INTRODUCTION

The acquisition, storage and distribution of essential information is a common problem in data processing. In order for the distribution of the data to be completed in a timely manner, the data must first be obtained, validated, edited, and integrated into a database as quickly as possible. A computerized information system with on-line access to a database, distributed processing, and an efficient database manager provides the framework in which to solve this problem. The system's design is crucial to its long range success, and a combination of the right technologies is needed to meet these requirements. Their use must be well thought out so that extensive data conversions are not required to go from one processing phase to the next. Future advances in technology must be considered so that hardware upgrades, new versions of operating systems, and future enhancements to the programming languages can be taken advantage of.

This paper will describe how an information system was designed to take full advantage of multiple software and hardware technologies in order to satisfy the acquisition, storage, and distribution requirements of a large information processing system. The SAS system became the pivotal tool in bringing multiple technologies together into one integrated information processing network. The system upon which this work is based integrated lasering technology and an extensive SAS-based reporting system with the Model 204 database management system. SAS code was also used as the interface to format laser commands, generate Model 204 commands for loading the database, and to perform quality control during data capture.

This information processing system was designed and developed by ORI, Inc. for the United States Department of Education. The Integrated Postsecondary Education Data System - IPEDS - processes up to ten surveys annually from over 14,000 institutions offering formal instruction beyond high school. The survey forms are electronically designed and printed with customized data on a Xerox 9700 laser printer, ready for assembly by a mailhouse. The response data, received on a variety of media, are stored in the Model 204 database management system to meet requirements for maintaining up-to-the-minute, fully-synchronized files that can be edited and reviewed by the survey managers. All detail and summary presentations of the data as well as end-user access are designed around a unique implementation of SAS/AF and SAS/FSP with the SAS Macro facility. (A paper presented at this conference "Making SAS Version 5 Work for You: The Design of a Dynamic Reporting System for all Users" describes this component in detail.)

The following topics will be addressed in the sections that follow:

- Design considerations: Acquisition, Storage, and Distribution
- Integrating the SAS system with Model 204
- Factors that lead to success

DESIGN CONSIDERATIONS

The acquisition, storage, and distribution of data must first be considered from the user's point of view. A system approach can then be taken, which separates these three areas into their external and internal "views". Finally, the software and hardware must be selected that most appropriately meets these requirements, and unites it into a complete information system.

Figure 1 presents the logical flow of data in an information system, as it might be viewed by an end user of the data. The data can be acquired through a variety of sources including interviews, survey forms, and electronic media such as diskette or magnetic tape. The data must be stored "in the computer" so that they can be reviewed, edited, and accessed electronically through video display terminals. The data are distributed as printed reports, and can also be made available on diskettes and magnetic tapes. Ad hoc analysis is usually required for users that are accustomed to working with microcomputers or high level programming languages such as SAS. This flow of data appears to the user as one system that always has up-to-date information.

Figure 2 presents a physical flow of the data as it would be envisioned by the system designers. It organizes the logical flow of data in the information system into the major system components. Again, we see the three essential aspects of information processing: acquisition (data capture), storage (data editing) and distribution (data reporting). All three components must add data to the database, as well as read and update existing information.

Figure 3 presents a design that breaks down the system components and identifies the programming languages and file interfaces that will be needed in order to implement the system. User access to the system is now presented through an external interface (often called a user-friendly menu). The system functions are presented to the user as the original user "views". The system design utilizes the DBMS for data storage and on-line editing, but relies on the SAS system for its strengths not only in analyzing the final results, but as the glue for tying the system together in reading external data files and reformating the data for storage in the database.
From the User's View

![Diagram showing the process from data capture to reports creation.]

FIGURE 1

From the System Designer's View

![Diagram showing the process from data capture to report publication.]

FIGURE 2
Acquisition of Data

Surveys, interviews and data tapes from state or federal agencies are often used to acquire data. The data to be added to the database must first be converted into electronic images (punch cards, tape, diskette) so that they can be accessed through a computer for storage, editing, and reporting.

In the system designed for the Department of Education, survey forms were created as electronic images using the Xerox Forms Descriptor Language (FDL), and stored on a Xerox 9700 disk storage unit. Electronic forms creation provides a method for turning the preprinted form into an image that can be distributed and stored electronically.

