Data Quality in the Small: Providing Consumer Information

Arnon S. Rosenthal, Donna M. Wood, Eric R. Hughes, Mary C. Prochnow
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
(813) 831-5535 (813) 835-4661 (fax)
{arnie, dwood, hughes, maryproc}@mitre.org

Abstract: Data quality, defined here as fitness for use, is increasingly seen as a serious problem in government and private sector databases. In this paper, we survey available techniques, and then describe our own work.

We are in the process of adapting general data quality techniques suited to large government relational databases, focusing on an aspect rarely seen in the literature, i.e., helping a user assess the quality of individual data records. Our emphasis is on developing solutions to the problem of providing better consumer information on each value used. We provide such information, so the consumer can determine whether the data are good enough for the intended purpose. The primary concern is with individual data items that drive major decisions, where erroneous data have high cost (e.g., human lives). The broad aim is to enable better decisions. A narrower aim is for consumers to trust data when appropriate, thereby reducing the incentives to ignore the data or expend effort on workarounds for data of unknown quality. This paper explains where our approach fits in the spectrum of data quality approaches, and describes a methodology for providing consumers with information needed to guide how they use each data value in making decisions. The methodology encompasses the following aspects:

- Providing an infrastructure to define, store, and make available quality attributes on various data records
- Obtaining values for quality attributes on important data granules
- Making the quality attribute values available to systems and people that use each data granule
- Tracking the impact of providing the quality values on decision-makers and decisions

Key words - data quality annotations, quality annotation methodology

1. Introduction

Data quality, defined as *fitness for use*, is increasingly seen as a serious problem in government and private sector databases. We are currently involved in a research project that has produced a methodology for providing consumers with information needed to guide how they use each data value in making decisions. This paper explains where our research fits in the spectrum of data quality approaches, and discusses our results and future plans.

2. Overview of Data Quality Approaches

There are two basic approaches to improving a system's data quality: *defect reduction* and *consumer information.*

Defect reduction efforts receive more attention in the literature. The mainstream of data quality research and products seems driven by data warehousing, enterprise resource planning systems, customer relations, and direct mail. For such efforts, one typically gathers impressions or statistics about the quality of large sets of data (e.g., all customer deliver-to addresses), the benefits of improved quality for each category, and the likely costs of improvement. One then alters the data acquisition and cleaning processes to improve the data values stored within the database [Red97]. Many government applications use non-rigorous, informal methodologies for defect reduction.

Consumer information efforts aim to make the existing data more usable, by adding information. One aspect is to better document how one interprets the *meaning* of the data (for example, just how 'Threat' is defined in Army applications or whether a French unit reports distance in meters, feet, or kilometers). Understanding the meaning is particularly important when connecting an automated application, which may not realize that 5 feet is a ridiculous distance for a tank sighting report. Because meaning is covered in the extensive data integration literature [Bln86, Rah01, Mil01], we will not consider it further here.

We focus instead on an aspect rarely seen in the literature, i.e., helping a data consumer assess individual data values. We are concerned with individual data items that drive major decisions, where erroneous data have high cost (e.g., loss of life). However, we find that the same quality measures can be used for both defect reduction and consumer information.

Our task therefore goes beyond the data quality marketplace. Traditionally, data are byproducts of providing goods or services; for our customers, information may be the primary product. Traditional efforts often use data for routine automated transactions; there is little human involvement with each data instance. Errors there are costly in the aggregate (e.g., wasting 10% of a direct mail campaign), but a single wrong data value rarely causes loss of life (or the equivalent in corporate motivation and survival, catastrophic financial loss). In government settings, a human typically inspects the data before a decision is made.

We aim to provide the consumer with a better picture of an individual item's quality so he can determine whether the data are good enough for the intended purpose. The broad aim is to enable better decisions. A narrower aim is for consumers to perceive the databases' contents as trustworthy, thereby reducing the incentives to ignore the data or expend effort on workarounds.

In other words, our methodology addresses both the quality of aggregate data sources as well as that of individual data items.

