RETHINKING DATA QUALITY AS AN OUTCOME OF CONCEPTUAL MODELING CHOICES

Research Paper

Roman Lukyanenko
Memorial University of Newfoundland, St. John‟s Canada
roman.lukyanenko@mun.ca

Jeffrey Parsons
Memorial University of Newfoundland, St. John‟s Canada
jeffreyp@mun.ca

Abstract: With the proliferation of unstructured data sources and the growing role of crowdsourcing, new data quality challenges are emerging. Traditional approaches that investigated quality in the context of structured relational databases viewed users as data consumers and quality as a product of an information system. Yet, as users increasingly become information producers, a reconceptualization of data quality is needed. This paper contributes by exploring data quality challenges arising in the era of user-supplied information and defines data quality as a function of conceptual modeling choices. The proposed approach can better inform the practice of crowdsourcing and can enable participants to contribute higher quality information with fewer constraints.

Key Words: Data Quality, Information Quality, Database design, Conceptual modeling, Crowdsourcing, Citizen science.

INTRODUCTION

Data quality is an important concern for organizations, individuals and societies [29, 31]. The quality of data has a direct impact on the quality of decisions made based on that data. This paper attempts to account theoretically for the impact of data modeling activity on the quality of data in a database and introduce additional and potentially significant modeling considerations into data quality research. The motivation for the research comes from the proliferation of *participative information systems*, which pose unique challenges to traditional conceptual modeling and data quality approaches. Close examination of the nascent participative domains can lead to advances in data modeling and shed new light into the nature of data quality in general. Much of the existing research on data quality has focused on traditional, corporate use of databases, in which data is typically stored in a highly structured form. Studies have explored such dimensions as accuracy, completeness, consistency, and fitness for use [e.g. 5, 29, 30, 57]. Prior research viewed users as information *consumers*, and considered data quality to be a *product* of an information system. Yet, the distinction between information consumers and creators is rapidly disappearing. As users become information creators, and large data sets are increasingly being generated by amateur and inexperienced users (e.g. social networks and crowdsourcing projects), database structures make it difficult to accommodate discretionary and often unstructured information without having to constrain user input. Managing the quality of semi-structured and unstructured information is emerging as a new research challenge [34].

We argue it may be possible to address some of the emerging concerns by changing the way information is collected and stored. This paper presents a conceptual modeling approach to data quality that promises both theoretical and practical advantages. We claim that data quality is, to a large extent, a function of conceptual modeling choices. In particular, the choice to record data in terms of classes has significant data quality implications. Once defined, classes affect the degree to which an information system is able

to reflect users‟ view of reality. Thus, conceptual choices *a priori* impose an upper bound on quality dimensions, such as accuracy, completeness, and representativeness. Relaxing the rigid constraints resulting from the use of "class-based" data models can help in capturing user input more objectively, leading to higher quality of stored data.

Keeping the user in mind is important in the context of data quality. Relational databases are being used for an increasingly diverse set of tasks and by different categories of users, each of whom may have his/her own criteria for data quality [4]. Theoretical work on data quality by Wand and Wang [54] emphasizes that it is impossible to anticipate all potential uses of data. As a result, the quality of information stored in a database depends on the design of the database itself. This conceptualization moves beyond prevailing "data-centric" approaches to data quality.

However, with the rapid growth of applications that allow users to create content, another issue needs to be recognized: ***it may be impossible to anticipate all kinds of information that users might want to record in the information system.*** Discretionary data input is growing and databases are being increasingly used to collect diverse information from broad audiences. One example of such data collection is crowdsourcing, the engagement of individuals in purpose-driven projects online [15]. Here, we focus on a particular kind of crowdsourcing – citizen science – the voluntary participation of amateur scientists in scientific endeavors [51]. Due to its potential to dramatically lower research costs and facilitate discoveries, citizen science is rapidly becoming popular. Recently, citizen science issues have been receiving increased attention from the IS community [e.g. 2, 32, 38, 60].

Citizen science illustrates the data quality challenges of crowdsourcing. Online citizen science projects, such as eBird (ebird.org) or iSpot (ispot.org.uk), attempt to harness valuable insights of ordinary people for use in scientific research. It is clearly difficult to *a priori* anticipate what kind of information non-experts can provide, and creating unnecessary constraints can undermine potential gains from such initiatives. Moreover, amateur observers are often unable to record information consistently and in compliance with the requirements of a given scientific domain, leading to what appears to be a tradeoff between levels of participation and data quality [40].

The prevailing method of storing information is recording individual data in terms of classes [cf. 42] and populating attributes that characterize classes with a set of values [see 10]. For example, Tsichritzis, and Lochovsky [53] define *datum* (data item) in a strictly-typed data models as members of an *a priori* category. Therefore "data that does not fall into a category have either to be subverted to fall into one, or they cannot be handled in the data model" [53, p. 8]. Similarly, Redman [47] defines data item in the context of data quality "as a triple <*e, a, v*>, where value *v* is selected from the domain of attribute *a* for the entity *e*" (p. 230). The class (also referred to as category, entity set, kind) is thus an important modeling construct. Classes act as filters upon information, essentially defining the types of information (e.g. attributes and attribute values) a system can store. Identification of classes is traditionally the first step of conceptual modeling [62, p. 221]. The central thesis of this paper is that both the process of creating classes, as well as the choice of specific classes, have a strong (and for reasons discussed below, negative) impact on data quality. This research further contributes by deriving data quality from the process of conceptual modeling and explores data quality deficiencies inherent in a class-centered database design.

The remainder of the paper is organized as follows. The next section provides a review of data quality literature followed by the motivation of the current research. We then discuss the theoretical foundation of the new approach and derive specific data quality propositions. A case study is presented to illustrate the new approach to data quality. The paper concludes with a general discussion and a summary of key findings.

REVIEW OF DATA QUALITY LITERATURE

Overview of Data Quality Research

A taxonomy of approaches to data quality has been proposed by Wang and Strong [57], who classified prior research on data quality as intuitive, theoretical, or empirical. Intuitive approaches often use a contextual or utilitarian definition of data quality and typically explore such dimensions as accuracy or reliability. Theoretical approaches, such as that of Wand and Wang [54], examine data deficiencies inherent in data products and attempt to derive data quality dimensions from fundamental theoretical principles. Empirical work, such as by Wang and Strong [57], empirically examines data quality"s nature and impact.

