Data Base Management Techniques To Ensure Project Integrity

Susan Campbell
U.S. Environmental Protection Agency

Jeffrey Finkeldey
Computer Sciences Corporation

INTRODUCTION

The Drinking Water Research Division (DWRD) of the Risk Reduction Engineering Laboratory, U.S. EPA, participates in numerous cooperative research projects with various utilities, universities, and firms. These studies produce large amounts of data which are used in the study of drinking water treatment technologies. Over the years, DWRD has had the opportunity to see the positive and negative sides of data base development and use with regard to project management. Prior planning can negate major problems at later development stages, while a lack of it can double or treble the amount of work to be performed in the long run. This paper will relate some of the authors' experiences while pointing out some of the management techniques to include and avoid when developing a research project.

OVERVIEW

The object of any data base management system is to provide answers to various questions related to the data, in effect creating useful information. In the case of DWRD's research projects, the main questions relate to different technologies as they affect drinking water quality, as well as the formation of chemical byproducts associated with the treatment process. How efficiently this broad range of chemical and engineering questions can be addressed is directly related to how well the pertinent information is organized and utilized. A brief overview of DWRD's data bases will be given as an example. The majority of DWRD's data bases are relatively large (average 46,000 records), and reside on the EPA's National Computer Center's mainframe computer. The workhorse programming language of the computer staff is SAS, due to the plethora of statistical applications available, and presentation quality graphics capability. DWRD uses an inhouse system, as well as batch and interactive SAS programs written for specific applications. Most of the data are related to removal technology performance, although a future goal is to include operational as well as cost data for the projects. Therefore, the basic data used by DWRD contain fields relating to the date and location the sample was taken, the experiment or phase in effect, the contaminants being tested for, the actual concentration of the sample, and any related notes. Of all the fields within the data bases, the concentration, or reading, field is the one in which accuracy is the most important. Statistics and actual data values from DWRD's data bases are used by the EPA as a basis for drinking water regulations. An incorrect value or mean could ultimately cause a recommended limit to be too tight or too lax. The location, phase, and date fields are also important because of the variations they can cause on the concentration value. Among the various uses of the data are reports of actual data values, statistical correlations, and graphic representations and comparisons. Many of these outputs are routine, and so are 'production' jobs. Most of the graphics, however, are written as needed, due to the changing requirements of the project managers, and newly evolving theories, etc.

Anyone involved in data base management knows the importance of planning ahead to achieve maximum efficiency and durability of the system. DWRD went through an extensive planning stage before its data base management system ever went into use - with over 500,000 records, they could not afford not to. And the system has maintained its usefulness for nearly 10 years, without a major rewrite or upgrade. With some of DWRD's other projects, however, the final product was a little less satisfactory. For example, a test case was implemented which utilized a personal computer and commercial data base management system (DBMS) to make simple analytical observations about factors related to water main breakage. The data were set up in two main files, which the DBMS was to join through an address field for use in the analysis. The two files contained a total of approximately 7000 records. However, the DBMS was never tested with files this large. Subsequently, when the analysis was performed, it was found the PC and DBMS could not handle those size files. Just to do a simple merge and print out (using indexed files no less!) took overnight. As a result the files had to be transferred to a mainframe and the analysis done from there. Fortunately, that project was merely a test case. Clearly, the necessary planning was not performed in the above case. However, since most of DWRD's applications use SAS on an IBM mainframe, such system problems are very rare. There is a need for planning though, when setting up a data base system, even when the hardware is entirely adequate.

Possibly the most basic aspect of the development stage is to make sure the data base will answer the questions it was designed to answer, and turn the data there into usable information. Besides the obvious matter of sampling necessary data, there is the concern of data integrity. Some problems which may hinder this are format incompatibility, holes in readings, excessive data entry errors, and out of range data.

An important area of consideration is that of field formats and consistency. If it is necessary (or even remotely possible) to compare fields from the same or different files, make sure the fields are compatible. SAS is very forgiving by automatically converting between numeric and alpha data, but many languages are not. Even values within the same field should either have a uniform format, or have a method to determine the relevant formats. For example, DWRD has created a data dictionary to document the various units of measure associated with each reading. Thus, explaining why one utility's contaminant readings are 100 times higher than another. Another example would be to make sure to use consistent values for street name suffixes (i.e., avoid 'str', 'st', 'st.') to prevent showing three separate streets instead of one during a search, merge, etc.

Another obstacle to the efficient utilization of a data base is created through gaps in field values. For instance, scientific data from a year long experiment is not as valuable if there is a month of data not taken or entered. Or consider comparing sales when receipts from some salesmen are absent. It is necessary to make sure data are sampled at regular intervals and encompass all variables to be compared to achieve maximum value. This is assured by prior planning when developing new systems, and also by working closely with the personnel in charge of compiling the data.

