Setting Up an Efficient Data Management System

MARY ANNE TREADWAY
U.S. Department of Commerce - NOAA
Southeast Fisheries Center
75 Virginia Beach Drive
Miami, FL 33149

INTRODUCTION

One of the problems of any research effort is the ability to have enough data at one's fingertips to perform a proper analysis. Since the introduction of computers, this problem has been redefined; the research community often has plenty of data, but the ability to find, understand, and relate it to project objectives as well as use it to the best advantage must be resolved. Successful data management systems have a number of elements in common which will be briefly discussed here.

PLANNING

The first step in building a usable system is designing a data management plan, including establishment of a functional organization capable of implementing this plan. The absence of such a plan is the major cause of failures in most data processing organizations. Guiding the planning process should be the policies and procedures of the parent organization; data processing is an integral part of most organizations, but in many cases this is not reflected in corporate policies. Depending upon budget, size of installation, and data processing needs, the system could be one person or a staff of more than 100. It is important to determine the scope of the work and then select a staff accordingly.

To centralize or not to centralize - that is the question! Following are some key points to consider relating to this highly debated topic:

Centralized Data Management

One system, located at a staff level, providing data processing support to one or more units within the organization. Advantages are:

1. Prevents a situation wherein one research effort hampers important projects for another.
2. Permits the design and use of common data bases and common standards for data entry and input validation.
3. Ensures future ability to maintain all systems and programs by using common standards for documentation.
4. Minimizes inefficient use of computer resources by using proper programming techniques.

5. Ensures implementation of mission goals by evaluating projects from an overall organizational perspective.
6. Avoids redundant development of similar systems for different division within the organization.
7. Offers opportunity for promotion of personnel, thereby attracting and retaining highly qualified staff by using a centralized unit with layering.

Decentralized Data Management

Data processing staff located within individual research areas. Advantages are:

1. Data processing staff located within each research unit are more attuned to scientists' needs, allowing them to create systems more suited to the particular research effort.
2. The staff has the ability to respond more quickly to emergencies and changes in priorities.
3. Closer association between the data processor and the scientist facilitates better education of the user on benefits and limitations of computer systems.

Conclusion

Centralization is usually preferable since it provides the greater benefits when considered in the overall mission goal; however, there is no easy answer or single "best" way to organize the data processing function. With the proper planning document outlining the goals of data processing, each organization can develop its own "best" solution, which often is some compromise between centralized and decentralized data management.

OPERATIONS: MAKING IT WORK

Developing computer systems that satisfy user needs, are delivered on schedule, and are within budget remains the most important task for the data processing manager. In many organizations, computer applications are the largest single consumer of funds. But, selecting applications and setting priorities can only be accomplished when upper management provides the focus and direction through careful planning and dedicated implementation.

Planning

Planning is best accomplished through an in-depth analysis of the desired end product: What does the user really need in order to accomplish his project? The plan, developed by the user, is a detailed document outlining the objectives, approach, and timeframe of the activity. Once documented, the information must then be presented to the data processing staff to determine:

1. Is there a system presently available that could provide some or all of the data?
2. Are there plans by other users to collect and automate data of this type?
3. Are the objectives, approach, and timeframe reasonable, given the hardware, software, and personnel resources available?

The answers to these questions will determine the next steps, *e.g.*, enhancement of current systems, creation of entirely new system, redesign of a data collection effort developed within a different unit, rethinking the original plan, etc. Without the planning document and good communication between the user and the data processing staff, systems tend to be developed in an ad hoc fashion and could result in duplication of data, collection efforts, and computer programming time and resources. The end result is a system which takes much longer to develop and implement, and may not provide the information needed.

Implementation

Once the planning document is completed and the data processing staff is "on board", execution of the plan must begin. The user's role is the most important here: he must convey to the computer personnel exactly what is needed, make changes wherever necessary, and learn how the system will function.

In the implementation phase, the computer programming staff begins the software development. There are generally six separate stages of software development, and it is imperative that complete documentation be developed and maintained for each stage:

Requirement Analysis

The planning document will serve as the basis for this analysis, which will focus on the interface between the computer system and the user. It can also serve as an aid in understanding both the problem and the trade-offs between conflicting constraints (essential requirements versus optional features, time, or space limitations; for instance: it is essential to carry species codes but may be optional to carry more than one type of species code; restrictions on implementation time may require certain reports be delayed; computer space requirements require the system to be implemented on a large machine rather than on a personal computer; physical location of data collectors and users require computer telecommunications to be available; etc.).

Specifications

This stage defines precisely what the computer is to do. What are the input and output? Are records kept on disk or on tape? How is the output to be formatted? What files are needed? How will they be accessed, updated, and archived?