Much data quality research considers data quality to be a multidimensional construct [24, 29, 44, 57]. One grouping of the dimensions based on data consumers" needs includes intrinsic (e.g. accuracy, believability, reputation), context (e.g. completeness, timeliness), representation (e.g. consistency in representation, ease of understanding) and accessibility (e.g. ease of accessing the data) factors [57]. Another way to classify dimensions is to consider some representing internal (design, implementation, operation) and some, external (use, impact, evaluation) views of information systems [54]. Following that distinction, an internal view can include such dimensions as accuracy, reliability, consistency, while an external view can be associated with timeliness, relevance, understandability, efficiency, and usefulness [for a complete list see 54]. There is also considerable variability in the scope and granularity of dimensions. For example, *completeness* can be viewed at an abstract level as the degree to which a database schema contains all required real world objects. At the same time, completeness can be treated similarly to column integrity, requiring column values to be drawn from a set of permissible, "lawful" values [30]. Definitions of accuracy, one of the most widely cited dimension, vary considerably, mostly due to the intended scope [54]. Similarly, Wand and Wang [54] lament, "there is no generally accepted notion of reliability and … it might be related either to characteristics of the data or of the system." The variability of the data quality concept and its dimensions makes attempts to develop a theoretical model especially important. Unlike ad hoc heuristics, a solid theory can guide practice and make evaluations of data quality more objective.

Data Quality as a Product, Users as Consumers

It is common to consider data quality a product or service produced by an information system and consumed by users. Ballou et al. [5] introduced the concept of *information manufacturing* to "encourage researchers and practitioners alike to seek cross-disciplinary analogies that can facilitate the transfer of knowledge from the field of product quality to less well-developed field of information quality." The term *information manufacturing system* designates "information systems that produce predefined information products" [5].

The manufacturing approach to data quality treats users as *consumers* of data and, as a consequence, of data quality. Thus, the marketing concepts of consumer needs and wants can be applied to data quality. Both marketing and psychology have long recognized a hierarchical nature of needs. Similarly, Wang and Strong [57] examined the impact of data quality on information users. Supporting the "fitness for use" definition of data quality, Wang and Strong [57] include use characteristics as data quality attributes and present a hierarchical grouping of data dimensions that correspond to certain quality needs. The manufacturing-marketing view of data quality has been extended in Kahn et al. [27] to include service quality. Both product and service aspects of information quality can conform to specifications or meet expectations.

In any production or manufacturing system, the setup of the system itself becomes important in shaping its product. Recognizing this, Wand and Wang [54] note "the quality of data depends on the design and production process involved in generating the data." In this design-centered view, data quality can be seen as the degree of discrepancy between the view of reality that can be inferred from an information system and what can be directly observed in the real world. This definition is grounded in Bunge's ontology [9, 55].

Our review of the literature shows that most studies consider data quality in the context of structured relational databases and use a manufacturing analogy for useful knowledge transfer. In contrast, we reconceptualize data quality based on an analysis of less structured datasets typical in participative domains for which user input is often difficult to anticipate. The next section discusses emerging issues that motivate this conceptual shift.

MOTIVATING THE CONCEPTUAL SHIFT IN DATA QUALITY

Design vs. Data-centric Data Quality

While data modeling is often mentioned in the context of data quality [e.g. 48], much data quality research focuses on issues arising after the design of a database is finished and the database is put into production. This is consistent with the view of *data quality as a product of consumption* and the fitness for use paradigm. In fact, Wang et al. [56] advanced an attribute-based data model to facilitate cell-level tagging of existing data sets with relevant data quality dimensions (e.g. "useful," "relevant," "timely"). Alternatively, a source can be attached to the values themselves [see 56].

Data-centric, *a posteriori* approaches have a number of shortcomings. First, scanning information to determine its degree of data quality can be costly, and often difficult. This is recognized by Inmon et al. [25], who suggest that a scouting of data before full scale consumption is more efficient. Second, such approaches cannot support early specification of data quality requirements [54]. Finally, the same piece of information can be "timely" and "accurate" in one situation and "outdated" and "incorrect" in another. Since it is difficult to anticipate all potential uses, and difficult to represent corresponding views of reality in a single data model, preferential treatment of some uses appears necessary. For example, a database of corporate assets that is intended for an accounting department may be modeled using classes and properties pertinent to the accounting domain. Those classes and properties, however, may neglect or be in conflict with some potential ad hoc uses of data by other corporate units. Thus, while the fitness-for-use paradigm is ostensibly data-centric, it appears to carry design-centric implications because the focus on use may introduce a modeling bias at the stage of systems analysis and design.

Addressing some of the shortcomings of a data-centric data quality is possible by exploring quality implications of conceptual modeling choices. This approach carries a number of desirable outcomes. First, certain quality standards and thresholds can be specified before data entry. Such proactive design can minimize future changes. Redesigning a database schema after application deployment can be extremely costly and highly undesirable. Second, as recognized by Wang and Strong [57] and illustrated by Wand and Wang [54], an *a priori* focus can uncover fundamental relations that transcend specific idiosyncratic implementations and data sets. The focus on fundamental conceptual modeling principles can be more universal and flexible, and thus potentially applicable to a greater spectrum of situations (e.g. citizen science or business).

Users as Information Providers: Beyond the Manufacturing Model

In the information manufacturing paradigm, data quality is a product consumed by users. This analogy has shown to be useful in knowledge transfer from the quality control field to IS. The manufacturing view, however, implies that a consumer is largely removed from the process of product creation. In marketing research, information asymmetry suggests that consumers do not know exactly how the product is manufactured and what features it may possess [26]. Yet, the relationship between product quality and the user may be different if the consumer shares the role of a product creator.