2.1 Project Setting

Our research focuses on two operational databases, their interactions with each other, and a subset of their data producers and consumers. In our domain, data necessarily give an imperfect picture of the external world. While we are performing informal studies for defect reduction, our efforts have focused on providing information so data consumers can more appropriately and confidently employ the data that are available.

For defect reduction, we conferred with data providers, system managers, and some users of these databases to identify individual data attributes whose quality was perceived as problematic. Several common issues emerged:

- The effect of semantics (data item meaning) on data quality
- The effect of business rules on data quality
- The cause and effect of inconsistencies between databases

- The effect of [poor] data quality on the enterprise
- The effect of database structure on data quality

We selected the quality measures (derived from the literature [Str94]) that would describe the problem to guide future efforts. The central idea is to allow consumers to see quality values for the data they retrieve. We define each step so that the process can be repeated. It is interesting that when compared with prior data quality methods (e.g., [Wang93]), the steps line up fairly exactly, but many of them took a radically different form.

Figure 1 illustrates two variations of the specific case that we investigate. In both instances, no quality annotations are provided in the database. When a consumer accesses data directly from the source, as depicted in the top flow of Figure 1, it is obvious that the consumer cannot ascertain quality. The second case, depicted in the lower flow of Figure 1, presents the problem of consumers accessing data that have been derived from a source database. In this case, even if the producer database contains quality indicators and the consumer is aware that the quality indicators exist, there is no mechanism to derive quality values as information flows from one system to the next.

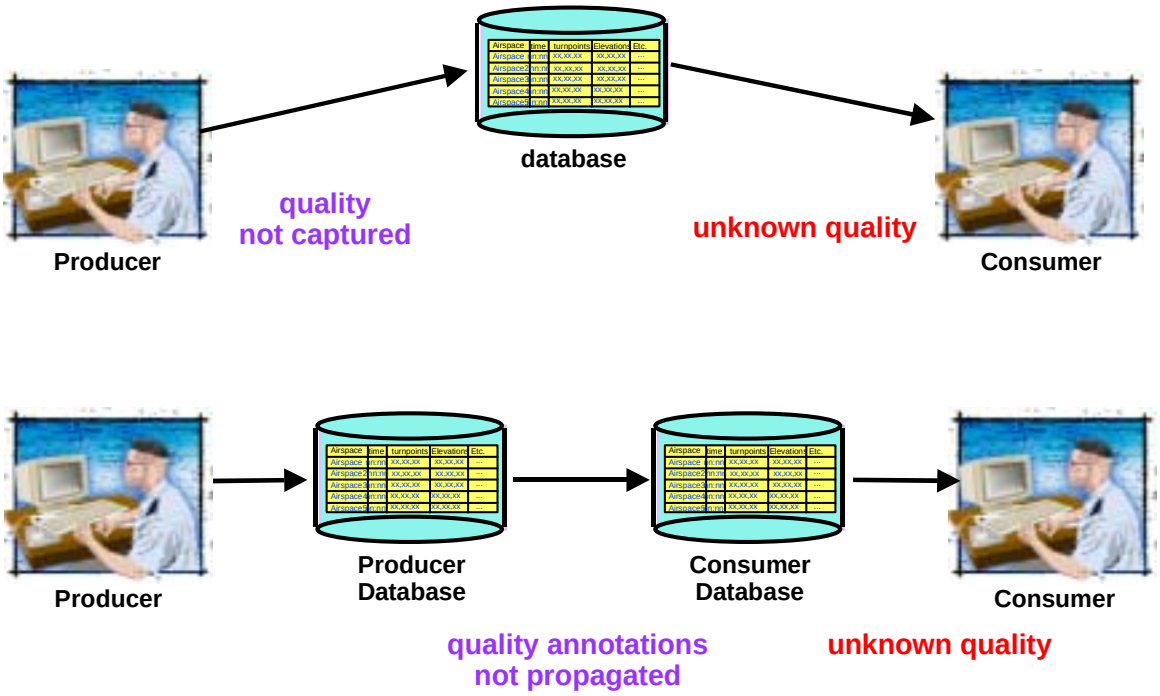

Figure 1. Problems with Understanding Data Quality

2.2 Overview of Methodology for Providing Consumer Information

Our research has revealed that in order to provide the data consumer with the information needed for critical decisions, these tasks must be accomplished:

- Provide an infrastructure to define, store, and make available quality attributes on various data items
- Obtain values for quality attributes on important data granules
- Make the quality attribute values available to users and systems
- Track the impact of providing the quality values on decision-makers and decisions

We assume that these technical tasks will be performed in the context of a business process reengineering effort designed to fix the problems with the processes that are used to provide, store, and access shared information. The next section provides more detail.

3. Methodology for Providing Consumer Quality Information

We begin with a data quality annotation infrastructure. The infrastructure's basic task is to allow administrators and other authorized users to attach values of data quality metrics, as annotations, to various chunks (*granules*) of the database, and to make this information available to other users of those granules. These concepts are defined below.

A *data quality (DQ) metric* describes the usefulness of some data. Popular examples include measures of accuracy, precision, source, completeness (for sets), and time of observation

[Fox95]; others may be derived from these; still others may be collected. Our infrastructure can contain quality metadata on quality values – after all, quality values are data. However, our investigations have not pursued this second order effect. In the future, we may track information *utility* (i.e., benefit of having it), both for ordinary data and for metrics; utility can be a function of quality.

An *annotation* is a triple, (annotated-object, annotation type, value) that is logically attached to some granule of a database. An alternative logical view might include some of the annotations as part of the regular database structure. The ordinary interface does not show whether an annotation is physically co-located with the annotated object (e.g., for image metadata) or stored separately. We use typed annotations so that many systems can use the same definitions of quality measures, which might be standardized for particular domains. The next section describes this concept in more detail.

3.1 Provide an Infrastructure to Define, Store, and Make Available Quality Attributes on Various Granules of Data

While not strictly part of the methodology, it is interesting to understand the infrastructure provided to support the data quality work. The primary requirement is to be non-intrusive. The infrastructure is able to employ existing data that provide quality information (e.g., dates of capture, error bounds), as well as store separately-provided knowledge, at cell, column, row, and table granularities. A more sophisticated infrastructure could capture knowledge as rules (e.g., If Year=1996 and company=MITRE then EarningsAccuracy < 0.95). Added metrics are stored separately from the application tables. The infrastructure provides operations to administer, update, and retrieve data quality metrics.

Administration comprises definition, annotation administration, and physical administration. One can define types of annotations (e.g., *accuracy, time-captured, source*) as ordinary data attributes. For each, one supplies a data type, value constraints, and prose describing its meaning. For each attribute that receives that type of annotation, an administrator specifies 1) rules that derive values from contents already in the database, or 2) that explicit storage be allocated. For example, one of our subject databases contains many attributes that describe how a datum was obtained, and an estimate of its currency. These are logically derived into annotations, but need not be physically replicated. Annotation administration controls are shown, by default, as part of the annotation user interface.

The infrastructure maintains the relationship between an annotation value and a granule in the database, e.g., a table, row, column, or cell. Annotations are updated as ordinary database data. Access permissions can either be derived from those for the annotated data, or managed as for any other data. For read, one can get annotations exactly on a granule, or include super- and/or sub-granules. The infrastructure provides a generic query interface that presents annotations as additional columns of the annotated table. The semantics are those of an ordinary database view.

Finally, we note that the infrastructure works for any kind of annotations one wishes to attach to data values – it is not specific to quality information other than the types of annotations we

have defined. Database researchers are making interesting progress on all sorts of annotations [Delc01, Bird00].

3.2 Obtain Values for Quality Attributes on Important Data Granules

Again, intrusion and extra work are minimized.

The first step is to determine what quality metadata is already provided in the database schema. If possible, we will get providers' agreements to continue supporting these attributes, and to provide fill for them. (One of our subject databases contains many attributes describing data acquisition and processing, and these provide much of the necessary information.)