With a total electronic system, form and data are merged so that the computer can be used to fill in various parts of the form. Since the data are merged with the form electronically, alignment is always exact. The computer system utilizes SAS Macros to compose the appropriate Xerox Laser printer commands for each form, and overlays the form images with information from the Model 204 database. Experience has demonstrated that this technique for customizing survey forms greatly reduces the respondents’ efforts, and insures that the forms will be returned quickly, since the respondent need only update a limited amount of data.

After the surveys are completed, they are mailed back to the Department of Education on either the original form, diskette, or magnetic tape. Each medium requires one customized quality control/pre-processing step to ensure that the required data elements (key fields defined to the DBMS) are present. Highly generalized SAS programs that make full use of the Macro language and Proc Format techniques are then used to transpose the data to the Model 204 input format. In this way only one update program is required for adding large volumes of data to the database.

Storage of Data

The data are loaded onto the database as the final stage of data acquisition. In designing an information system, however, the storage of the data should not be limited to how the data are formatted by a DBMS. If data are changed during updates or imputations, can these events be stored with the data? If there is a power failure, or a system "crash" can the current data be recovered, as well as the day's editing events? What means exist for backup? Can data be edited as the database is loaded? Data editing must include the determination of missing values and outliers, tolerance checks and numerical evaluations.
In the system presented as a model for this paper, database management systems were evaluated with regard to the items listed above. An additional requirement was that the DBMS interface well with the SAS system, since the end-users were familiar with SAS and had already targeted the reporting system for SAS implementation. The decision to use Model 204 was based on the following features:

- Access time
- Relational nature of the IPEDS data matched the Model 204 structure
- Ability to store multiple occurrences of values
- Immediate availability of Model 204
- Interactive editing with lockout at the record and field level
- Acceptance of all values (whether character or numeric)

The data editing component of a large information system should be activated every time a value is changed on the database. This ensures that the database is always synchronized with all system events. Each editing error should be noted with a "flag", and summary error reports should be generated after each new batch of data is added to the database. Status fields can note such events as mailout, receipt, presence of errors, and follow-up. Audits of updated data should be maintained so that prior values of fields can always be retrieved.

All data in the Department of Education System are edited during the loading of the database as an "off-line" batch process. However, error resolution takes place on-line at video display terminals. Using an interactive system of menus and "help" screens that are programmed into the Model 204 DBMS, editors can systematically resolve some errors and identify the need for follow-up for apparent outliers and erroneous responses. Errors that are flagged can be overridden if the data are verified as correct. Data that satisfy tolerance checks are accepted and data out of tolerance are flagged. Again, all data and all flags can be overridden. Error resolution continues until a "clean" database is obtained.

Distribution of Data

The goal of the distribution phase is to disseminate results as soon as possible after the close of data collection. The distribution of data includes publishing final reports and database files as well as extracting database components for end-user analysis. Interim status reports on data collection are also crucial products as the process of gathering information progresses. All system output - distribution tapes, reports, analysis tables - should be integrated into the computer system.

The SAS system was selected for the entire reporting component of IPEDS for a multitude of reasons:

- End users were familiar with it
- Historical data already existed in SAS data sets
- A sophisticated, yet generalized reporting system could be designed so that it could be maintained by the Department of Education staff.

The resulting reporting system for the Department of Education was designed for a wide variety of "users". Extracts of the database are taken periodically so that the reporting system can provide the IPEDS user with hand-on access to analysis reports, status reports, and preliminary data results. The reporting system makes full use of the latest features of the SAS system including SAS/AF and SAS/FSP. Reports that allow survey managers to monitor the status of the system, as well as both summary and detail reports on the data are incorporated into a large report library. A set of standard reports can be easily accessed by all end users. The report formats can be modified through a customizing menu option that allows for a variety of aggregations, subsetting of data based on certain parameters, and variable name selection. As end users identify useful reporting tools, the programs that are written to create new reports can be made a part of the computer system by inclusion into this report library.

INTEGRATING THE SAS SYSTEM WITH MODEL 204

The IPEDS processing system integrates the three independent software products-SAS software, Model 204, and Xerox FDL with the Xerox 9700 Laser printer and the IBM 3083 under MVS. At the core of the system described in this paper, is the Model 204 DBMS which must manage a database in excess of 25,000 variables. The User Language of Model 204 is the primary software for working directly with the data. It is an easy language for skilled programmers to learn. It is designed for acquiring data from the database by selecting key variables, manipulating the data, and subsetting or joining data to existing lists of data. On-line access to the data is accomplished through use of the Model 204 User language to generate screens, provide for help menus, and display the data. A special scrolling system is incorporated into the system so that the screens can be rolled forward and backward, right and left. This technique, which is automated in products such as SAS/FSP has been programmed in Model 204 to function like a spreadsheet.
In designing a system with more than one end-user component, it is important that:

- All user presentations "look" as though they are created from identical software.
- Terminal (PF) keys have identical functionality.
- Scrolling works the same way - screens are "rolled" rather than completely replaced.
- Help screens and menus are always available.