The last decade has seen a rapid rise of Web 2.0 technology, which embraces user-supplied information and increased user interaction. Successful Web 2.0 applications (e.g. Facebook, Wikipedia, Youtube, and Twitter), have many millions of users. Ordinary people are becoming more comfortable in the new role of information creators. Concepts such as *customer-driven innovation* [37] and *crowdsourcing* [15] are being actively explored. A growing body of research aims to harness collaborative and participative computing for business needs [3, 37]. Yet, active solicitation of content from users carries a new set of challenges related to data quality, leading to a new data quality research frontier. This is evident from research and practice of IT-driven citizen science.

Data Quality Challenges of Discretionary Data Collection

With the evolution of Internet technologies, it has become easier for ordinary people to participate in scientific projects, known as citizen science [51]. Humans can be effective sensors of their environment [21] and human volunteers are now engaged in a variety of scientific projects online – from folding proteins to finding interstellar dust; from identifying birds to classifying galaxies [see 23]. Yet, given the expertise and language gap between scientists and ordinary people, information transfer in citizen science projects is not straightforward. Despite the potential of citizen science, serious doubts about quality of citizen science data preclude it from playing a more important role in research and decision-making [see, for example, 17]. Although intrinsically motivated, volunteers may have little invested interest in projects run by scientists and as one volunteer remarked: "Despite your best efforts, your mind wanders. You start thinking about lunch or whatever" [23, p. 686]. According to Foster-Smith and Evans [19] while citizen scientists can offer insights and generate new ideas, their lack of training and expertise often results in inconsistent and incorrect data [see 11, 17, 59].

There is no universally acceptable approach to improving the quality of citizen science data. Research on database information quality offers little guidance to deal with specific challenges of citizen science data quality. Citizen science is based on the recognition that non-experts can possess valuable scientific information. Yet, this information comes from users who do not necessarily understand the scientific domain, its language or structure. It has been suggested that an information system should faithfully represent reality [58]. For example, the theoretical framework by Wand and Wang [54] measures data quality as a discrepancy between digitally transformed and directly observable reality. In crowdsourcing each user may have a unique view of reality. This means that the same crowdsourcing project may need many data models, each with own data quality specifications. The absence of a theoretical framework underpinning quality of user-generated data means that practitioners have to rely on ad hoc heuristics.

One of the most popular ways to increase data quality is to train volunteers. The focus on training and procedures is advocated by Dickinson, Zuckerberg and Bonter [13], and by Foster-Smith and Evans [19]. Training however can sometimes introduce biases as participants may guess the objective of the study and overinflate or exaggerate information [1].

Another approach to increasing quality is expert verification. For example, in a project where volunteers were asked to identify carcasses of by-catch and beached birds, the results were later verified by experts [22]. However, with the increasing size of data sets [59] extensive expert verification of user-supplied data is unrealistic and in many ways contrary to the spirit of citizen science.

Exploiting the advantages of Web 2.0, collaboration between citizen scientists has been suggested as a key to increasing data quality. Seeing trust as a proxy for data quality, Bishr and Mantelas [7] propose trust and reputation model for classifying knowledge. This approach is the basis for a UK-based iSpot, a website that relies on social networking for collaborative identification of species [52]. While the social networking/trust approach appears promising, it has a number of serious limitations. Despite being likened to the "scientific peer review process" [7], social networking is appropriate only for popular citizen science projects with a significant user base. Web sites with a small number of users will not have sufficient user activity per unit of data to ensure adequate scrutiny. The peer review process also raises a philosophical issue of whose reality is being represented and stored: the original user who submitted data or the expert user who verified and corrected it? We elaborate on this issue below using a case study.

In summary, conventional wisdom in citizen science holds that, in order to increase the quality of the supplied information, the experience and expertise of the information creators must be enhanced (through training or verification by experts or peers). Yet, training can be expensive and, since only a small number of people can be experts in something, this implies that the best data quality can come from a limited number of people. Such an approach can thereby severely limit the potential scope of citizen science. We argue, however, that this can be avoided by changing the way data is collected and stored. In fact, the challenges of citizen science projects help reveal a general principle of data quality: **data quality is a function of the data structures used to hold user-supplied information**.

RECONCEPTUALIZING DATA QUALITY

In order to understand and address the emerging data quality challenges, several assumptions need to be clarified. First, unless the interface is supportive, it is unlikely that uncommitted users such as citizen scientists will communicate all the information they might want to. Coleman et al. [11] described a neophyte volunteer as one who "uses … information provided at a given Website without question." For example, a typical citizen science project may ask a user to classify an observation at the *species* level (e.g. eBird.com). Suppose a user can only be sure of the higher *genus* level. In this case, a user might therefore choose not to participate, or might make a potentially incorrect guess. Alternatively, suppose a user knows not only the species, but also a *variety* or *subspecies*. To satisfy the interface requirements species classification is enough. In this case, finer-grain and potentially valuable information may be lost unless a user puts in extra effort to communicate it.

Second, we assume that the data collection process and the user interface are strongly influenced by the underlying data structure. Application design typically follows database development and closely reflects objects defined in a database. Ultimately it is the database schema that impacts the information collection activities [32].

Wand and Wang [54] define data quality ***deficiency*** as "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" (p. 89). Using this definition in a real context suggests the need to examine the impact of *classification*, the process by which data storage is organized.

The prevailing method of storing information in databases is recording instance information in terms of (usually one) *a priori* defined class [cf. 42]. Classification is a fundamental activity in which humans engage to manage what some call "infinity" of real world stimuli [49]. Raven et al. [46] put it more bluntly: "Man is by nature a classifying animal." While much of the database design research draws parallels between computer classes and human cognitive processes, few have noticed fundamental differences. Although humans and databases use classes for the same reasons [cognitive economy,

inferential utility, see 12, 39, 43], in each case the classification *process* and its *consequence* are not the same. *We claim that this difference is a fundamental cause of many data quality problems in modern relational databases*.

Proposition 1: Data quality in an information system is necessarily reduced whenever a class is used to *store* instances.

