Published in DM Review in August 2004.
Printed from DMReview.com

Data: An Unfolding Quality Disaster

by Thomas C. Redman

Summary: *No industry, company within any industry or any department within any company is immune to the effects of poor quality data. While most effects are barely observable, the cumulative impact of poor data quality is enormous.*

It is trite to observe that data is a critical asset in the information age. Data is the "facts and figures" associated with customers, products and services, market and financial performance - indeed, every aspect of life in the information age. It is used to conduct every operation, no matter how mundane, and it is a crucial input to decision making and planning. Further, the sheer quantity of data acquired and stored by companies and government agencies is growing by leaps and bounds. The following quote of Lou Gerstner illustrates the point: "Inside IBM, we talk about 10 times more connected people, 100 times more network speed, 1,000 times more devices and a million times more data."[1] Additionally, increasingly more data is published on the Internet as heretofore proprietary databases are made available to the public. There is no end in sight to any of these trends.

It is becoming increasing clear that much (probably most) data is of poor quality. Some data is simply incorrect, other data is poorly defined and not understood by data customers; still other data is not relevant to the task at hand. The impact is enormous. Poor quality data is at the root of many issues of national and international importance that dominate the news for weeks at a time. Fortunately of course, most data quality issues are more mundane. However, in aggregate, they may be even more costly.

Of course, most data quality issues do not announce themselves as such - many people and organizations are not aware of the importance of the issues. This article aims to shake them from their slumber. It presents a high-level synthesis of so-called data quality disasters and everyday issues that bedevil organizations. The most important point is that poor data quality is an unfolding disaster.

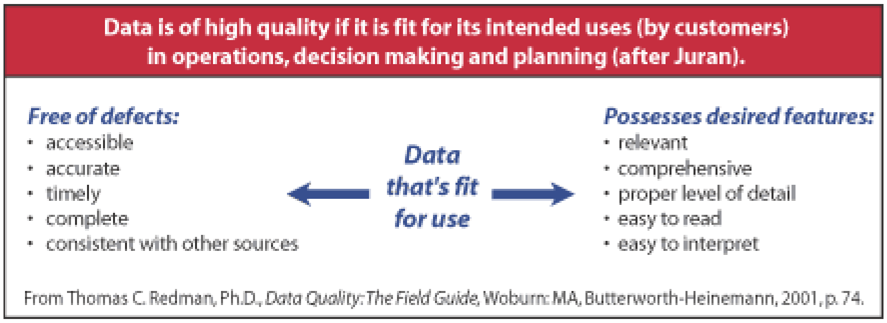
- Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.
- Most data quality issues are hidden in day-to-day work. If they think about it at all, most people and organizations conclude that poor data quality is just a fact of life.
- From time to time, a small amount of bad data leads to a disaster of epic proportions. There is no way to tell when or where the next disaster will occur.

This article focuses solely on building awareness. It stops short of offering prescriptions - they are obvious. They involve extending the tried and true methods of quality management into the realm of data. We do not claim that doing so is easy - data differs from manufactured products in critical ways. However, the extensions have been made and are described in recent books by myself, Michael Brackett, Larry English, David Loshin, Richard Wang and others. Organizations that have applied those prescriptions diligently have made enormous improvements.

The next section of this article defines data quality. The following two sections describe recent data quality disasters and mundane data quality issues, respectively. The section after that synthesizes estimates of the cost of poor data quality (COPDQ) to support our overall estimates.

Data and Data Quality Defined

After J.M. Juran, we define "data to be of high quality if they are fit for their intended uses in operations, decision making and planning."[2] (See Figure 1.) While there are, quite literally, hundreds of dimensions of data quality, a relatively few dimensions are most important in practice. Almost all customers want data that is relevant to the task at hand, easy to understand and correct.



Data is of high quality if it is fit for its intended uses (by customers) in operations, decision making and planning (after Juran).