Beyond this, we intend to capture wholesale rules that describe all the instances provided by a data feed. This is much cheaper than manually creating each instance. In one of our subject databases, this approach can be used to derive completeness and consistency measures, and estimate precision and accuracy for geospatial coordinates.

To plan gathering of further quality information, one works from two sides – need and ease of capture. For need pull, we determine what quality data would make a difference, and be desirable to obtain. As a form of push, we capture quality annotations that are cheap to get (e.g., time of entry, source).

Builders of data capture software have the option of enforcing data constraints, which sometimes improve data quality. These include value constraints and referential constraints (i.e., that subsidiary data must refer to entries already in the table). Ease of use must be considered. In Desert Storm, data providers disliked the constraints, and moved much of their content to free-text fields. An alternative, supported by our approach, is to record constraint violations as annotations for later attention, e.g., to check for alternate spellings.

But even the best data capture software cannot provide fill where none is available, nor recheck to determine if an office has moved or a company has a new vision. Some observations are inherently unreliable (e.g., number of people in an organization). For these cases, quality metrics should be provided.

Where part of a record fails the quality checks, we want a means of capturing the good part. (In the past, one government system lost considerable data that its data-passing interfaces found somehow faulty; the overall effect of these interfaces on data quality was detrimental.) One approach is to set "bad" values to null, with an annotation holding the suggested value so it is not lost. Automated applications will need to be null-aware, i.e., to behave correctly with null data.

3.3 Make the Quality Attributes Values Available to Users of Each Data Granule (Including Both Humans and Queries)

We explore two means of providing quality values to users -- non-intrusively. Figure 2 illustrates our concept. The producer provides quality annotations, which are captured by our tool in a separate but related database. These annotations are then propagated either directly to the

consumer, or to the consumer database. We note that the consumer needs control over whether screen space is devoted to these extra columns when displaying the results of database queries. In both cases, we have provided a very basic implementation. We also modified two existing user interfaces: one for querying, and the other for map display. In both cases, once the implementers understood the user interface's code, a few hundred new lines sufficed. We are convinced that a fuller implementation need not be very difficult. Generic interfaces for annotations should be provided for the most common forms, i.e., relational (which we have completed) and Extensible Markup Language (XML).

3.4 Track the Impact of Providing the Quality Values on Decision Makers and Decisions

We anticipate that it will be very hard to track the impact of quality metadata on user decisions. Several techniques seem natural. For now, we lean toward using only the first, which is least intrusive:

- Use interfaces that make display of quality values optional, generating different queries based on what quality values the user wants retrieved. We can track whether users include quality annotations in their displays (though not its influence on their decisions).
- Survey users about what they use and how valuable it is.
- Provide a box for rating the utility of metadata, as part of the user interface. (For example, Amazon.com lets users rate the utility of feedback from a reviewer.)