Since many classes can be used to represent the same phenomena, it is unclear which class is better. Parsons and Wand proposed cognitive guidelines for choosing classes that could be used for "reasoning about classification, and their violation indicates something is „*lost*" from a cognitive point of view" [41, p. 69; emphasis added]. Choosing the "wrong" one means that information stored will be deficient with respect to perceived reality. Extending this idea further, we claim that using classes for storage will *always* fail to fully capture reality, no matter how "good" the chosen classes are.

Proposition 2: An instance can never be fully represented by a class.

Any complex object has a large number of features and no one class can encompass them all. In fact, storing instances in terms of classes *always* means that some potentially valuable properties are sacrificed for the cognitive efficiency provided by classification. For example, if we define a class *student* (assuming it has no subclasses), then any individual instance of that class will possess *only* those attributes that are part of the class definition. This also means that *all other* potentially useful attributes will be lost. To classify is to abstract from the diversity of the world by focusing on properties that satisfy some utility. In this process, certain properties that are not of immediate interest are neglected. Yet, they can be invoked at a later time should it be necessary. Here lies the difference between *human* and *computerized* representation. When humans classify, they *focus* on relevant features *but remain aware of other ones*. In contrast, when we *record* data into a database, the information that is not committed to storage is lost. We call this loss *the loss of properties* principle, which is a corollary to Proposition 2.

Corollary 1: Every time an instance is stored as a member of a class, loss of properties occurs.

The following scenario illustrates the loss of properties principle and its implications. Suppose a citizen scientist observes a kind of bird that he/she has not seen before. Lacking expertise, a non-expert may resort to analogies (*seagull-like*), basic-level categories (*birds*) or superodinate categories (*living thing*s, *animals*) to classify or simply reason about the observed phenomenon. Suppose, later he or she enthusiastically tells friends about a recently sighted *unusual seagull* (the category that best describes the observed phenomenon). As a consequence, the friends can visualize the birds and infer properties *unusual seagulls* may possess (e.g. lay eggs, live by the sea, eat fish, can fly) without having to observe them or being explicitly informed of what they are. Later, the same citizen scientist comes across a similar-looking bird in a birding field guide, and quickly notices the striking similarities to the birds observed before. This happens because the human brain retains many details that persist over time [8]. After examining the field guide, the non-expert knows that the observed birds were actually *northern gannets*. In other words, the instance has been *reclassified* based on the originally observed features. The fact that the original class was *unusual seagulls* does not preclude humans from retaining those features that, at a later time, allow the observation to be classified as *gannet*. Seagulls and gannets are different classes and storing instances as one or the other class carries different data quality implications on dimensions such as accuracy. In contrast, natural classification does not impact data quality in the same way. Thinking of gannets as seagulls does not preclude humans from capturing features that were different from a typical seagull and using the class *seagull* to efficiently communicate information that is appropriate for the casual conversation (i.e., fit for use).

The above example illustrates why classification works well for humans. However, it may be ill-suited for

the prevailing methods of recording data in databases. Cognitively, classifying an observed phenomenon does not preclude humans from retaining individual details not implied by the class. A database is different in four important ways.

First, in a strictly-typed data model, a class definition is a *hard* constraint on the types of data that can be recorded. A data model cannot accept a gannet as a seagull unless all the attributes supplied match the conceptual definition of a seagull. Once a data entry operator chooses seagull as a class, all gannet-like attributes that do not "fit" the schema definition of a seagull will be rejected by the system. After data is captured, it may be taken at face value: "seagull", and there will be no way to ever reclassify as it as gannet. In this sense, storing instance data as a member of a class means some individuality carried by properties not included into the schema is irreversibly lost.

Second, the details of a classification decision (i.e., uncertainty) are not stored with the data. The issue of dealing with database uncertainty is being increasingly researched in part to provide support for the Semantic Web [20, 33]. Here, we emphasize two different aspects of uncertainty: the uncertainty of classifying real-world phenomena by humans and the related uncertainty of matching the classification hypotheses with (usually one) data storage equivalent. In the above scenario, a user was clearly unsure of the initial classification. In fact, this uncertainty might trigger a quest for a better answer, which can also explain why some details irrelevant to the original classification were retained. It appears humans know the choice is not final, and retain as much "evidence" as possible for a later time. Databases can accommodate some of the uncertainty using fuzzy data models [61]. Based on Proposition 2, however, there can be many probabilistic class memberships for any given instance, each with different degrees of correspondence to the set of classes considered by a user and thus, with different data quality implications (discussed later).

Third, both databases and human memory can be used for later information retrieval and decision-support. Data warehouses powering business intelligence are frequently designed using predefined class-based structures. The quality of warehouse data is of a paramount concern [16, 25]. When data is aggregated into warehouses, it needs to be transformed (e.g. using ETL tools) to conform to the unified structure. This process is known as *schema matching* and is considered to be "extremely difficult" with dearth of a universal theoretical foundation and a large number of ad hoc heuristic solutions [for review of the discipline see 14, 45]. Propositions 1 and 2 discussed in the present paper suggest that the process of data transformation can lead to inherent deficiencies of the aggregate product. Such deficient data can lead to ill-informed decisions.

Finally, data quality is impaired by the *requirement* of information systems to classify at a prescribed level, which does not exist in reality. Classification is intuitive for humans [46]. Yet, classification is not always *possible* at the level of specificity defined by the database schema, and while a user is flexible with the level or classification granularity, the schema is usually not. In the case of gannets, a user could not objectively classify at the species level, yet many information systems require or imply exactly that [40]. Further, soliciting a probability of classification from a user, which may be necessary to support fuzzy data models appears counterintuitive to the way humans think. Humans may sense uncertainty, but translating it into a numerical equivalent (e.g. 90% certainty the observed *bird* is a *seagull*, and 10% certainty it is something else) is both awkward and arbitrary.

In the real world, humans employ a large number of alternative categories they feel more comfortable with. In fact, depending upon a particular goal, humans may create ad hoc categories, which never existed before [e.g. "things to take from a burning house", see 6]. Unlike in a database schema, ad hoc classes are usually discarded after use. Clearly, human categorization is dynamic and flexible in ways that are difficult to *a priori* predict and probabilistically quantify. In a database, however, the choice is usually (1) select classes (usually one) defined by the database schema, (2) choose "other" or "unknown," or (3)

refrain from data entry. And while "other" or "unknown" seem to be the optimal choice for strict data model uncertainty, it is probably the least desirable: little inferential utility can be drawn from a category which potentially lumps together dissimilar objects. The other two options engender a potentially incorrect guess or constrain participation.

