MANAGEMENT OF EVOLVING ECOCLOGICAL DATA SETS WITH SAS: AN OPEN-ENDED MANAGEMENT APPROACH

Michael P. Farrell
A. Dale Magoun
Karen Daniels

Environmental Laboratory
United States Army Corps of Engineers
Waterways Experiment Station
Vicksburg, Mississippi 37180

ABSTRACT

Often in ecological pilot field studies, decisions as to data base management, structure, analyses and report generation can only be speculative since the data base requirements often have to be developed in this phase of the study. In these cases, the lack of appropriate information prior to the study cannot justify detailed and extensive pre-planning of research data base management strategies. The potential liability of open-minded management of data bases has to be minimized against the need to be able to modify content and structure of large data bases. For many complex and large scale ecological field studies being conducted at the Waterways Experiment Station SAS 78 under T50 provides the framework for making open-minded research data base management systems practical and cost effective.

INTRODUCTION

The General Problem

The foundation of sound research data base management (RDBM) is detailed planning, or so says the texts on the subject. One who manages such a data base is usually barraged by the apparent need for flow charts, pert diagrams, cross-reference libraries, directional dictionaries, and a host of other "aids" designed to increase efficiency and/or ensure a final product that will accomplish the goals of the project.

Plans on how to recover data sets, sorting strategies, merging and updating capabilities, and inter-computer exchanges must be obligated well in advance with few allowances for "lurking variables" or any of the other forgotten monsters associated with RDBM. As if the RDB manager was not faced with enough problems, he must also give detailed breakdowns on development time, personnel and computer costs, and the lead time necessary to develop the application programs even though the project may be several years away with the possibility of personnel turnover and/or hardware changes.

With the background described herein, the RDB manager thought he/she was prepared to tackle his/her job during the second and part of the third generation of computing—the golden age of

Present address: U. S. Army Corps of Engineers San Francisco District, San Francisco, California

PORTTRAN, COBOL, and the boards of application programmers. What is difficult to understand is the continuation of this RDBM strategy among RDB managers today—the current generation of computing.

Informal polls among RDB managers indicate that an increasing number of RDBM systems are not being developed as outlined in many of the leading references in this field. For example, flow charting, one of the backbones of the industry, has been shown to be an academic exercise with little application or help to real world complex RDB problems. Pert diagrams are very informative after the final report is written. Dictionaries, fixed sorting strategies, cumbersome merging capabilities, data base recovery problems, etc., are all dinosaurs, extinct but with varying degrees of belief by some RDB managers that they may still roam CPU's.

Most RDB managers find it difficult, if not impossible, to cost justify more than a cursory RDBM plan. Time estimates for development are usually too long and must be reduced. Furthermore, cost estimates may be very inaccurate when many research programs cannot identify all the variables nor data base formats that will ultimately be necessary for analysis and report generation.

Even though the previously mentioned problems exist in RDBM applications, there is a discernible naivety associated with many of the currently published reports on problems in RDBM. One of the principle reasons for this discrepancy may be the different approaches in RDBM found among applied contract researchers with strict budgetary constraints and university based research programs with their more liberal approach to computer related costs. Although free computer time at the universities is becoming very rare, there still exists such a price differential as to perhaps subliminally encourage many of the RDBM strategies popularized in reference material.

Another area of concern in RDBM is the vexing problem demonstrated at its worst by the apparent trend among RDB managers trained as programmers to view most projects as unique with unique solutions. The list of in-house developed data base management programs written in PORTTRAN
and/or COBOL specifically for a project must be
horrendous. Our experience could supply a
long list of these database management struc-
tures so specific as to preclude any general use
and often, only meeting a small proportion of the
needs of users. Cost overruns, programming,
problems, or changes in the project's emphasis.
On the other hand, this
tendency to "reinvent the wheel" does not seem as
popular among RDB managers trained in areas other
than programming, and aware of the new applica-
tion programs currently available on lease or
purchase options.

The Environmental Sciences and Research Data
Base Management

Unlike many of the other sciences, the en-
vironmental sciences have most of the problems
identified previously and project goals are often
defined more appropriately after the study be-
comes operational. Many times, at the startup of
a project, variable selection and RDB formats are
often tentative due to the unknown biological
complexity that may be encountered. Potential
ways to summarize the data base are usually more
numerous than programming, and development of even a simple RDB structure is
usually non-existent. RDB managers frequently
become involved with a project only shortly
before data collection with the subsequent need
for immediate data summarization so that the
project may be modified before the next scheduled
sampling period. In such an environment, where
the RDB manager faces a project that will provide
answers by an iterative process, and where there
will be major changes in the data base content
and structure, the manager cannot hope to spend
many days planning the specifics of the RDB and
only infrequently can the cost of development of
such a RDB program be justified.