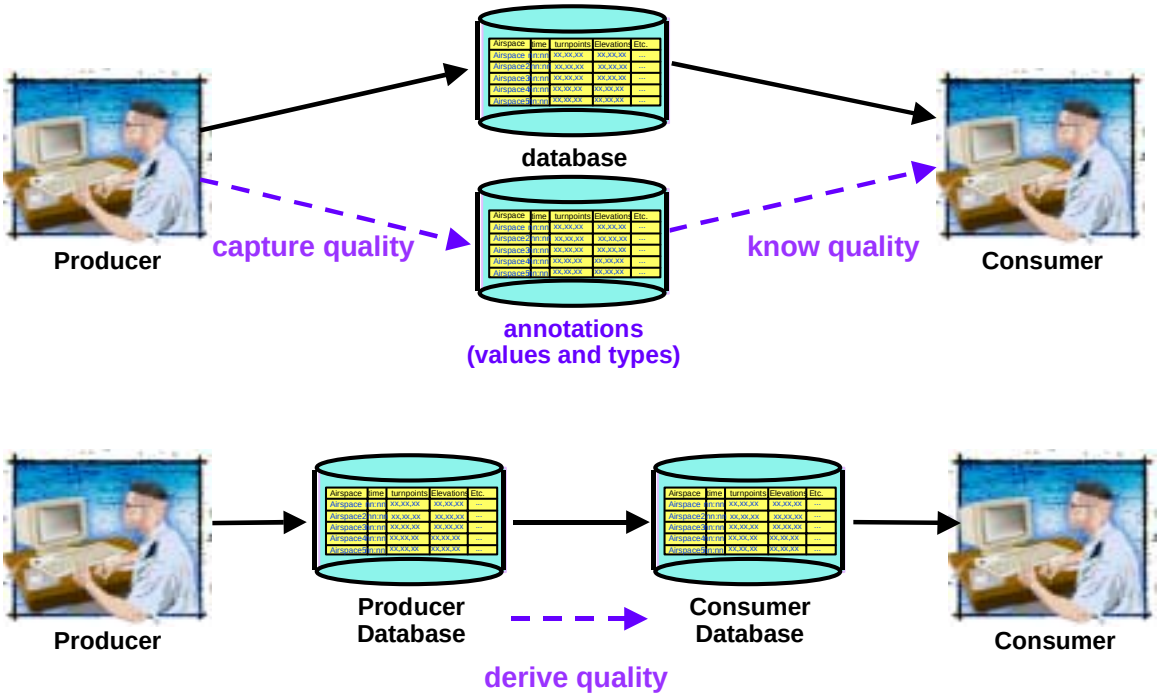

Figure 2. Benefit of Consumer Quality Information

4. Conclusions and Future Work

We have shown how database systems can be extended to manage data quality annotations on base data. Our critical next step is to engage real users, and improve the approach based on their feedback.

Figure 3 illustrates some additional future directions. We hope to provide the means for creating and managing tailored views (comprising both query and display) for communities of interest (COIs), and for adding data quality capabilities to these views. We aim to reduce the cost and delays in producing and maintaining tailored interfaces, thereby enabling better ones to be provided. To do so, we plan to build a componentized view capability that can address both query and display aspects of an interface. We will show how this view capability will allow COIs to form dynamically, collaborate through a view, and update data via the view. We also intend to investigate mechanisms that allow users to provide feedback on quality annotations. In addition, we will consider techniques to dynamically choose source information with quality annotations considered.

Portions of this paper have been derived from A. Rosenthal et al., "Methodology for Intelligence Database Data Quality", AFCEA Database Colloquium, August 2001.