A choice of a specific class can have varying impact on data quality. Different designers can model the same reality differently. As Parsons and Wand [42] observe, an "information model is constructed to reflect the views of a single user at a given point in time." The instances that are being recorded in a database can belong to multiple classes. While there have been attempts to support multiple classification, the prevailing database practice is to reserve one class.

Data Quality as a Gradient Fit

In a general sense design-centric data quality can be understood as a gradient fit between a data storage "container" and its real-world source. The degree of the fit affects such quality dimension as *accuracy* and *completeness*. The gradient fit can be seen at class, attribute and value levels of information conceptualization.

To understand the class-instance fit, let us consider a case of data values first. Suppose a financial analyst wants to record exchange rates of 0.567 and 0.54 in a data field that holds two decimal places. In this case only the second value will be a "perfect" fit for that data container, while the first one will have to be rounded. Value-based data quality is gradient in nature. Assuming that a rounding rule to the hundredth decimal is triggered, values of 0.569 and 0.566 will have different degrees of *information loss* (i.e. 0.001 and 0.004 respectively).

This relationship is often difficult to quantify. Suppose an account manager responsible for placing credit offenders on prepay needs to ensure that an extremely high risk customer stays on prepay indefinitely. A credit database contains *date off prepay* field, which is linked to certain business rules (e.g. allowing customers to use credit). Thus, a specialist may choose to assign a value far in the future as a proxy for *infinity*. In this case, years 2015, 2020 and 2120 will have different degrees of correspondence to what the data creator had in mind. The third option is more satisfactory, while the first one is less so. Thus, we can consider the three stored values, 2015, 2020 and 2120 as having different degrees of correspondence to the data creator‟s intentions and thus, different *degrees of data quality*.

A similar relationship between stored and real-world phenomena exists at the class level as well. Any given instance has different degrees of correspondence to a given conceptual class. According to Proposition 1 and 2, no class can be a perfect fit for an instance.

Well-researched notions of prototyping and typicality are analogous to the gradient nature of data quality at the instance-class level. Observations such as "*robin* is a more typical *bird* than *penguin"* suggested gradient structure of categories for a number of cognitive researchers [28, 50]. Rosch and Mervis [50] defined prototypes as "best examples of a category" (p. 574), and later Rosch [49] argued that "to increase the distinctiveness and flexibility of categories, categories tend to become defined in terms of prototypes or prototypical instances that contain the attributes most representative of items inside and least representative of items outside the category" (p. 31). Barsalou [6] suggested that gradient structure is also applicable to non-natural, *ad hoc* categories (e.g. *things to take from a burning house*).

While prototyping research has been questioned and is not universally accepted by the cognitive community (e.g. category composition [18] and others issues, see [35]), it provides a useful analogy and enables knowledge transfer from a well-developed field of psychology into the field of data quality. The notion of gradient structure of categories is consistent with Propositions 1 and 2. Since a class can never

fully represent an instance due to individual differences of each instance, instance individuality assures different conceptual correspondence between an instance and any class.

Proposition 3. There are different degrees of conceptual fit between any instance and any given class.

Corollary 2: Data quality of instances in class-based data structures exhibits a gradient nature.

Consider several examples. For a university, students who are registered for full-time course work are more "typical" than those taking one or two correspondence courses. Thus, a decision can be made to model full-time students and correspondence students as separate conceptual classes, which affects the kinds of properties that will be collected and retained by an information system. For a library, a person who borrows books and has an account with a library is a more typical exemplar of a *patron* than a person who visits the library"s webpage and searches its catalog. Since there can be a degree of engagement with library services and each engagement may be unique, who gets recorded and what properties get stored becomes discretionary. In citizen science, identifying *robins* may be easier than identifying *boreal lichens*, thus *robins* appear to be better fitting for the *species* level of specificity, given a normal level of user expertise. In each case, instances exhibit similar data quality patterns as values and dates discussed earlier: some fit better than others into their intended storage containers. Unlike conceptual values, where a real-world value of 1960 can have a perfect equivalent in a database, in a class-instance pair no perfect fit is possible. Some properties are invariably lost. And much in the same way as information is lost when storing 2050 as a proxy of infinity or 0.6 as the result of rounding 0.56, more properties are lost when the conceptual fit between an instance and a class becomes coarser.

DATA QUALITY CASE STUDY

We present a case study to illustrate how data quality is affected by the way information is stored.
Suppose a number of unusually looking *barn swallows* decided to nest in a neighborhood full of citizen scientists. Multiple sightings took place by observers with different levels of domain expertise. Each sighting has been recorded in a hypothetical database shown in Figure 1. The structure is typical of an ecological citizen science database and is based on the authors" correspondence with the Cornell Lab of Ornithology (eBird.org). The design also uses prevailing ER model and asks users to classify observed phenomena at a species level. Below we consider two possible scenarios.

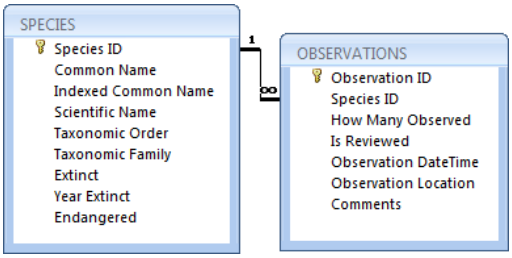

Figure 1. Typical ER diagram of a citizen science project

Scenario 1. Citizen scientist John is a domain expert. John goes to a popular online birding website to record his sighting. John chooses Barn Swallow *(Hirundo rustica)* in the interface and populates remaining attributes of the *OBSERVATIONS* table. While this in some sense is an *accurate* observation, it is still deficient. When John records a sighting of a *barn swallow* a SpeciesID of barn swallow is entered in the table holding transaction details. This suggests that the schema describing barn swallows as a class is sufficient to represent the exact individual observed. In other words, the same schema can be used for

all future observations of barn swallows no matter how many different attributes may be present. Inherent to class-based data models, the individuality of observed attributes escapes structured storage.