What is expected of a RDB manager faced with
the uncertainties of managing an environmental
data base? He is expected somehow to provide a
project with a skeleton of a data management
system that can add broad ranges of new vari-
ables, reformat existing variables, perform
analyses that are not anticipated, provide com-
puter generated tables and copy-ready figures
that will be formatted at the conclusion of the
study, and, in general, provide immediate answers
to a time sharing system. But have the capa-
bilities of reducing the cost of large complex
analyses via batch operations. Superimposed on
the above list, the system development must not
demand excessively large and complex programming
tasks. The system must be cost effective with
costs ranging from S-5/LUX of the total project
funds—the lower the percentage the better.

Furthermore, the need for a system analyst to
manage and construct the data base is by project
definition "counterproductive." A person de-
voted to RDBM is often considered an "unneces-
sary burden" to the project and may jeopardize
the project's financial capability to measure
other "important variables" or eliminate some
other aspect of the project. If the project group
has a statistical-type it is usually that
person who is nominated for managing the data
base. Unfortunately, an ability in statistics
and experimental design does not insure skills
required for a RDB manager. In the final analy-
sis, RDBM and programming requirements usually
fall on the person in the project whose claim to
the position lies in the fact that he is the
only one in the group who realizes that computers
do not burn kerosene!

Now that we have outlined the problem we
feel are currently associated with RDBM on a
general scale, this paper presents an approach
we have used in a large, multidisciplinary and
complex study associated with the Environmental
Water Quality and Operational Studies (EWQOS).

The long-term field studies on the
Mississippi River, a part of EWQOS, are being
conducted in an attempt to ascertain detailed
information regarding the environmental effects
of selected types of structures that are used by
the Corps of Engineers to maintain the navigation
channel and flood control on the Nation's largest
waterway. The studies are being conducted by
personnel of the Environmental Laboratory at the
Waterways Experiment Station and are planned to
cover the six year period from FY 78 to FY 83.

An overview of the long term field studies can
be found in Vol. E-78-4, Nov 1978 of the infor-
mation exchange bulletin on EWQOS.

Because of the complexity of the Mississippi
River project and the fact that little informa-
tion was available prior to the extensive field
studies planned by the Corps, a six month pilot
study was initiated to help refine the general
plan of study and to provide base line data that
was needed to evaluate variable selection and
potential experimental designs. In relationship
to the RDBM aspects of the project, the situation
faced by the project management team was quite
familiar and reflected most of the difficulties
of managing research data bases associated with
changing and extremely complex environmental
data sets.

THE SYSTEM SELECTION

Because of the research status of our in-
stallation, the usual decision as to the choice
of appropriate software based on hardware avail-
ability did not have to be entertained. We have
the ability to utilize in-house computing capa-
bilities and/or purchase computer time from
commercial vendors holding G.S.A. contracts.
Therefore, the selection of an appropriate soft-
ware system was of primary importance and was
unencumbered by the usual hardware consider-
ations.

To accomplish the short-term objectives of the
pilot study and maintain the proper perspec-
tive of a long-term commitment to the overall
study, the program software system had to meet
the following five major system selection cri-
tera: (1) vendor support of the system's
software including programming applications, analysis programs and help trouble-shooting user applications; (2) provide RDBM capabilities that are easily programmed (user oriented), flexible and hierarchical in that canned instructions exist (e.g., sorting, merging, updating), but user programmed instructions are also supported (e.g., input, output, quality control checking); (3) provide a basic complement of statistical analysis routines (i.e., means, standard deviation, analysis of variance, regression, etc.) plotting and charting capabilities, and more advanced programs which may be available in the system, programmable within the system, or available in other packages that interface with the parent system; (4) provide a common syntax for batch and time shared operation; (5) cost effective, not only in terms of computer costs (e.g., core, CPU, I/O) but for the personnel time needed for implementation and maintenance.