Free of defects:

- accessible
- accurate
- timely
- complete
- consistent with other sources

Data that's fit for use

Possesses desired features:

- relevant
- comprehensive
- proper level of detail
- easy to read
- easy to interpret

From Thomas C. Redman, Ph.D., *Data Quality: The Field Guide*, Woburn: MA, Butterworth-Heinemann, 2001, p. 74.

Figure 1: Data Quality

As with the quality of manufactured goods, high-quality data stems from well-defined and managed processes that create it, store it, move and manipulate it, process and use it. Thus, data quality involves "getting the right and correct data in the right place at the right time to complete the task at hand."

Data Quality Disasters in the News

For the past several years, data quality disasters (though not, of course, labeled by the news media as such) have occurred with striking frequency. These disasters have dominated the national and international news for weeks. The next several paragraphs highlight five data quality disasters in chronological order as they appeared in the media.

In May 1999, during the Bosnian War, the United States inadvertently bombed the Chinese Embassy.[3] The bombing stemmed directly from a data error. The "facts" associated with what was located in the intended target were simply out of date. Instead of a legitimate target, the Chinese Embassy was bombed and three Chinese citizens were killed.

The data quality disaster of the year 2000 was the presidential election. Most people know a sketch of the facts. The election hinged on the vote counts for George Bush and Al Gore in Florida. For weeks, the national and international press followed the machinations of the candidates and various levels and branches of the Federal and Florida governments as they pressed their cases, counted and recounted votes, tried to decide whether a "hanging chad" signified voter intent and maneuvered for advantage with the Supreme Court. In the end, of course, the State of Florida certified that Mr. Bush had carried the state. He won the election.

Since the election, a number of organizations have reexamined both the results and the underlying processes. Most conclude that George Bush was indeed the winner in Florida.[4] However, the deeper analysis of voting processes (voter registration, ballot design and testing, vote counting and so forth) reveals fundamental issues. For example, a CalTech-MIT report on the quality of the electoral process concluded that the vote could be accurately counted only to within two percent nationally. Results may be even worse in some locations.[5]

One might take comfort if aggressive efforts to rectify voting irregularities had proven successful. Not so. In the recent recall election in California, two independent studies found that more than 383,000 votes - 4.6% of those cast - did not have a valid vote on the recall.[6]

Incorrect and/or misleading corporate financial reporting is our next example. The public became aware of the issue with the collapse of Enron and the subsequent fall of Andersen, its auditor. It appears that out-and-out fraud, not simple error, lies at the heart of the Enron debacle. However, Enron is just the tip of the financial reporting iceberg. Literally hundreds of companies have restated earnings over the past several years.[7] This data is simply not accurate. Further, key data may be omitted and the data provided may not be defined clearly enough for the customer - the potential investor - to get a clear picture of corporate performance. As a result, the general public has lost faith in corporate America. Indeed Hank Paulson, Chairman of Goldman Sachs, noted that it was the largest crisis of confidence in 50 years.[8]

Fourth is Jésica Santillán. In a desperate attempt to save her life, doctors gave her a new heart and lungs. She was given a Type A transplant; unfortunately, however, her blood was Type O. Her body rejected the organs and she lapsed into a coma-like state. Subsequent attempts to save her, with other organs, also failed.[9] The cost - at least one life. Unnamed others who might have received the organs given Jésica may have died as well. Jésica's case is but an example. The Committee on Healthcare in America estimates as many as 98,000 unnecessary deaths per year due to error, and poor data quality contributes in many cases.[10]

Our final example involves failures within the intelligence community. Some speculate that had various agencies shared data the September 11, 2001, attacks on America might have been prevented. As this is written (May 2004), the National Commission on Terrorist Attacks Upon the United States, whose final report will appear before publication of this article, is expected to document a long series of blunders by the FBI, CIA and others.[11] It is critical that their recommendations ensure that the right (and correct) data is in the right place at the right time to prevent future attacks.