Suppose, however, that John also notices unusual coloration of the swallows and describes it in the comments. Comments are unstructured and difficult to analyze. As a result, the fact that birds were unusual may never be uncovered and other observations of the same abnormality will not be linked to each other. Further, aggregating text fields is challenging. Yet, such individual variations can be important and valuable for science. For example, it has recently been noticed that "male barn swallows from Chernobyl have a pale red coloration compared to males from a control area" [36].

Such individual attributes can also be captured using extra columns, but this is inefficient as most other records in the table will have corresponding *null* values. Each individual has something unique, but creating an extra column each time a new attribute needs to be stored is not realistic, as this means the schema is changed and all the objects that depend upon it need to be updated.

Alternatively, an additional table of attributes can be stored with the observations table, allowing none, one, or many additional attributes to be recorded with each observation (Figure 2). Yet, this can cause redundancy as some of the additional *ad hoc* attributes can conceptually overlap with those used in the original class definition. For example, a user may observe how many legs something has, or what color it is and supply these attributes to be recorded in the *OBSERV-ATTRIBUTES* table. Yet, such attributes may be redundant if *SPECIES* table already include them in the schema.

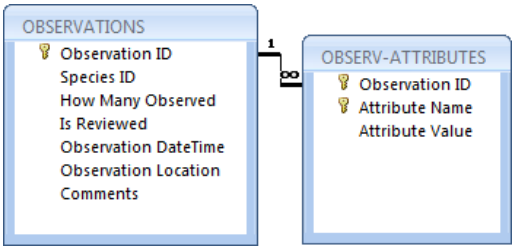

Figure 2. Possible solution to capture individuality

Scenario 1 shows that, in spite of John"s domain expertise, data supplied by John lacks individuality due to the way information is stored. If at a later date attributes other than those defined by the class *barn swallow* are required, a database may fail to provide them. Further, if a different view of reality becomes important, classifying a bird at the species level may no longer be sufficient. Thus the existing database structure focuses on one prescribed aspect of the world and ultimately underrepresents the observed reality.

Scenario 2. Unlike John, Jane is an amateur birder. Jane observed the same birds, but does not know what they are. The field guide does not help, as they are somewhat different from the birds described in it. Without domain expertise, it is difficult for Jane to classify at the required level of specificity. She knows the phenomenon is a *bird*, but the implied objective of the system is to classify at a lower level. Several options are possible.

(1) Jane can record species as *unknown* and let others decide on what it is, and possibly *reclassify* it. This is advocated by the UK-based citizen science website iSpot [52]. This approach has a number of limitations. If Jane"s sighting remains unnoticed by the experts, it will add to a category with little inferential value. It is also possible that one expert says it is a *barn swallow* while two more suggest that it is not, and the information system will have to make a choice to store it as an *unknown* or *barn swallow*. In each case, none of the options are optimal [15]. But even if the experts are successful at reaching a consensus, it means it is not Jane"s, but *others'*, view of reality that will be *recorded* and

used for decision-making purposes by the database. Consequently, the database will fail to faithfully represent the *original* view of reality.

(2) Another possibility is that Jane will simply choose the species that she thinks is the best choice in order to satisfy the interface requirements. How good this choice is remains unknown to the information system. Typical applications are not concerned with recording the human rationale for classifying, but are instead concerned with its outcome. Data can be verified by experts after creation, but this may not be realistic in large data sets. How much of this kind of data of dubious quality (guesses to satisfy the requirement to classify) exist in the modern databases and are routinely used at face value has not, to our knowledge, been surveyed.

(3) Finally, Jane may opt not to participate, as she may reason it is objectively impossible to classify at the required level. This means that while accuracy of the overall dataset is unaffected the data completeness dimension is lowered. The database represents fewer phenomena in the real world and the reality that is stored is biased towards (1) instances that were easier to classify and (2) instances recorded by more expert or assertive users.

Thus, in both Scenario 1 and Scenario 2 data quality suffers with respect to accuracy, completeness, and representativeness. In each case, the root cause is the need to record an instance as a member of a class.

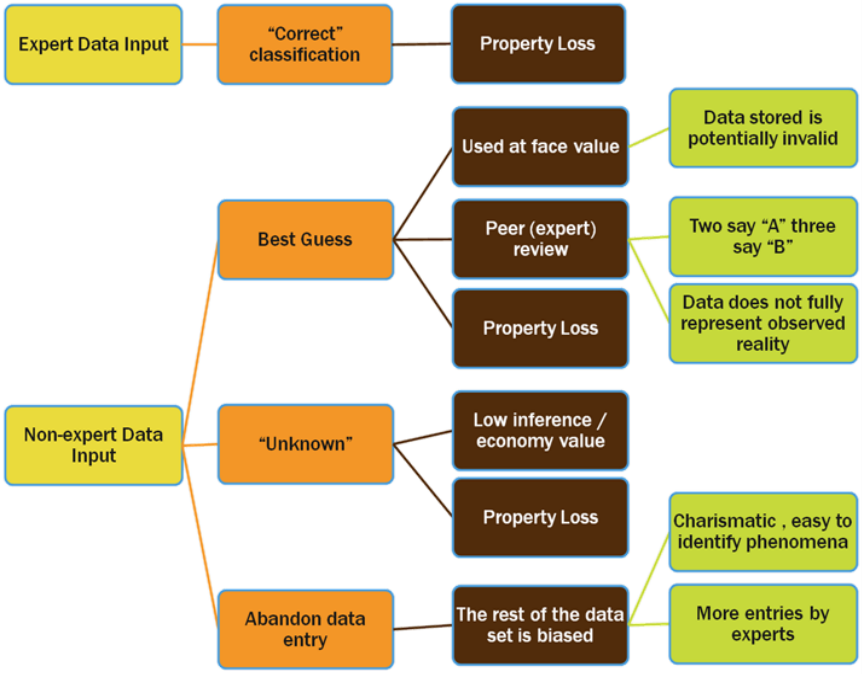

Figure 3. Data quality scenarios

The two scenarios are summarized in Figure 3, which shows both the universality of property loss and potential ways information can be distorted due to the way it is stored. The diagram clearly implies that no matter what the scenario is, the resultant data set is adversely affected.