After reviewing our needs, the selection of the appropriate software was initiated by the decision that an in-house development of application programs (i.e., data handling and analysis) was not a cost effective alternative for our project. Considering development time estimates, personnel commitments and costs, especially for the changing status of the pilot study, an in-house effort would immediately exceed one of our prime goals of keeping RDBM at 5-10% of the project's funded level.

In the author's opinion one system, the Statistical Analysis System running under MVS in a TSO environment, adequately addressed each of the potential problem areas associated with both the pilot study and the long term study and met favorably with the system selection criteria outlined previously. Furthermore, at least one staff member was very familiar with the system and how it could be efficiently applied. For these reasons SAS was selected for use in our study.

PLANNING VERSUS OPEN-ENDED MANAGEMENT

All RDBM structures must be planned. What is perhaps not clear is the amount and direction of planning necessary after an advanced analysis package has been selected. Detailed planning of RDBM's appears to be inversely proportional to the degree of which the selected package meets the system selection criteria outlined previously. If adherence to the criteria is high, then planning the RDBM can be minimal with sessions devoted to determining output formats and requirements, and any specialized analysis programs needed but not contained in the package. On the other hand, when a selected package has major omissions in relationship to the system selection criteria—when adherence is low—planning time is usually increased with the emphasis on more of the basic problems of RDBM such as variable input formats, internal file construction, sorting, merging, updating, etc. Therefore, adherence to the system selection criteria permits the RDB manager to be more involved with the end-product requirements of the study such as copy-ready graphical displays, computer generated tables, and quality control assurances.

In turn, the scientist involved with reporting the findings of the study benefits from this new end-product orientation of the RDB manager. Now the scientists can become more involved with interpreting results of the study, as opposed to editorial requirements mandated without an efficacious system. Furthermore, decisions that were previously based on inflexible computer program requirements can be modified so that the emphasis is placed on the needs of the scientists. As a result, efficiency is gained in the field operations where the majority of cost are usually involved without additional cost to the RDBM program.

We have coined the term "open-ended research data base management (OE/RDBM)" to describe this philosophy of RDBM that supports a minimal planning effort and in which the emphasis is placed on computer related needs at the completion of the study. Obviously, an OE/RDBM strategy will not work for all types of programs, nor for all levels of experience in executing the selected package. In addition, OE/RDBM may be precluded as a working system due to the emphasis of the program and/or the degree to which the selected package meets the system selection criteria. We felt, however, that SAS is very amenable to OE/RDBM and that it reduces many of the potential hazards that could be encountered with an open-ended management system. We found that the key element in using SAS for OE/RDBM is a sound understanding of the structure of a SAS data set and the data set formats needed by a procedure to produce the desired results. Knowledge of the capabilities of SAS and of the lack of certain capabilities is also instrumental in avoiding the old adage that "there's no way to get there from here..." after a RDBM has been implemented.

THE COST FACTORS

While OE/RDBM using SAS is an appealing alternative to RDBM based on philosophical grounds, the cost differential of a OE/RDBM system may be even more attractive. We recognize the difficulty of assigning dollars to inflation factors, time estimates, computer equivalents, etc., but even a simple, basic cost analysis of RDBM alternatives provides some striking arguments for alternative approaches in RDBM.

Table 1 shows the comparative cost analysis of a system using the OE/RDBM approach and SAS with a system developed within an organization at the program language level (FORTRAN). Both systems have been evaluated using commercial and private computer cost estimates. The manpower cost estimates are based on a skill level of a GS-12 rating including overhead burdens. All values are estimated on the basis of our experience and reflect a detailed knowledge of the specific project's requirements which are not presented in this paper. There are many other costs that could be included in the analysis (e.g., maintenance, office costs, leased on-line devices, terminal leases, communication costs, programming support) but have been omitted because of the variable nature of these costs.
and in an effort to maintain simplicity.

From Table 1 it appears that using OE/RDBM approach and SAS can result in substantial savings to the project on the order of 50-60%, irrespective of that computer affiliation and computer time, with a minimal dollar outlay for a OE/RDBM system on a personal computer. As might be expected, analysis costs are approximately the same for both systems. The major difference between the two systems is the additional cost of manpower and computer time in the development of an in-house system—approximately 85% of the total RDBM costs of an in-house system and only about 60% of the cost of an OE/RDBM system utilizing the capabilities of SAS. It should also be pointed out that the two systems would not be equal in terms of data base management or analysis capabilities. In our example, the OE/RDBM system using SAS would provide about 50% more capabilities in both management and analysis.