A third problem to avoid is excessive data entry error. This can be caused by a variety of factors but is best avoided through close supervision of the entry personnel, and strict adherence to standard data entry protocols by the entry personnel themselves. If more than one person is responsible for entering data, make sure the work is divided and assigned well ahead of time to prevent reentry and omission of data. At various times, DWRD has college students perform data entry. On one occasion, two students were entering data on different days and did not record their daily progress. The end result was a duplicate year of data entered, with one year omitted, and a lot of extra effort.

Lastly, be aware of the problem of out-of-range data. Although a value may be appropriate for the size and type of field it occupies, be aware of the possible value range for the field. For example, although a value of '25' is a perfectly acceptable numeric value, it would not be correct for a pH reading. This can be combatted by the entry personnel having an idea of the possible values they will encounter.

Verification is a very important step in finding problems in the data. Usually DWRD will verify data immediately after entry by randomly checking every 10 or 20 values manually. Even so, incorrect data can still slip through. Another helpful verification technique is to use SAS Graph procedures to plot data sets using various BY values and sorted in various ways to show trends and correlations. By using PROC MEANS, summary statistics for large data sets can quickly show problems relating to out of range values and gaps in data. PROC FREQ is useful to find errors when variables are interrelated. Most parametric statistical tests (such as Analysis of Variance, t- tests, least squares regression) rely on the underlying populations and/or error terms, etc. being normally distributed. Plots of the data, error terms, etc., are done before a formal statistical test for the normality assumption. SAS can easily do plots, as well as provide formal statistical tests for normality, such as the Kolomogorov-Smirnoff under the PROC UNIVARIATE procedure. Many statistical tests also rely on the assumption of independence of the error terms. Again, plots of the error terms are done before a formal test of independence. Plots of residuals can usually reveal correlations among the error terms over time, among other undesirable patterns. Histograms are invaluable for checking assumptions of symmetry between two groups of data. Formal statistical tests to check assorted assumptions are normally done following a visual examination of the data.

Assuming the data has been entered correctly, now consider how it will ultimately be used. DWRD exports data to interested parties, and compatibility becomes a major concern. When files are sent out on tape, for instance, the tapes are set up as generically as possible to facilitate easy use by the recipient. A million-record research data base is of no use if it is only accessible by those users with an XYZ system. The sooner a data base is usable the better.

Just as important as managing the data itself is managing and interacting with the people related to the various stages of generating and using the data. By having an intermediary to work with the parties involved, the entire data base development and operation becomes much easier in the long term. The data base manager should meet with all levels of personnel, including those requesting such a data base, those generating the data (in the case of experimental data), and those producing the outputs and using the system. An important concept to keep in mind is that the more each individual knows about the data base, its uses, and structure, the more efficient and productive the entire system will become.

The data base manager must have concrete ideas for the data and system well before development begins. The manager should always consult with the person(s) requesting the system and find out what ultimate uses the system will have, and what it is expected to produce. By finding out the potential uses for the data and system, the manager can better tailor the final product to the individuals needs. This prevents the problem of variation between what the requestor wanted, and what the developer thought they would want.

The manager/developer should next consult with the person(s) responsible for generating the data. Here is where many of the possible format problems can be avoided, and by working with the technicians the collecting of superfluous data can be avoided. This is also an opportunity for the developer to gain insight into the real workings of the project, as opposed to only the project's ultimate goals. Here important information regarding problems, modifications, etc. can be obtained first hand.

Finally, during actual development of the data and system, the manager should strive to have the data processing staff as knowledgeable as possible about the data. Also, the data entry personnel that are used should be well informed regarding possible values, formats, etc. This alone prevents many data entry errors. By knowing the data well, the programmers and users of the system can better interact with management to process requests.

CONCLUSION

In summary, this paper has briefly tried to stress the importance of the data base manager to become more involved in the personnel management aspect as well as the data base aspect of creating an information system. There are specific problems which confront anyone involved in this endeavor relating to the data and system itself, as well as the personnel involved. By judicious use of various SAS procedures, problems with the IS data can be quickly discovered and rectified. Another way to attack the problems is to prevent them from happening. This is aided by having the persons involved in the project become familiar with the data. By clearing up misconceptions in the minds of management as well as technicians, professionals, and clerical staff, all involved are better able to utilize the data to maximum benefit.

ACKNOWLEDGEMENTS

The authors would like to thank John Ireland of DWRD and Diane Westendorf and Deanna Wild of Computer Sciences Corporation for their input and assistance.