DISCUSSION AND CONCLUSION

In this paper we offered a theoretical approach to data quality that derives data quality from the prevailing practice of storing information in terms of classes. We motivated a reconceptualization of data quality by examining the difficulty of accommodating data from the emerging domains such as crowdsourcing. While we used the citizen science domain to illustrate the deficiencies of class-based storage, it should be

noted that it may be applicable to any complex domains, including those with established structure. In the business environment, for example, a *customer* may have different definitions across the enterprise. Without details behind each record, simply querying *Customers* table may provide a misleading picture about the number of customers a company has and lead to ill-informed decisions.

Unlike biology, the business domain often operates with more abstract objects. Much of the theoretical foundation on classification, however, uses material objects (e.g. bird, rose, tree) to understand principles of human cognition. Such material objects are observable, and have a stable and often intuitive set of intrinsic attributes. In contrast, business classes often describe invented entities (e.g. account, customer, contract). Forming class definitions of such phenomena can be difficult without domain expertise. Building a database system that avoids the pitfalls of forced classification has the potential to save training expenses while improving the quality of corporate data.

Such a system may be based upon the instance-based data model (IBDM) developed by Parsons and Wand [42]. In an instance database users are not required to classify observed phenomena and can record any observable attributes associated with the information they are contributing. This addresses the consequences of Propositions 1 and 2. By allowing instances to exist independent of classifications, a database does not place an *a priori* constraint on the potential information that can be stored. Thus, citizen scientists and corporate database operators alike can supply attributes based on their respective levels of domain expertise. Once several attributes are recorded, the system can match them with pre-existing sets of identifying attributes for a phenomenon (such as biological species), and either infer a species or ask for additional attributes that could also be automatically deduced from those previously supplied. The final attribute set can potentially match to a class or simply record data without classifying it. Doing so avoids inherent data quality deficiencies of the class-based models.

In many classification situations, a mismatch is possible between the observer"s view of a phenomenon and the rigid model of the database schema used to store information about phenomena. As end-users are generally unable to change the way information is stored, they have no choice but to comply and force-fit their observations to the structure imposed by the schema. The result is stored information that may inaccurately represent the perceived reality, or that is biased towards user expertise and easier classification choices. By considering the impact of conceptual modeling on data quality, the research and practice of database design can focus on better ways to store user input and make both participative and corporate data sets more reflective of the domains they aim to represent.

REFERENCES

[1] Aaron, W. E. G., Tudor, M. T. and Haegen, W. M. V. The Reliability of Citizen Science: A Case Study of Oregon White Oak Stand Surveys. *Wildlife Society Bulletin*, *34* (5) 2006, pp. 1425-1429.

[2] Alabri, A. and Hunter, J. Enhancing the Quality and Trust of Citizen Science Data. in *Proceedings of the IEEE eScience 2010* (Brisbane, Australia, December 8 - 10, 2010), 2010.

[3] Andriole, S. J. Business Impact of Web 2.0 Technologies. *Communications of the ACM*, *53* (12) 2010, pp. 67-79.

[4] 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.

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

[6] Barsalou, L. W. Ad hoc categories. *Memory & Cognition*, *11* 1983, pp. 211-227.

[7] Bishr, M. and Mantelas, L. A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal*, *72* (3) 2008, pp. 229-237.

[8] Brady, T. F., Konkle, T., Alvarez, G. A. and Oliva, A. Visual long-term memory has a massive storage capacity for object details. *PNAS Proceedings of the National Academy of Sciences of the United States of America*,

