

# Ten Mistakes to Avoid in Data Quality Management

## *Foreword*

The corporate data universe consists of numerous databases linked by countless data interfaces. While the data continuously moves about and changes, the databases and the programs responsible for data exchange are endlessly redesigned and upgraded. Typically, this dynamic results in information systems getting better—while data quality deteriorates. This is unfortunate, since quality is what determines data's intrinsic value to businesses and consumers. Information technology magnifies this intrinsic value. Thus, high quality data, combined with effective technology, is a great asset, while poor quality data, combined with effective technology, is an equally great liability.

Yet we tolerate enormous inaccuracies in databases and accept that most of them are riddled with errors—while corporations lose millions of dollars because of flawed data. Even more disheartening is the continuous growth in the magnitude of data quality problems, fostered by exponential increase in the size of databases and the further proliferation of information systems.

This *Ten Mistakes to Avoid* offers advice on how to steer clear of common pitfalls and build an efficient data quality management program.

## ***Mistake One: Inadequate Staffing of Data Quality Teams***

“Who should be responsible for data quality management?” is a frequently asked question in the data quality profession. Uncertainty exists partly because the profession is still in its infancy; no clearly defined group has the appropriate expertise and responsibility. Even companies that form data quality departments often staff them with employees who have expertise in general IT and data but who have no specific data quality knowledge.

Recently, I received an e-mail from one of my conference class attendees, an employee of a household name corporation, who wrote: “I am new to the data quality management world and have found myself in charge of enterprise-wide data quality here at ....” This single sentence explains the major challenge to our profession: Most people in charge of data quality initiatives lack data quality experience. As a result, data quality management programs tend to follow one of two scenarios.

1. Data quality initiatives fall into the laps of technical people within the IT group. For example, an attendee of another of my classes—a database administrator—was asked by her boss to outline a data quality assessment strategy. Why her? According to her boss, she was picked because data quality assessment involves writing queries, manipulating data, and understanding databases—skills that were part of her resume. Such reasoning makes as much sense as asking me to be a sports reporter for the Chicago Tribune because I can type, I've published some articles, and I watch sports a lot from the comfort of my living room couch.
2. Initiatives are spearheaded inside business units by the data users. This scenario appears to make some sense. The data users themselves can tell good data from bad, and because they are the ones most in need of quality data, business departments sometimes initiate their own data quality management projects. Of course, the problem is that business users lack technical expertise, which is why I get this question at almost every class: “Is there a tool that can manage data quality without any custom coding or querying?” My answer does not please those asking the question. I tell them, "Data quality management is an IT discipline and requires IT expertise."

As it takes two to tango, so a data quality management team must include both IT specialists and business users. In addition, a team needs data quality experts—those who have firsthand experience in designing, implementing, and fine-tuning data quality rules and monitors.

## ***Mistake Two: Hoping That Data Will Get Better by Itself***

One of the key misconceptions is that data quality will improve by itself as a result of general IT advancements. Over the years, the onus of data quality improvement was placed on modern database technologies, better information systems, and sophisticated data integration solutions. I remember an HR executive telling me that because his company had implemented a modern HR and payroll system, the company's data quality problems were solved. "PeopleSoft does not have data quality problems," he said confidently. How could I respond? I said, "I cannot tell you whether or not PeopleSoft has data quality problems, but your company certainly does, and they are not going away. In fact, they are about to get much bigger." He did not believe me; people do not like those who predict rain on a sunny day. A few years later, that company's new HR executive called me for advice about their data quality strategy.

In reality, most IT processes affect data quality negatively. New system implementations and system upgrades are a major source of data quality problems. Data integration interfaces create thousands of errors in the blink of an eye. Even routine data processing is prone to error. Thus, if we do nothing, data quality will continuously deteriorate to the point at which the data becomes a huge liability. The only way to address the data quality challenge is by a systematic, ongoing program that assesses and improves existing data quality levels, and continuously monitors data quality and prevents its future deterioration as much as possible.

### ***Mistake Three: Lack of Data Quality Assessment***

Nearly all data quality management programs focus on data quality improvement. A major obstacle on the path to higher data quality, however, is that most organizations—aware of the importance of data quality—are unaware of the extent of the problems with their data. Their knowledge of data quality problems is usually anecdotal, rather than factual. Typically, organizations either underestimate or overestimate the quality of their data, and they rarely understand the impact of data quality on business processes. These two pitfalls cause the failure of many BI projects. Furthermore, data quality improvement initiatives, when put in place, often fail because no method is provided for measuring data quality improvements.

