular data processes and

Table 2. Quality dimensions for data values

Dimensions	Target description	Typical datum measure	Typical database measure	Related notions
Accuracy	Accurate or correct	size of error	fraction incorrect	precision, reliability
Currentness	current	how far out-of-date	fraction out-of-date	age, timeliness
Completeness	complete	Y/N	fraction incomplete	duplication
Consistency	consistent	Y/N	fraction inconsistent	integrity

data processing systems. Our discussion here is the foundation for constructing measurement and management systems for data quality control and improvement.

Our discussion of definitions of "data" and its quality dimensions are motivated by *conventional* data and database systems. We do not consider the consequences for data quality of such newcomers as object-oriented and multimedia databases. We have found in our work that these new sorts of database systems are still in the research stage, and not widely deployed, whereas conventional database systems are the fabric of business processes. Hence our work has focused on current needs. However, it is important to investigate the applicability of both definitions and quality dimensions to these new kinds of databases as they are adopted.

5. CONCLUSION

Several years ago, Kent (1986) made the following comment:

> While the advanced wave of research and technology is moving on to object-oriented and knowledge-based systems, we still haven't come to satisfactory closure on many questions regarding conventional databases.

This paper addresses several foundational issues regarding data and data quality in an effort to bring them to closure.

The first part of the paper considers several approaches to the notion of data. All but one fail most of our six criteria of adequacy. However, we find that the approach to data championed in the database community is adequate in most ways, and with a few more definitions, meets our criteria. Thus we adopt the position that a modelling activity must be applied to some portion of the world to generate *attributes* as well-defined sets of *values*, and *entity classes* as sets of attributes. Particular objects, their properties, and their relationships to one another are modelled as triples of entities, attributes, and values. An entity-attribute-value triple in a model is a *datum*, and *data* is a collection of datum triples. Since data are abstract, they must be represented in some way. A *data representation* is a set of rules for recording triples on some medium, and a *data recording* is an instance of such a representation.

This characterization of data suggests three major data quality concerns: the adequacy of the modelling, the adequacy of the representation and recording, and the adequacy of the triples in the data model. The latter is the focus of much current data quality improvement efforts, and so the second part of the paper considers the quality dimensions of data values. We define and discuss important issues of data accuracy and correctness, precision, reliability, currentness, age and timeliness, completeness, integrity, and consistency. Topics not addressed in this paper, but still important, include issues surrounding the quality of data modelling and representation, data quality measurement theory and practice, and consideration of special data quality issues that may arise for new kinds of database systems, like object-oriented and multimedia systems.

Acknowledgements—We thank an anonymous reviewer who made valuable contributions and suggestions for improvement.

REFERENCES

ANST/ASQC (1978). *Standard A3-1978: Quality systems terminology*. Milwaukee, WI: American Society for Quality Control.

Bendel, A. (1991). Principles of reliability databases. In A.G. Gannon & A. Bendel (Eds.), *Reliability databanks* (pp. 8–11). New York: Elsevier Applied Science.

Blumenthal, S.C. (1969). *Management information systems*. Englewood Cliffs, NJ: Prentice-Hall.

Bolour, A., Anderson, T.L., Dekeyser, L.J., & Wong, H.K.T. (1982). The role of time in information processing: A survey. *ACM SIGMOD Record*, *12*(3), 27–50.

Brodie, M.L. (1978). *Specification and verification of data base semantic integrity*. Doctoral dissertation. Toronto: University of Toronto.

Brodie, M.L. (1980). Data quality in information systems. *Information and Management*, *3*, 245–258.

Burch, J.G., Strater, F.R., & Grudnitski, G. (1983). *Information systems: Theory and practice*, 3rd ed. New York, NY: John Wiley & Sons.

Chapple, J.N. (1976). *Business systems techniques*. London and New York: Longman.

Codd, E.F. (1981). Data models in database management. *ACM SIGPLAN Notices, 16*(1).

Date, C.J. (1986). *An introduction to database systems*, Vol. 1, 4th ed. Reading, MA: Addison-Wesley.

Date, C.J. (1983). *An introduction to database systems*, Vol. 2, 3rd ed. Reading, MA: Addison-Wesley.

Davis, G.M., & Olson, M.H. (1985). *Management information systems: Conceptual foundations, structure, and development*, 2nd ed. New York, NY: McGraw-Hill.

Davis, C.H., & Rush, J.E. (1979). *Guide to information science*. Westport, CT: Greenwood Press.

Dorn, P.H. (1981). Business information in the eighties. In A.E. Pappenheim (Ed.), *Business information systems*, INFOTECH State of the Art Report, series 9(7) (pp. 245-260). Maidenhead, Berkshire: Pergamon Infotech.

