Zen and the Art of Information
What is (Data) Quality?

There I stood, seemingly paralyzed by the persistence of The Question. All features of the room receded into the distance, the ambient sounds first amplified to a deafening din and then reverberating into a diffuse whisper, and then silence. Again and again I was tortured by my seeming inability to answer the basic question "What is Data Quality?" Then I had a cup of coffee and I felt much better.

Happily, the pursuit of superior data quality, and thus useful business information, is a lot less torturous than Robert Pirsig's journey in *Zen and the Art of Motorcycle Maintenance*. But we can reach a similar Zen-inspired conclusion: data quality is more of a journey than a destination. Why is that? It is because getting useful information greatly depends on human effort and time. In time, data (and thus information) will change.

For example, consider the value of your stock portfolio . It can change in a moment due to a profit warning here, a lawsuit there, or a new invention on the horizon. So it goes with all sorts of information in the enterprise. In a matter of seconds, a happy customer can turn into your worst nightmare, or a waffling potential customer can put you over quota because of that extra donut they had that morning.

This article is the first in a two-part series about data quality. In this issue, we'll consider the implications of data quality challenges both for business intelligence and for other areas. In the next issue, we'll talk about the remarkable rise in of one of the most practical and impactful examples of data quality, data hubs.

The Information Equation

But beforeBefore we go any further, we need to make a distinction between data and information. These terms are so commonly used interchangeably that they are both weakened for the purposes of our discussion. So to bring some clarity to this situation, I will defer to Ali El Kortobi and Paul Narth, from the Oracle Warehouse Builder development team. Paul and Ali think about information quality a lot and don't require coffee to get them going about it.

The first thing Ali wrote on the white board when I interviewed him for this piece was:

$$\text{Data} \neq \text{Information}$$

Okay, fair enough. But the Ali went on to make a distinction he then made aboutbetween data and information that was so elegant and clear, that I've named a theorem after its inventors. Here is the El Kortobi-Narth theorem:

$$\text{Information} = \text{quality (data + metadata)}$$

In plain English, this states that **Information** (that is, human-human-readable, no-batteries-required information) **equals Data** (**Data** (that our computers love to collect and trade), **plus MetadataMetadata** (which is data about data, or context), **with *quality* applied**. Now for you mathematicians out there don't try to go home and prove this, but let us accept this at face valua. value. The clear distinction between data and information is the application of context and quality.

Data collection is easy, and has gottenbecome steadily easier and cheaper. With seemingly every web site, application and home appliance just dying to tell us how and what they're doing, our friends at EMC, NetApp, and Seagate will sleep soundly tonight

Of course, the creation and maintenance of metadata will greatly effectaffect the information theorem,. Thus it is the quality component of the equation that is the hard (and costly) part. According to a study by the Data Warehousing Institute, U.S. businesses lose more than $600 billion each year due to data quality problems

Quality is Not a Place

Data quality has long been associated with data warehousing, the idea being that the elusive "single source of truth" can be established by extracting data from many source systems, then cleansing (or transforming) the data and loading *over there in that system.* It's nice to be able to sit next to a machine and say, "I've got good data quality here!"

But this notion no longer sufficiently tells the complete data quality story. Data quality can no longer be considered a *many-to-one* proposition. IT professionals increasingly act on the premise that numerous core systems, not just data warehouses, require high quality data to create valuable information. Thus, this *many-to-many* approach to data quality requires the tools and techniques to evolve so that the aggregate level of data quality is higher. To stick with the motorcycle maintenance metaphor, the machine reaches peak performance when all of the individual components – such as the clutch, pistons, gears, and steering perform at – perform at their best.

