Data, Information, Knowledge (Newsletter)*

April 19, 2010

All regulated industries – and their regulators – depend on data. Yet data (observations and measurements) are meaningless until transformed to information and used to create knowledge and insight. Too often the tools used to collect, analyze, store, and manipulate regulatory data are inappropriate because this subject is rarely included in educational curricula or on-the-job training. Required education in statistics, modeling, data storage, and data presentation are not universal. If your position depends on knowledge from observed and measured data you will be well served by better understanding of the issues involved.

Data collection methods, sources of error, accuracy, and precision require separate consideration.

Analyses of data to yield information depend heavily on descriptive and exploratory statistics. The tool most frequently used in business for these analyses is the spreadsheet. Unfortunately, spreadsheet formulas are not necessarily appropriate or correct. Lotus 1-2-3® used the population formula for calculating standard deviation, not the sample formula, and produced inaccurate results. Excel® uses the annuity formula (interest received on an invested amount) for net present value (NPV) and produces incorrect results of a loan's NPV (interest paid on a received amount). We trust these tools because we both expect them to be correct and we cannot delve into the guts of every technical application to ensure it is correct.

Because tools such as spreadsheets are widely used for number crunching they are also used for data storage. Almost always it becomes apparent that this is the wrong tool for the job. The appropriate tool is a well designed database, yet database design (like mathematical and spatial statistics) is not widely taught. While a database table looks like a spreadsheet (both have rows and columns) there are critical differences.

Spreadsheet cells are independent of each other. Formulas link cells by location (column letter and row number), and this location can change if a column is independently sorted. Consider a spreadsheet page used as an office directory. Columns contain last name, first name, office number, and telephone extension. You enter the information for everyone in your office, and

sort it alphabetically by clicking on the last name column, selecting "Sort" from the menu, and pressing the [Enter] key. You now have a Rubik's cube with the last names in alphabetic order, but all other columns unchanged. This happens all too often. Even if everyone accessing that spreadsheet remembers to sort all columns, problems still arise. Consider querying monitoring data to prepare a permit renewal application. You want to know how many monitoring locations were out of compliance, how many times each location was out of compliance, what parameters were involved, the months (or other sampling period) when this occurred, and sorted by frequency of occurrence. Perhaps this could be done with spreadsheet data, but it is the type of query commonly seen in a relational database application.

In a database table the attributes ("columns") in each row are a connected unit. In the above example, sorting your office directory table alphabetically by last name keeps all other attributes associated with that last name. Regulatory data categories (such as permits, monitoring locations, and parameters) are in separate tables related to each other. Tables can each contain millions of rows (try that with a spreadsheet!) and return query answers quickly using indices into the data. Multiple users can view data while one user is updating a table or entering a new row; relational databases are inherently multi-user. Highly complex queries return results that lead to greater understanding of the dynamics of the data and more insight to guide decisions. Databases can also be "mined" for subtle patterns and relationships not immediately obvious. Companies with years of monitoring data can explore these hidden patterns and relationships to make better informed management decisions.

Use of relational databases benefits both regulators and the companies they regulate. Examining trends, patterns, and complex relationships allows regulators to set realistic constraints and surety levels because they are not working blindly and compelled to identify worse case scenarios to use as the basis for permit conditions and bonding.

Spreadsheets can be great for financial analyses with relatively limited data, but relational databases are necessary for complex, extensive data sets such as are generated by regulatory permits and compliance monitoring data. The potential knowledge and insights gained from regulatory data in a database is extensive and fiscally valuable.