While these figures are tentative and subjective to our own biases and experiences, there appears to be a potential financial savings that is substantial enough to warrant detailed cost analysis at the local project level. We feel that the OE/RDBM approach in conjunction with SAS will be difficult to ignore in programs that have to be cost-effective.

**THE PRACTICAL CASE—ELPROG**

The Environmental Water Quality Operational Studies loading program (ELPROG) was developed as an OE/RDBM system in SAS and consists of approximately 2500 lines of SAS procedure instructions and SAS programming statements. The impetus for the development of ELPROG rests in the biological, physicochemical and RDBM complexities involved with the pilot study phase of the long term field studies. For instance, the four major areas of the project that deal with the various aspects of the ecology of the Mississippi River have a total of over 500 variables that potentially could have been measured. Therefore, because of the uncertainties involved in variable selection, sampling methodologies, analytical techniques, and the minimal effort that was allocated for RDBM, it was felt that no other approach seemed reasonable with as high a probability of success as using OE/RDBM approach.

The initial version of ELPROG was developed with data input being through a common TSO file containing the card image data from all field measurements and laboratory analyses. ELPROG 1 inputs the TSO file supplying variable names and labels according to the type of data being processed (e.g., fish, benthos) and outputs these observations to a particular SAS data set depending upon the type of data identified previously. This flow of data is shown in Fig. 1. ELPROG 1 also made quality control decisions; when a variable for a given observation failed a quality control check, a flag variable associated with a particular variable was turned on for visual inspection of the printed SAS data set. The temporary, but saved SAS data set was edited using PROC EDITOR and/or UPDATE and usually, after several iterations, an error free SAS data set was stored for processing by specific SAS analysis programs.

While the details of the loading program ELPROG 1 are beyond the scope in this paper, several programming decisions were made prior to implementation of the system and are of interest. Briefly, ELPROG 1 used a conditional input routine that keyed on a specific column for a unique character code to identify the input string and the appropriate data base for the observation. No alphanumeric codes were carried into the analysis programs. For example, all fish species codes were processed through a taxonomy macro that supplied the appropriate genus, species, official common name, etc. for the observation. Likewise, location codes, gear identification, time, dates, etc. were all translated to English equivalents.

During the first few months of the pilot study, variable lists changed on all input data sheets. As expected, however, the study followed a path that reflected the iterative nature of problem identification and solution. ELPROG 1 handled the many changes in the data base easily. Editing the temporary SAS data set for errors using PROC EDITOR and UPDATE was easily accomplished but the amount of editing necessary to correct the data sets indicated that stricter controls were needed on data sheets prepared by the field crew.

Two major design problems were experienced during the longevity of ELPROG 1. The first problem was based on the assumption that the card image, TSO file containing laboratory and field data would never have to be edited. In planning the OE/RDBM structure, it was originally thought that the saved SAS data set created after editing would be the only data set necessary for analysis and report generation. This assumption proved to be erroneous. The kind and complexity of logic errors experienced in the raw data file were not anticipated. The TSO file had to be edited, or else a SAS program developed to handle all the encountered logic errors in the laboratory and field data sheets—a formidable task that would have conservatively taken 60% of our RDBM budget. With few other options available, editing under TSO/MVS began. After the first major editing session, another problem emerged—the cost of editing under TSO. Confounded and added to the cost of TSO editing was the fact that the new data files needing editing were random in relation to the position of an observation with erroneous information. Under these conditions, costs were astronomical in terms of manpower and computer dollars, and impossible to justify when data summarizations and analyses might have to be reduced.

The second major design problem was the dropping of alphanumeric codes associated with many descriptive variables before the final SAS data set was saved. The overhead of carrying long, alphanumeric variables along with each observation and the additional cost of sorting
or merging on these expanded variables unnecessarily inflated data base management costs.

It should be pointed out that the problems associated with ELPROG were design problems—data flow problems—and not an inability to meet new database requirements by reprogramming the system. We also feel that the OE/RDBM approach used in this project, with its limited planning activities, did not add to the probability of encountering the problems described above. On the contrary, the lack of any significant alteration of the data base structure and the ease of which the next version of ELPROG was implemented encourages the use of OE/RDBM using SAS.