Oracle has been working on data quality techniqueschallenges for years. Oracle Warehouse Builder is one of the more visible products in this area, a product that Gartner recently moved into theits coveted leadership quadrant for **e**xtraction, **t**ransformation and **l**oad (ETL) tools. This is partly a result of a market catching up with a technology. As Gartner thatputs it, "Market demand is taking extraction, transformation and loading tools into new application areas" as these types of tools are applied "beyond the domain of business intelligence."[1]

This idea is not lost on the Oracle Warehouse Builder (OWB) development team. The main reason that OWB is steadily gaining accolades and new users is our increasing focus on the 'T' in ETL. extractingExtracting (and by extension, loading) data to and from different computer systems is a mature, well-understood process.

[1] Friedman, T. and Gassman, B. *Magic Quadrant for ETL,* Gartner, Inc., *2H04*.

The Oracle database, the application server and various tools have been refined so that data is wonderfully portable and mobile. Moving data and application instances, tables, tablespaces, files, even entire systems and data centers has been highly automated and simplified in more recent generations of our products.

But it is the "T" that looms large. Transformation: a longish word that represents the gargantuan problem of reconciling how myriad applications and data sources can agree. It is this huge collection of data streams, application interfaces and target systems for this data that has multiplied and evolved almost beyond reckoning.

So the notion of a data warehouse, data mart, operational data store or what-have-you as being the ultimate recipient of good data quality and the ultimate repository of a single source of truth is not as satisfying as it once was. In fact, the name "Warehouse Builder" no longer really captures what this product delivers and what it is being used for. Maybe thatthe name should .be changed

Quality on the Grid

Anyone who attended this year's Oracle OpenWorld conference sawenjoyed a detailed view of Oracle's grid computing strategy. We focused on new application design, integration topics and middleware (after givinghaving given grid infrastructure a very thorough treatment the priorprevious year).

fromFrom top to bottom, grid computing is all about achieving better information and better overall quality: quality of service, data quality, maintaining quality in rapid application development, and hopefully better quality of life for mustthe people who maintain IT systems. In infrastructure terms, Oracle's grid mantra of *consolidation*, *standardization* and *automation* is about reducing complexity, reducing the number of variables and creating an environment where better data quality is easier to attain.

But easier does not mean easy. And here we get to my main point for data quality: everywhere people and technology meet, and everywhere two or more technologies meet, you have a data quality opportunity. A choice of opportunities, in fact: you have an opportunity to apply a data quality discipline, or an opportunity to pollute your data streams.

Now, when I say, "everywhere people and technology meet" I am referring to the user interface and the act of data entry. All we can really do here is develop better front-ends, add better error checking mechanisms, and train our people as best we can. But when two or more technologies meet, I am talking about integration.

In a very frank article in the November 2003 issue of Business Integration Journal, author and technology architect Russell Levine talks about "The Myth of Disappearing Interfaces." He states that, "The dirty little secret of integration is that no technology is ever going to resolve the semantic data mapping issue." Levine goes on to say, "Data mapping requires intimate knowledge of the data and how it's used. This can be

accomplished only with the network of knowledge and analytic processing power possessed by higher-order, carbon-based life forms. It takes time and effort."

Damn! I guess we'll have to wait a bit longer for completely automated application-integration tools, flying cars, and hotels on the moon. For the time being, the best we can hope for is a technology infrastructure that doesn't get you out of bed at two in the morning and a good set of tools that gets the most out of that brain of yours.

On the Road Again

Data quality has evolved and must continue to evolve into a ubiquitous operation, a *de rigueur* discipline, a Zen-like practice;. It should be as much a part of your IT operations as your system backup and recovery strategy or your hardware maintenance schedule. And although achieving high levels of data quality can at times seem like what Pirsig would call "a continually receding horizon where perfection is impossible," take heart. Knowing that you've got a smoothly running engine, dry pavement in front of you and that you are heading in the right direction, you might as well enjoy the scenery. It's bound to improve.

[All due acknowledgement to Mr. Pirsig for my ofsomewhat haphazard references to his grand work. If you haven't read it, put it on your reading list. It's a classic.]