Most Data Quality Issues are More Mundane

Fortunately, most data quality problems do not make the national news. They are much more mundane, but perhaps no less costly. Most are simple errors in databases that cause an invoice to be incorrect, direct mail to be thrown away, the wrong product to be sent, inventory to be off and so forth. The interested reader may find dozens of examples in the data quality texts noted earlier.

Other data quality problems are more subtle. Many companies have different divisions, and they would like to link data about customers so that they can cross-sell or offer customers better deals, but the data is simply unfit for doing so. The various divisions employ different data formats, they model customers differently and the data is erred, making linkage impossible.

Poor data quality affects decision-makers and planners as well. Marketers, for example, continually complain that they can't understand what's going on in the marketplace and that they simply don't have the data to see how well their products are doing compared to their competitors' products. This hinders their ability to define and implement marketing strategies. Other decision-makers have similar complaints.

Poor data quality also hinders the implementation of new technologies. Data warehouses, enterprise systems and customer relationship management systems have all been bedeviled by poor data - in many cases causing the new technology to fail altogether. The promised gains are never realized.[12]

Finally, company image may be hurt by even mundane data quality issues (we do not consider a restated financial report to be mundane). Some examples: incorrect prices on Amazon.com, where a 1GB memory module normally listed at $999.99 was on sale at Amazon.com for $19.99; hotel rooms at W Hotels sold for $59 instead of $259; and United Airline tickets selling for $5.[13, 14, 15] We may expect similar occurrences as more and more data is exposed to customers via the Internet.

The Cost of Poor Data Quality

Cost of poor data quality (COPDQ) analyses are difficult to conduct. A few costs, such as the cost of error detection and correction, can be measured. Other costs, such as the cost of customer dissatisfaction, are tougher. COPDQ is, at best, an approximate gauge; however, it can be a useful one for understanding the magnitude of the problem.

Consider first the cost of efforts to find and fix errors. While organizations do, from time to time, conduct massive clean-up exercises, most efforts to find and fix errors are embedded in day-in and day-out work. Over the years, we developed the Rule of Ten: If it costs $1.00 to complete a simple operation when all the data is perfect, then it costs $10.00 when it is not (i.e., late, hard to interpret, incorrect, etc.).

The Rule of Ten makes clear why even a few data errors are so costly. If bad data impacts an operation only five percent of the time, it adds a staggering 45% to the cost of operations.

In my latest book, I suggested that a minimum COPDQ for the typical company is

10% of revenue. That estimate was based on a few proprietary studies (that are now somewhat dated) and much anecdote. It is almost certainly too low. A recent survey conducted by The Data Warehousing Institute estimates that in the United States, $611 billion a year is lost as a result of poor customer data (name and address data).[16] This estimate includes the costs associated with our direct mail example. Because the GDP is approximately $10 trillion, poor customer data alone accounts for 6% of the GDP as the cost of poor data quality; and customer data makes up only a fraction of an organization's data.

As noted, it is more difficult to estimate other costs of poor data quality. Customers are often surprisingly unforgiving of simple data errors. They reason, "If you can't get my address right, why should I trust you to perform a complex service?" They then take their business elsewhere. This cost is very difficult to estimate.

It is logical to suppose that poor data leads to poor decisions. While bad decisions do occur, it appears to us that the more frequent problem is that managers delay making decisions or don't make them at all. Even more critically, decisions that aren't supported by fact are much more difficult to carry out. In the face of clear data, most organizations align in support of a decision. Without such data, the people within the organization know that the decision-maker is relying solely on his or her intuition. Those whose intuitions lead them to different conclusions simply do not line up. All of these costs are extremely difficult to estimate.

It is possible, in principle at least, to estimate the COPDQ of technology failures as the cost of those technologies; however, the real cost is much greater, as the organization doesn't realize the benefits of the new technology.[17]

It is even more difficult to estimate the real costs associated with each of the data quality disasters cited. For example, the United States paid $27 million to the Chinese government for the bombing of the Chinese Embassy. The easily measured COPDQ is $27 million. However, the real cost is much greater. The U.S. military, intelligence community and entire government were embarrassed. Even more costly, Sino-American relationships were set back for years. Worst of all, three families lost loved ones. These costs are incalculable. Similarly, the costs due to challenged elections, lost confidence in the financial markets and medical errors are unknowable, but staggering.