Assessment is the cornerstone of any data quality management program. It helps describe the state of the data and advances understanding of how well the data supports various processes. Assessment also helps the business estimate how much the data problems are costing it.

Data quality assessment that is comprehensive and recurring allows the business to:

- Set expectations for data quality in new DW and BI applications and reduce the number of unwelcome surprises when the data quality fails to support the new applications
- Plan and prioritize data cleansing initiatives and evaluate the potential ROI of data cleansing
- Understand root causes of existing data problems and investigate ways of improving data collection processes
- Monitor ongoing data quality, identify new problems that creep into the databases, and observe and manage data decay

### ***Mistake Four: Narrow Focus***

Systematic data quality management efforts originated in the 90s from parsing, matching, standardizing, and de-duplicating customer data. Over the years, great strides have been made in this area. Modern tools and solutions allow businesses to achieve very high rates of success. A good number of organizations have implemented these solutions by now, and it is fair to say that overall corporate customer data quality is at the highest level ever. This progress makes many organizations feel good about their data quality management efforts.

Unfortunately, the same cannot be said about the rest of the data universe. Data quality has continually deteriorated in the areas of human resources, finance, product orders and sales, loans and accounts, patients and students, and myriad other categories. Yet these types of data are far more plentiful and certainly no less important than customer names and addresses.

The main reason we fail to adequately manage quality in these data categories is that their structure is far more complex and does not allow for a “one size fits all” solution. More effort and expertise are required, and data quality tools offer less help. Until organizations require data quality management programs to focus equally on all of their data, we cannot expect significant progress.

### ***Mistake Five: Bad Metadata***

The greatest challenge in data quality management is that actual content and structure of the data is rarely understood. More often, we rely on the theoretical data definitions and data models. Since this information is usually incomplete, outdated, and incorrect, the actual data looks nothing like what is expected. The solution is to start data quality management programs with extensive data profiling—the term used to describe a collection of experimental techniques aimed at examining the data and understanding its actual structure and dependencies.

Comprehensive data profiling includes the following components:

- **Subject profiling:** Examines subjects in different tables or on different systems and helps to find where the information about each subject is stored.
- **Relationship profiling:** An exercise in identifying entity keys and relationships as well as counting occurrences for each relationship in the data model. It is necessary to validate existing relational data models or build them when none are available.
- **Attribute profiling:** Examines values of individual data attributes and provides information about frequencies and distributions of their values. It helps to identify meaning and allowed values for an attribute.
- **Timeline profiling:** Looks for patterns in historical data, such as temporal distribution of the data, patterns of values for different time periods, etc.
- **State-transition model profiling:** Examines lifecycle of state-dependent objects and provides actual information about the order and characteristics of states and actions. It helps build or validate state-transition models.
- **Dependency profiling:** Uses various pattern recognition techniques to find hidden relationships between attribute values.

### ***Mistake Six: Ignoring Data Quality During Data Conversions***

Data warehouses begin their life with data conversions from various operational databases—usually a rather violent beginning. Data conversion usually takes the better half of the implementation effort and almost never goes smoothly.

Every system is made of three layers: database, business rules, and user interface. What users see is not what is actually stored in the database, especially in older “legacy” systems. During data conversion, the data structure is usually the center of attention. The data is mapped between old and new databases. However, since the business rule layers of the source systems are poorly understood, this approach inevitably fails.

Another problem is the typical lack of reliable metadata about the source database. Consider how often we find value codes in the data that are missing from the mapping documents—all the time. When such a basic component is incorrect, how can we believe any metadata? Yet, over and over again, data conversions are made to specifications built on incomplete, incorrect, and obsolete metadata.

The quality of the data after conversion is directly proportional to the amount of time spent to analyze and profile the data and uncover the true data content. Unfortunately, the common practice is to “convert first and deal with data quality later.” The ideal data conversion project begins with data analysis, comprehensive data quality assessment, and data cleansing. Only then can we proceed to coding transformation algorithms.

### ***Mistake Seven: Winner-Loser Approach in Data Consolidation***

Most data warehouses draw data from multiple operational systems. The need to consolidate data from multiple sources adds the new dimension of complexity to basic data conversion, as the data in the consolidated systems often overlap. There are simple duplicates, overlaps in subject populations and data histories, and numerous data conflicts.