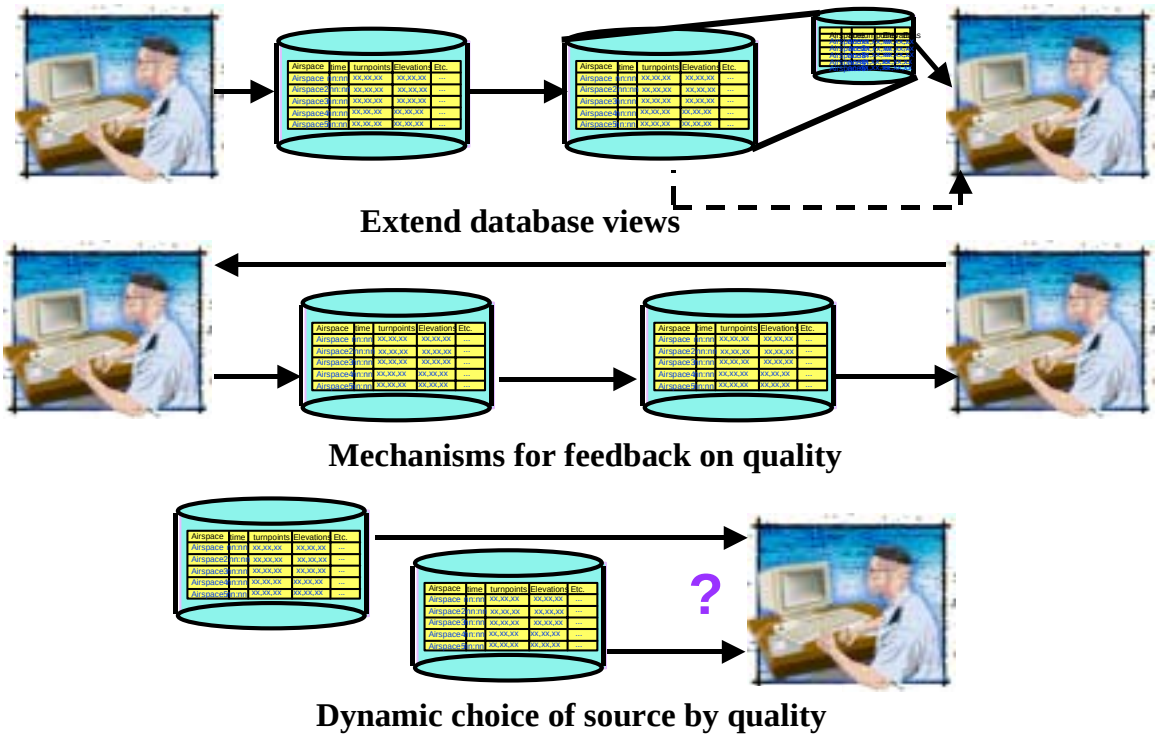

Figure 3. Future Directions

References

[Bird00] S. Bird, P. Buneman, W-C Tan, "Towards a Query Language for Annotation Graphs," *Conference on Language Resources and Evaluation 2000.*

[Bln86] C. Batini, M. Lenzerini, and S. Navathe, "A Comparative Analysis of Methodologies for Database Schema Integration, " *ACM Computing Surveys*, 18(4):323-364, 1986.

[Delc01] L.Delcambre, D. Maier, et. al., "Bundles in Captivity: An Application of Superimposed Information," *IEEE International Conference on Data Engineering 2001*.

[Fox95] C. Fox, A. Levitin, and T. Redman, "The Notion of Data and Its Quality Dimensions," Massachusetts Institute of Technology (MIT) Sloan School of Management TDQM-95-08, February 1995.

[Mil01] R. Miller, M. Hernandez, L. Haas, L. Yan, C. Ho, R. Fagin, L. Popa, "Clio: A Semi-Automatic Tool For Schema Mapping," *ACM SIGMOD Record, web edition, March 2001*. http://www.acm.org/sigmod/record/issues/0103/index.html

[Rah01] E. Rahm, P. Bernstein, "On Matching Schemas Automatically," Tech. Report 1/2001, Comp. Science Dept., U. Leipzig, Feb. 2001, http://dol.uni-leipzig.de/pub/2001-5, to appear in VLDB Journal.

[Red97] T. Redman, "Data Quality for the Information Age," Artech House, 1996.

[Str94] D. Strong and R. Wang, "Beyond Accuracy: What Data Quality Means to Data Consumers," MIT Sloan School of Management, Cambridge, MA TDQM-94-10, October 1994.

[Wang93] R. Wang, H. Kon, S. Madnick, "Data Quality Requirements Analysis and Modeling," *IEEE International Conference on Data Engineering 1993.*

Appendix A Comparison With Other Methodologies

To understand the novelty of our work, we compare with a methodology motivated by finance and industrial databases [Wang93]. The points of difference are:

- Government organizations often have individual data values that drive important decisions, and are evaluated by humans, rather than by tools.
- Some methodologies suggest filtering out data that do not meet standards. In many government applications, one uses the best data available – but more cautiously.
- Aging is a major problem with much of our data.
- Coverage can be *very* sparse, and too costly to increase.
- Other methodologies require the data administrator to estimate quality, since data are entered by clerks. With professional information analysts, one may often get good estimates.
- Our research is investigating the improvement of an existing system, not the design of a new system from scratch.
- Our methodology has no need to treat subjective and objective metrics differently.