

Data Quality: Should Universities Worry?

A popular exercise in many corporate training courses goes something like this: The class members are asked to imagine that in one of the company offices is a fine antique French desk, recently purchased for \$20,000. On top of the desk sits a brand-new laptop computer, complete with all the bells and whistles, that cost \$3,000, along with a diskette that cost \$1. The diskette contains the only known list of the names and purchases of the organization's fifty largest customers. Now, each trainee is told, a fire has started and the trainee can save the desk, the computer, or the diskette—but only one of the three. Which should be saved?

Naturally, almost everyone in the class chooses to save the diskette. They all immediately recognize that it is not the diskette that is worth saving but rather the data that it contains. Almost everyone intuitively concurs that the data are worth more than \$20,000. The exercise illustrates a point that most people understand but don't think about too often: that data are an extremely important asset. When people do think about this, everyone realizes that the organization with the best data wins wars, crafts the best strategies, makes the best decisions, knows the most about customers, and keeps costs down.

Despite this realization, the quality of most data is extremely low. Poor-quality data come in many forms—missing, late,

and incorrect data are obviously of poor quality. So are poorly defined data, data that are difficult to interpret, and data that are “not quite relevant” to the task at hand. The list can go on and on. Conversely, the goal is to have the right (and correct) data in the right place at the right time so that people can complete operations, make decisions, and plan.

Poor-quality data are the bane of business and government. Do bad data plague

- Incorrect and missing data, causing the misallocation of funds
- Miscommunication of acceptance decisions
- Concerns about grade inflation and the interpretation of an “A”

Of course, examples from outside academia have made even bigger headlines. Indeed, poor-quality data lie at the root of issues that have captured international

attention and will not let go. Two examples suffice. First, as this is written, a U.S. congressional commission is investigating the 9/11 disaster. Some speculate that had the FBI and the CIA shared data, they could have pieced together the clues and prevented the September 11 attacks. In the second example, a simple data error resulted in a mismatched heart and lungs being mistakenly transplanted into seventeen-year-old Jéscica Santillán, eventually killing her. Unnamed others who might have received the organs given Jéscica, but who did not, may have died as well.

Fortunately, most data-quality issues don't make the front page. But virtually every organization is bedeviled by bad data—increasing costs, angering customers, compromising decisions, and making it more difficult for the organization to align departments. One study estimates the costs to U.S. business at over \$600 billion per year for customer data alone.¹

These examples are particularly

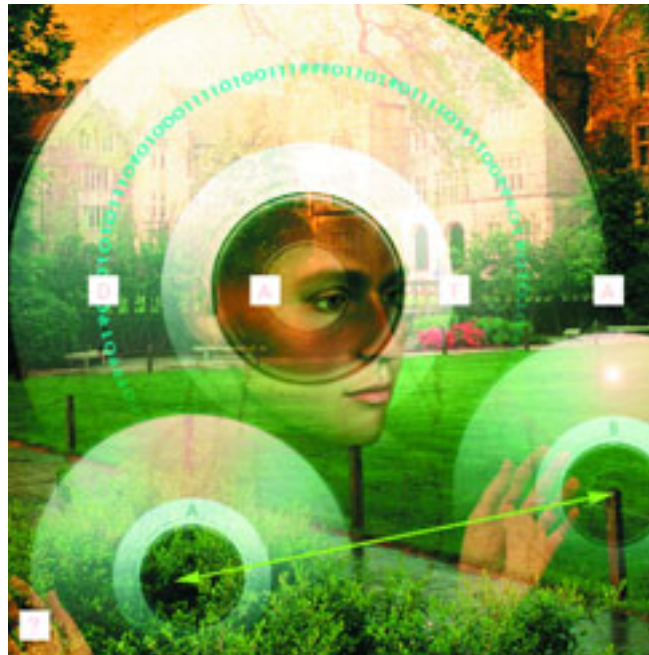


Illustration by Paul Wattson. © 2004

academia as well? A quick search of the “bad data quality makes the news” file turned up numerous examples:

- Incorrect test scores and the resulting impact on students
- Misclassifications of students in ways that lead to the misinterpretation of statistical results

timely in light of Nicholas Carr's *Does IT Matter?* The book follows up on his controversial essay "IT Doesn't Matter," published in the May 2003 issue of *Harvard Business Review*.² Whether or not one agrees with Carr about the future of information technology, one must agree that data quality demands greater attention. But what should college and university IT leaders—already beset by the numerous and conflicting demands of researchers, educators, administrators, and students—do about data quality? Indeed, why should they care? Aren't data the purview of others?

There are no easy answers to these questions. Many factors influence the course of action that any organization should take. However, three simple steps can help lay the framework for an overall data-quality strategy.

The first step is to ask some rather basic questions:

- How much data does the organization have, how fast is it creating new data, and how many redundant copies are there?
- Which data are most important?
- Are there policies that define who is (or which departments are) accountable for these most important data? Are these policies adhered to? (Note: most people assume that the answer to the first question is "the CIO.")
- Are the data of high quality?
- Are sufficient precautions in place to ensure that data are kept secure, held private, and cannot be manipulated?

If the answer to all these questions is "Yes," there is probably no need to take the next step. But if the answer to any question is "No," the second step is to get a feel for data quality. It is usually best to do a "deep dive" into a couple of problems. One area that often proves fruitful is billing data. Incredibly, many organizations don't bill for all they are owed. The reasons are numerous (too many customers to bill, too many departments contributing charges, too many systems that don't talk to each other), intertwined (poorly defined processes and unclear accountabilities), and apparently paradoxical (after all, we are talking about money!). Discovering and sorting the reasons will

offer a window into the joys and perils of data-quality management.

One of the more fascinating perils is confusing "data" and "IT." Versions of the following occur with stunning frequency: "The BS [billing system] has screwed up financial aid once again. We'll have to get that new ES [enterprise system]." But the system is not the root cause of poor financial-aid data, and a million-dollar "solution" misses the point. A recent case study involving Stanford University illustrates how a misalignment between "systems" and "process" can prove costly.³

The third step is to conduct an improvement exercise. Many colleges and universities offer courses on quality improvement, Six Sigma, or manufacturing excellence. The techniques taught in these courses are directly applicable to data. It is important to pick an area that is big enough to matter but not so large as to be intractable. The aforementioned area of billing data is often a good choice. Interestingly, those who are billed are adept at finding instances of overbilling and at demanding corrections, but almost no one notices (or points out) underbilling.

So, improving billing data often has the direct result of reducing underbilling. Best of all, the improvements go right to the bottom line.

These three steps will not solve the data-quality problem. But they will help give colleges and universities enough background to craft a strategy for doing so. After all, it is the \$1 diskette that needs to be saved from the fire.

Notes

1. Wayne W. Eckerson, *Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data*, TDWI Report Series (Chatsworth, Calif.: The Data Warehousing Institute, 101 Communications, 2002), <<http://www.dataflux.com/data/dqreport/pdf>>.
2. Nicholas G. Carr, *Does IT Matter? Information Technology and the Corrosion of Competitive Advantage* (Boston: Harvard Business School Publishing, 2004); Nicholas G. Carr, "IT Doesn't Matter," *Harvard Business Review*, May 2003.
3. Debbie Gage, "Stanford University: Hard Lesson," *Baseline*, vol. 1, no. 31 (June 2004), <<http://www.baselinemag.com/article2/0,1397,1609128,00.asp>>.



Thomas C. Redman is President of Navesink Consulting Group, based in Little Silver, New Jersey.