Elmasri, R., & Navathe, S.B. (1989). *Fundamentals of database systems*. Redwood City, CA: Benjamin/Cummings.

Fox, C., Gandel, P., and Frakes, B. (1988). Foundational issues in knowledge-based information systems. *Canadian Journal of Information Science, 13*(3), 90-102.

Fry, J.P., & Sibley, E.H. (1976). Evolution of data-base management systems. *ACM Computing Surveys, 8*(1), 7-42.

Fidel, R. (1987). *Database design for information retrieval*. New York, NY: John Wiley & Sons.

Flavin, M. (1981). *Fundamental concepts for information modeling*. New York, NY: Yourdon Press.

Hoare, C.A.R. (1975). Data reliability. *ACM SIGPLAN Notices, 10*(6), 528-533.

Imielinski, T., & Lipski, W. (1984). Incomplete information in relational databases. *Journal of the ACM, 31*(4), 761-791.

Jardine, D.A., & Matzov, A. (1988). Ontology and properties of time in information systems. In R.A. Meersman & A.C. Sernadas (Eds.), *Data and knowledge* (DS-2) (pp. 173-188). Amsterdam: North-Holland.

Juran, J.M. (1964). *Management breakthrough*. New York, NY: McGraw-Hill.

Kent, W. (1978). *Data and reality*. Amsterdam: North-Holland.

Kent, W. (1986). The realities of data: Basic properties of data reconsidered. In T.B. Steel & R. Meersman (Eds.), *Database semantics* (DS-1), (pp. 175-188), Amsterdam: North-Holland.

Kleijnen, J.P.C. (1980). *Computers and profits: Quantifying financial benefits of information*. Reading, MA: Addison-Wesley.

Langefors, B., & Samuelson, K. (1976). *Information and data in systems*. New York, NY: Petrocelli/Charter.

Langefors, B., & Sundgren, B. (1975). *Information systems architecture*. New York, NY: Petrocelli/Charter.

Laudon, K.C. (1986). Data quality and due process in large interorganizational record systems. *Communications of the ACM, 29*(1), 4-18.

Lee, R.M. (1988). Logic, semantics, and data modeling: An ontology. In R.A. Meersman & A.C. Sernadas (Eds.), *Data and knowledge* (DS-2) (pp. 221-244). Amsterdam: North-Holland.

Liepins, G.E., & Uppuluri, V.R.R. (Eds.) (1990). *Data quality control: Theory and pragmatics*. New York, NY: Marcel Dekker.

Loebl, A.S. (1990). Accuracy and relevance and the quality of data. In G.E. Liepins & V.R.R. Uppuluri (Eds.), *Data quality control: Theory and pragmatics* (pp. 105-144). New York, NY: Marcel Dekker.

Mealy, G. (1967). Another look at data. *Proceedings AFIPS 1967 Fall Joint Computer Conference*, 525-534.

Morey, R.C. (1982). Estimating and improving the quality of information in a MIS. *Communications of the ACM, 25*(5), 337-342.

Naus, J.I. (1975). *Data quality control and editing*. New York, NY: Marcel Dekker.

O'Neill, E.T., & Vizine-Goetz, D. (1988). Quality control in online databases. In M.E. Williams (Ed.), *Annual review of information science and technology, 23* (pp. 125-156). New York, NY: Elsevier Science Publishing.

Parnas, D.L. (1977). The influence of software structure on reliability. In R.T. Yeh (Ed.), Current trends in programming methodology, Vol. I (pp. 111-119). Englewood Cliffs, NJ: Prentice-Hall.

Redman, T.C. (1992). *Data quality: Management and technology*. New York, NY: Bantam Books.

Sy, K.J., & Robbin, A. (1990). Federal statistical policies and programs: how good are the numbers? In M.E. Williams (Ed.), *Annual review of information science and technology, 25*(pp. 3-54). New York, NY: Elsevier Science Publishing.

Tasker, D. (1989). *Fourth generation data*. Englewood Cliffs, NJ: Prentice-Hall.

Teskey, F.N. (1989). User models and world models for data, information, and knowledge. *Information Processing & Management, 25*(1), 7-14.

Tsichritzis, D.C., & Lochovsky, F.H. (1982). *Data models*. Englewood Cliffs, NJ: Prentice-Hall.

Wilson, L. (1992). Devil in the data. *Information week*, August 31, 48-54.

Yovits, M.C. (1981). Information and data. In A. Ralston (Ed.), *Encyclopedia of computer science and engineering* (pp. 714-717). New York, NY: Van Nostrand Reinhold.