The current version of the ELPROG system, ELPROG_2, has hopefully eliminated the problems experienced during the genesis of the initial system (Fig. 2). One of the major changes incorporated into ELPROG_2 was the shift in the manner of handling errors in the data base. Major logic errors are now changed in the raw data file (TSO file) and a new SAS data set is compiled. For minor infractions, both the raw data file (TSO file) and the temporary SAS data set are modified, and the data flow continues to the storage of the final SAS data set. The other major change in ELPROG_2 is the substitution of expanded alphanumeric variables immediately before analysis or report generation routines. In this way, sorting and merging on numeric variables and/or short alphanumeric variables results in better timings, disk pack utilization, and cost. Storage requirements are also reduced because the overhead cost of carrying redundant descriptive variables with each observation has been eliminated.

SUMMARY

The need for new methods in research data base management is acute. Historically, RDBM has been based on planning and in-house development; however, this approach has become partially antiquated. New approaches in RDBM are needed to cure problems, but to be effective, these cures must be applied during the development of the data management system. The approach described in this paper, open-ended research data base management, is intimately related with the capabilities of SAS. Under these conditions, OE/RDBM using SAS appears to be a practical and cost effective alternative for managing large, complex research data bases. The practical experience gained from the implementation of ELPROG supports the usefulness of OE/RDBM and demonstrated the feasibility of the method for RDBM.

REFERENCES

A litany of authors, reports, and textbooks could be provided to support or negate most of what we have stated in this paper. We believe it unfair to ascribe to a particular author our misrepresentation of his/her approach to research data base management. Suffice it to say that we obviously did not invent the approach described in this paper. Bits and pieces from many authors in many disciplines aided in the development of the concept of open-ended research data base management. To all those authors from whom we "borrowed" we want to extend our thanks for doing much of the legwork and providing the stimulation needed to actually try doing something with practical overtures.

ACKNOWLEDGMENTS

We would like to thank the personnel of the Environmental Monitoring Group, Environmental Laboratory, Waterways Experiment Station for their patience during the genesis of this data base management program. Without their efforts the ability of the ELPROG system to cope with difficult RDBM tasks would never have been tested. Support for the senior author during the development of the management system was made possible by an IPA appointment with the U. S. Army Corps of Engineers from Miami University, Oxford, Ohio.

DISCLAIMER

The contents of this report are not to be used for advertising, publication or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such commercial products.
Table 1. Cost analysis of an inhouse research data base management system compared to the cost of an open-ended research data base management system (OE/RDBM).

A. Inhouse development in FORTRAN

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>inhouse</td>
</tr>
<tr>
<td>Manpower</td>
<td>2</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>275</td>
<td>275</td>
</tr>
<tr>
<td>Developmen</td>
<td>3</td>
<td>15/30</td>
<td>15/30</td>
<td>15/30</td>
<td>10/20</td>
<td>70</td>
<td>150</td>
</tr>
<tr>
<td>Analysis</td>
<td>4</td>
<td>8/12</td>
<td>10/13</td>
<td>10/13</td>
<td>20/24</td>
<td>58</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>403</td>
</tr>
</tbody>
</table>

B. OE/RDBM with SAS

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>inhouse</td>
</tr>
<tr>
<td>Manpower</td>
<td>4</td>
<td>14</td>
<td>18</td>
<td>28</td>
<td>18</td>
<td>37</td>
<td>105</td>
</tr>
<tr>
<td>Developmen</td>
<td>3</td>
<td>5/10</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>9</td>
</tr>
<tr>
<td>Analysis</td>
<td>4</td>
<td>8/15</td>
<td>10/15</td>
<td>10/15</td>
<td>20/30</td>
<td>58</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>172</td>
</tr>
</tbody>
</table>

1 Cost are estimated in $1000.

2 Manpower cost based on a GS-12 skill level for a 12 month period.

3 Development cost for inhouse computer/commercial computers.

4 Manpower cost base on a GS-12 skill level for 3, 4, 4, and 8 months respectively for years 1-5.
Fig. 1. Data flow through ELPROC 1 for a particular sampling period. Only two of the major work groups are shown (i.e., BENTHOS = benthic invertebrates and FISH = adult and larval fish). DS - data set. PGM - program. OE/RDBM - "open-ended research data base management".
Fig. 2. Data flow through ELPROC 2 for a particular sampling period. Only two work groups are shown (i.e., BENTHOS = benthic invertebrates and FISH = adult and larval fish). DS = data set. PGM = program. MACROS = SAS program macro's. OE/RDBM = "open-ended research data base management."