To summarize, a COPDQ figure of 10% of revenue is easily defended. Indeed, easily measured costs for customer and billing data alone account for 8% of revenue. We suggest 20% of revenue as a better estimate of the total cost of poor data quality, based on the assumption that unmeasured costs are at least as great as measurable costs.

A Wake-Up Call

Poor quality data is the norm. No industry, company within any industry or any department within any company is immune to its effects. Most issues are mundane. Alone, they are barely observable. However, their cumulative impact is enormous. They:

- Increase cost - at least 10% (and probably as much as 20%) of revenue.

- Anger customers.
- Increase the difficulty of decision making.
- Make it more difficult to implement new technologies.
- Put company image at risk.

Unfortunately, occasionally, bad data causes enormous damage. In some cases, the damage is confined to individual organizations as their data quality woes are discussed in the media. Other times, the damage is on a national scale. In the last 40 months, there have been at least five disasters; and there is no way of knowing when the next will occur.

References:

1. McDougall, Paul. "More Work Ahead," Information Week, December 18-25, 2000, p.22.
2. Juran, Joseph M. and A. Blanton Godfrey, Juran's Quality Handbook, Fifth Edition, p. 2.2, McGraw-Hill, 1999.
3. Meyers, Steven Lee. "CIA Fires Officer Blamed in Bombing of Chinese Embassy," The New York Times, April 9, 2000, p. A1.
4. Cauchon, Dennis and Jim Drinkard, "Bush Still Wins Under 2 Most Common Standards," USA Today, May 11, 2001, Page A1.
5. "Voting: What is, What Could Be," CalTech-MIT Voting Technology Project, July 2001.
6. Konrad, Rachel. "383,000 Missing Votes in California Recall," AOL News, Associated Press, October 10, 2003.
7. McNamee, Mike, Paula Dwyer, Louis Lavelle, Christopher H. Schmitt, "Accounting Wars," BusinessWeek online, September 25, 2000.
8. McGeehan, Patrick. "An Unlikely Clarion Calls for Change," The New York Times, June 16, 2002, p. 3-1.
9. Archibold, Randal C. "Girl in Transplant Mix-Up Dies After Two Weeks," The New York Times, February 23, 2003.
10. Institute of Medicine, "To Err is Human, Building a Safer Health System," National Academy of Sciences, 1999.
11. Shenon, Philip. "9/11 Panel May Not Reach Unanimity on Final Report," The New York Times Online, May 26, 2004.
12. Girard, Kim. "Blame Game," Baseline, March 2002; Ben Worthen, "Nestle's ERP Odyssey," CIO, May 15, 2002; Brian Caulfield, "Facing Up to CRM," Business 2.0, August/September 2001.
13. Wolverton, Troy. CNET News.com, "Cheap RAM a fleeting memory on Amazon," The New York Times Online, March 28, 2001.
14. Costello, Jane. "W Hotels Room Rate Mistake Benefits Some New York Guests," The Wall Street Journal On Line, January 3, 2002.
15. Tedeschi, Bob. "Trying to Keep Computers in Line," The New York Times on the Web, May 20, 2002.
16. Eckerson, Wayne W. "Achieving Business Success through a Commitment to High Quality Data," TDWI Report Series, The Data Warehousing Institute, 2002, p. 5.
17. Bosch, Rob. "$44 Billion Spent on CRM in 2000 - And Customer Relationship Management Is Worse Than Ever," *DM Review*, July 2001.

Dr. Thomas C. Redman, president of Navesink Consulting Group, is an internationally recognized author and speaker. Redman's latest book is ***Data Quality: The Field Guide*** *(Butterworth-Heinemann, 2001). He can be reached at tomredman@dataqualitysolutions.com or 732-933-4669.*