105 (38) 2008, pp. 14325-14329.
[9] Bunge, M. A. *The furniture of the world*. Reidel, Dordrecht; Boston, 1977.
[10] Chen, P. P.-S. The entity-relationship model - toward a unified view of data. *ACM Trans. Database Syst.*, *1* (1) 1976, pp. 9-36.
[11] Coleman, D. J., Georgiadou, Y. and Labonte, J. Volunteered Geographic Information: The Nature and Motivation of Producers. *International Journal of Spatial Data Infrastructures Research*, *4* 2009, pp. 332-358.
[12] Corter, J. and Gluck, M. Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, *111* (2) 1992, pp. 291-303.
[13] Dickinson, J. L., Zuckerberg, B. and Bonter, D. N. Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, *41* 2010, pp. 112-149.
[14] Doan, A. and Halevy, A. Y. Semantic-integration research in the database community - A brief survey. *AI Magazine*, *26* (1) 2005, pp. 83-94.
[15] Doan, A., Ramakrishnan, R. and Halevy, A. Y. Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, *54* (4) 2011, pp. 86-96.
[16] English, L. P. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. Wiley, 1999.
[17] Flanagin, A. and Metzger, M. The credibility of volunteered geographic information. *GeoJournal*, *72* (3) 2008, pp. 137-148.
[18] Fodor, J. A. *Concepts: where cognitive science went wrong*. Clarendon Press, 1998.
[19] Foster-Smith, J. and Evans, S. M. The value of marine ecological data collected by volunteers. *Biological Conservation*, *113* (2) 2003, pp. 199-213.
[20] Galindo, J., Urrutia, A. and Piattini, M. Representation of fuzzy knowledge in relational databases. in *Proceedings of 15th International Workshop on Database and Expert Systems Applications*. 2004, pp. 917-921.
[21] Goodchild, M. Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69* (4) 2007, pp. 211-221.
[22] Hamel, N. J., Burger, A. E., Charleton, K., Davidson, P., Lee, S., Bertram, D. F. and Parrish, J. K. Bycatch and beached birds: Assessing mortality impacts in coastal net fisheries using marine bird strandings. *Marine Ornithology* 2009, pp. 41–60.
[23] Hand, E. People power. *Nature*, *466* (7307) 2010, pp. 685-687.
[24] Hoxmeier, J. A. Typology of database quality factors. *Software Quality Journal*, *7* (3) 1998, pp. 179-193.
[25] Inmon, W. H., Strauss, D. and Neushloss, G. *Data quality in DW 2.0*. Morgan Kaufmann, 2008.
[26] Johnson, A. R. and Folkes, V. S. How consumers' assessments of the difficulty of manufacturing a product influence quality perceptions. *Journal of the Academy of Marketing Science*, *35* (3) 2007, pp. 317-328.
[27] Kahn, B. K., Strong, D. M. and Wang, R. Y. Information quality benchmarks: product and service performance. *Commun. ACM*, *45* (4) 2002, pp. 184-192.
[28] Lakoff, G. *Women, fire, and dangerous things : what categories reveal about the mind*. University of Chicago Press, Chicago, 1987.
[29] Laudon, K. C. Data quality and due process in large interorganizational record systems. *Communications of the ACM*, *29* (1) 1986, pp. 4-11.
[30] Lee, Y. W., Pipino, L. L., Funk, J. D. and Wang, R. Y. *Journey to Data Quality*. The MIT Press, 2006.
[31] Lorence, D. P. The perils of data misreporting. *Communications of the ACM*, *46* (11) 2003, pp. 85-88.
[32] Lukyanenko, R., Parsons, J. and Wiersma, Y. Citizen Science 2.0: Data Management Principles to Harness the Power of the Crowd. in *Design Science Research in Information Systems and Technology (DESRIST-2011)*. 2011, pp. 465-473.
[33] Ma, Z. M. *A Literature Overview of Fuzzy Database Modeling*. IGI Global, 2009.
[34] Madnick, S. E., Wang, R. Y., Lee, Y. W. and Zhu, H. Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality*, *1* (1) 2009, pp. 1-22.
[35] Margolis, E. and Laurence, S. *Concepts: core readings*. MIT Press, 1999.
[36] Moller, A. P. and Mousseau, T. A. Albinism and Phenotype of Barn Swallows (Hirundo rustica) from Chernobyl. *Evolution*, *55* (10) 2001, pp. 2097-2104.
[37] Nath, A. K., Singh, R., Iyer, L. S. and Ganesh, J. Web 2.0: Capabilities, Business Value and Strategic Practice. *Journal of Information Science & Technology*, *7* (1) 2010, pp. 22-39.
[38] Nusser, S., Miller, L., Clarke, K. and Goodchild, M. Geospatial IT for mobile field data collection. *Communications of the ACM*, *46* (1) 2003, pp. 45-46.
[39] Parsons, J. An Information Model Based on Classification Theory. *Management Science*, *42* (10) 1996, pp. 1437-1453.
[40] Parsons, J., Lukyanenko, R. and Wiersma, Y. Easier citizen science is better. *Nature*, *471* (7336) 2011, pp. 37-

37.
[41] Parsons, J. and Wand, Y. Choosing classes in conceptual modeling. *Communications of the ACM*, *40* (6) 1997, pp. 63-69.
[42] Parsons, J. and Wand, Y. Emancipating Instances from the Tyranny of Classes in Information Modeling. *ACM Transactions on Database Systems*, *25* (2) 2000, pp. 228.
[43] Parsons, J. and Wand, Y. Using cognitive principles to guide classification in information systems modeling. *MIS Quarterly*, *32* (4) 2008, pp. 839-868.
[44] Pipino, L. L., Lee, Y. W. and Wang, R. Y. Data quality assessment. *Communications of the ACM*, *45* (4) 2002, pp. 211-218.
[45] Rahm, E. and Bernstein, P. A. A survey of approaches to automatic schema matching. *The VLDB Journal*, *10* (4) 2001, pp. 334-350.
[46] Raven, P. H., Berlin, B. and Breedlove, D. E. The Origins of Taxonomy. *Science*, *174* (4015) 1971, pp. 1210-1213.
[47] Redman, T. C. *Data driven: profiting from your most important business asset*. Harvard Business Press, 2008.
[48] Redman, T. C. *Data quality for the information age*. Artech House, 1996.
[49] Rosch, E. Principles of Categorization. in Eleanor Rosch & Barbara Lloyd (Eds.), *Cognition and Categorization*. John Wiley & Sons Inc, 1978, pp. 27-48.
[50] Rosch, E. and Mervis, C. B. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7* (4) 1975, pp. 573-605.
[51] Silvertown, J. A new dawn for citizen science. *Trends in Ecology & Evolution*, *24* (9) 2009, pp. 467-471.
[52] Silvertown, J. Taxonomy: include social networking. *Nature*, *467* (7317) 2010, pp. 788-788.
[53] Tsichritzis, D. C. and Lochovsky, F. H. *Data models*. Prentice-Hall, 1982.
[54] Wand, Y. and Wang, R. Y. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39* (11) 1996, pp. 86-95.
[55] Wand, Y. and Weber, R. An Ontological Model of an Information-System. *Ieee Transactions on Software Engineering*, *16* (11) 1990, pp. 1282-1292.
[56] Wang, R. Y., Reddy, M. P. and Kon, H. B. Toward quality data: An attribute-based approach. *Decision Support Systems*, *13* (3-4) 1995, pp. 349-372.
[57] Wang, R. Y. and Strong, D. M. Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.*, *12* (4) 1996, pp. 5-33.
[58] Weber, R. Still Desperately Seeking the IT Artifact. *MIS Quarterly*, *27* (2) 2003, pp. 183-183.
[59] Wiersma, Y. F. Birding 2.0: citizen science and effective monitoring in the Web 2.0 world. *Avian Conservation and Ecology*, *5* (2) 2010, pp. 13.
[60] Wiggins, A. and Crowston, K. From Conservation to Crowdsourcing: A Typology of Citizen Science. in *44th Hawaii International Conference on System Sciences (HICSS-2011)* 2011, pp. 1-10.
[61] Zvieli, A. and Chen, P. P. Entity-Relationship Modeling and Fuzzy Databases. in *Proceedings of the Second International Conference on Data Engineering (1986)* 1986, pp. 320-327.
[62] Zwass, V. *Foundations of information systems*. Irwin/McGraw-Hill, 1998.