Design

During the design stage, algorithms required by the specifications are developed and the overall structure of the computer system begins to take shape. The system must be divided into small modules, each with its own constraints—function, size and speed.

Coding

The coding stage is where the computer programmer actually writes the instructions in computer language (FORTRAN, COBOL, BASIC, etc.), that will be carried out each time the program is executed. Coding has been mastered better than any other stage of software development. Unfortunately, many people think this stage is the only stage in implementation of computer systems, which has led to misunderstandings and frustrations on the part of many users.

Testing

The testing stage may require as much as 50% of the total development effort. Inadequately planned testing often results in either extremely late deliveries or systems that do not function properly. Testing must be accomplished at the module level (subroutines within a computer program), the individual computer program level, and the system level (all programs together: edit, update, reports), making sure that every statement is executed at least once by test data, every path through the program is executed at least once by the test data, and for each specification of the program, the test data must demonstrate that the program performs the particular specification correctly. The user must be very involved in the test stage, designing the test data and "signing-off" after each successful test is accomplished.

Operation and Maintenance

In a large scale computer system such as Fisheries Statistics, two-thirds of the man-hours involved in the life of that system are devoted to operations and maintenance. No computer system is without change. Because users rarely know exactly what they really need, or because certain analyses point out needs for additional analysis, users often request changes to the system. Errors that are missed in testing will often be discovered later. Requests for rush development and implementation usually trade early execution for a much more extensive maintenance operation.

Conclusion

The implementation of a computer system is a complex and time consuming task. Without the proper planning, communication, cooperation, and resources it cannot be accomplished effectively or efficiently.

GETTING THE MOST FROM PERSONAL COMPUTERS

We have briefly looked at planning, organization, and making the system work, but I have not yet discussed how to cope with the demands of the personal computer. The arrival of personal computers has made users more aware of the need to understand all computer systems that affect them. Users of personal computers frequently contact data processing personnel to request assistance; time which was devoted in the past to designing and implementing large-scale computer systems is now given to assisting users in developing "computer literacy."

The data processing organization can be very beneficial in translating "computerese." Bits, bytes, kilobytes, ROMs, RAMs, operating systems, 80286 processors, tracks, sectors, graphics boards, serial vs. parallel, adaptors, etc. are all terms in a seemingly foreign language but necessary vocabulary for new users.

Personal computer users are primarily interested in "off-the-shelf" programs that provide spreadsheets, data management, charting, graphing, word processing and communications. Users must eventually learn to perform such tasks as making backup disks, revising file names, deleting files, etc. Some type of training sessions must be made available to assist in bridging this information gap, but this objective can be accomplished through a variety of techniques: in-house seminars, formal training sessions by outside consultants, or purchase of video tapes. This area of training is a very important one and should not be overlooked.

Once the personal computer is in place and the user is trained, management's task is to control the operating environment. The most significant areas of control are:

Security

Data security controls for personal computers using either mainframe or microcomputers are not very effective. Both data and computers are moving into offices; with the distribution of computing resources to the end user, control responsibilities are also distributed.

Data Integrity

The reliability and availability of data depend primarily on the guarantee that data will be properly inputted and maintained. Data can be uploaded or downloaded between the personal computing system and the mainframe. Mainframe control over input, editing, accuracy, and completeness is usually not duplicated at the personal computer level and the possibility exists that some new data input via the personal computer could invalidate the organization's data bases.

Efficiency

Efficiency of a system can be derailed by many factors: attempts by unskilled or unqualified persons to program complex systems; "re-inventing the wheel" by duplicating programs and procedures already accomplished elsewhere in the organization; and buying personal computers without first determining user requirements (the "buy now, decide how to use later" syndrome). These are just a few of the factors that can damage the efficiency of a system.

Backup and Recovery

Backup refers to copying files from one disk to another so that in the event of one disk failing, chances are good that the file(s) are not lost. Files for backup purposes are not limited to just data, programs need to be backed up also. Most users seem to learn the importance of backup only after experiencing some catastrophic failure. A plan needs to identify critical personal computing functions and coordinate a system for backup of these functions.

System Compatibility

Increased sharing of data and computer programs dictates that personal computer systems must be truly compatible. Most manufacturers do not build devices that are readily compatible with one another and therefore different devices cannot readily exchange data or programs. Failure to plan for future integration needs will limit access and exchange of needed data.

Senior management needs to be involved at every step to ensure effective, controlled use of personal computing technology. Sound procedures and controls should promote, rather than discourage, its acceptance.

CONCLUSION

An efficient data management system benefits everyone in the organization, not just the end user. Coordination and cooperation between upper management, the end users, and the data processing organization are the keys to a successful system. When these entities work together to scope the project, assist each other, resolve problems, and share information, the result can only be an efficient and effective data management system.