The traditional approach is to set up a winner-loser matrix indicating which source data element is picked up in case of a conflict. For instance, date of birth will be taken from System A if present, from System B otherwise, and from System C if it is missing in both A and B. This rarely works because it assumes that data on System A is always correct—an illogical assumption. To mitigate the problem, the winner-loser matrix is usually transformed into a complex conditional hierarchy. At some point, the decision tree becomes so complex as to be impossible to manage; it yields good results for only the simple indicative data elements. The approach inevitably fails for complex historical data, such as event histories and state-transition histories.

The correct approach to data consolidation is to view it in a similar light as data cleansing. We select one of the data sources as the primary data source for each data element; design a comprehensive set of tests comparing the data against other sources, and then use these additional data for “data cleansing.” Once the data in the primary data source is correct, we convert it to the target database.

### ***Mistake Eight: Inadequate Monitoring of Data Interfaces***

It is not uncommon for a data warehouse to receive hundreds of batch feeds and uncountable real-time messages from multiple data sources every month. These ongoing data interfaces can be usually tied to the greatest number of data quality problems. The problems tend to accumulate over time and there is little opportunity to fix the ever-growing backlog as we strive toward faster data propagation and lower data latency.

Why do the well-tested data propagation interfaces falter? The source systems that originate the feeds are subject to frequent structural changes, updates, and upgrades. Testing the effect of these changes on the data feeds to multiple independent downstream databases is a difficult and often impractical step. Lack of regression testing and quality assurance inevitably leads to numerous data problems with the feeds anytime the source system is modified—which is all of the time!

The solution to interface monitoring is to design programs operating between the source and target databases, which are entrusted with the task of analyzing the interface data before it is loaded and processed. Individual data monitors use data quality rules to test data accuracy and integrity. Their objective is to identify all potential data errors. Advanced monitors that use complex business rules to compare data across batches and against target databases identify more problems. Aggregate monitors search for unexpected changes in batch interfaces. They compare various aggregate attribute characteristics (such as counts of attribute values) from batch to batch. A value outside of the reasonably expected range indicates a potential problem.

### ***Mistake Nine: Forgetting About Data Decay***

If a caterpillar has turned into a butterfly but is still listed as a caterpillar on the finch's menu, the bird has a right to complain about poor data quality. In human affairs, people move, get married, and die without filling out all necessary forms to record these events in each system where their data is stored.

The data is accurate only if it represents real world objects. This assumes perfect data collection processes, of course, and in reality, object changes regularly go unnoticed to computers. Thus, accurate data can become inaccurate over time, without any physical changes made to it.

In this age of numerous interfaces across systems, we rely largely on changes made in one database to be propagated to all other databases—which does not always happen. For instance, interfaces often ignore retroactive data corrections. Alternatively, IT personnel may make changes using a backdoor update query, which, of course, does not trigger any transactions to the downstream systems.

Whether the cause is a faulty data collection procedure or a defective data interface, the situation when the data gets out of sync with reality is rather common. The solution to the problem is recurrent data quality assessment and sample comparison against trusted sources. This provides information about the rate of decay and shows the categories of data that are most prone to quick decay. Such knowledge can be used to improve data collection procedures and data interfaces.

### ***Mistake Ten: Poor Organization of Data Quality Metadata***

Data quality initiatives produce enormous volumes of valuable metadata. Data quality assessment tells us about existing data problems and their effect on various business processes. When done recurrently, assessment also shows data quality trends. Data cleansing determines causes of errors and possible treatments. It also creates an audit trail of corrections so that, at a later point, we can discover how a particular data element came to look the way it does. Interface monitoring identifies ongoing data problems and tells about data lineage, as does data conversion and consolidation.

The common problem in data quality management is inadequate architecture of the data quality metadata repositories. Data quality assessment projects routinely generate innumerable unstructured error reports with no effective way of summarizing and analyzing the findings. Data cleansing initiatives typically lack audit trail mechanisms, and ETL processes often lack data lineage information. As a result, the value of the data quality initiatives is greatly diminished. In the worst-case scenario, the projects are totally abandoned.

The solution is to design a comprehensive data quality metadata warehouse (DQMDW), which is the collection of tools for organization and analysis of all metadata relevant to or produced by the data quality initiatives. It is a rather complex solution, combining elements of object-oriented metadata repository with analytical functionality of a data warehouse. However, in absence of a well-designed DQMDW, data quality metadata will suffer from the very malady it is intended to cure—poor quality.

### ***About the Author***

Arkady Maydanchik is a recognized practitioner, author, and educator in the field of data quality and information integration. His data quality methodology and his breakthrough *ARKISTRA* technology have been used to provide services to many organizations. Arkady is the author of *Data Quality Assessment for Practitioners*, is a frequent speaker at various conferences and seminars, and contributes to many journals and online publications.