The IPEDS system modeled its Model 204 screens after the same, automatic features of SAS/FSEDIT.

The two major areas where SAS programs were used in conjunction with Model 204 were in the formatting of the data to be loaded onto the database, and for generating routine Model 204 procedures. The SAS System became the "glue" that turned the components of information processing into an integrated system.

The loading of data into the Model 204 database is a process that begins with a SAS program, and ends with a file of data fully formatted for a database update.

The "raw" data originate in a hierarchical structure. Whether they are submitted on paper and require keypunching, are submitted on tape or diskette, the "lead" record determines what data follow. Format tables are then used to translate the data, by column location, and create the variable names used on the Model 204 database.

"Named output" (VNAME=xxxx) is created - a format that the database can accept. Control characters, and special spacing requirements (including the location of "blanks" in the data stream) are added to the output file. The system commands that are needed to identify the files to be updated are also composed in the SAS output data step.

An Information system usually contains a series of "rules" for editing the data on the database. When translated into modular software programs or procedures, the rules can be applied to many variables requiring the same "edit" (e.g. - a tolerance test).

In the Model 204 system designed for IPEDS, these procedures were identical for hundreds of variables, but required as many unique procedures, since each variable name (field name) was different and resulted in different editing messages. A SAS program was designed to generate the procedure code for each edit. The same formats designed for loading the data were modified, so that they could be used to write the customized procedures for each variable in a series of data.

Creating a SAS dataset from a Model 204 file can also be achieved easily. The only requirement is that the data are added to an existing SAS dataset. The dataset can be empty (i.e. have no observations), but the variable names must already be defined prior to using the technique described below.

A Model 204 record does not have to contain all of the variables that are defined for a given file. If data values were not present, the variables do not exist. An extracted Model 204 can be easily generated in a named output format: field name = value. While the extract creates output for only those values that are present, the SAS system can "read" the output as named input and create a rectangular file (i.e. a SAS dataset). Furthermore, the SAS input statement need only list one variable named, but all name variables on the record will be read.

The versatility of base SAS software in reading a wide variety of data structures is extremely valuable when working with a database management system. A DBMS often does not have the flexibility in its programming language to handle complex, hierarchical data structures. The SAS System "glues" all of the pieces together.

FACTORS THAT LEAD TO SUCCESS

Requirements of an automated data collection system can be separated into two major categories: processing the data, and long term viability of the software.

Successfully processing the data requires the right design of the system. Designing an information system requires vision, for the automation of a data acquisition system is a major undertaking. Many years of planning will go into a project like this even before a computer system can be designed. A state-of-the-art system that attempts to solve all of the problems inherent in the current, manual methods is always the goal of the new project.

Integrating several technologies while automated processes are added to the data collection efforts, presents numerous system design challenges as well as implementation hazards. It compounds the complexities inherent in the transition from manual to automated processes. Prototype programs and interim database development must be carried out during the system specification stages to guarantee a final, viable design. Due to the ease in which SAS programs can be used to prototype systems, the SAS system is an important development tool during the design of the system.

Besides the magnitude of the processing of the data, one must consider the problems inherent in maintaining the software. Will the system have long term viability?
The data processing system implemented for the Department of Education addresses the issue of long-term viability with a modular approach to the three critical processing areas of acquisition, storage, and distribution. The system design was established along a modular approach to accommodate the variety of hardware and software that would meet the Department of Education's immediate needs and long range goals for IPEDS. The critical software for processing the surveys is independent of the way the Database Management System stores the data. The interface to the DBMS is a single path. If the DBMS were to change at a later time, the single path component would be the only system structure requiring a major overhaul. The SAS programs that handle the transition from "raw" data to formatted data, and vice versa, can always be changed to work with a new DBMS. However, such a change would never impact either the originators of the data or the final published results.

This paper has described the design and development of a survey processing system that integrated the SAS and Model 204 Systems with Xerox 9700 laser printing technology. The system became operational in September, 1986, but it will require several survey cycles before the end-users' needs are fully met and the anomalies of the data are straightened out. An ever-changing universe of institutions requires a database that can easily manage a large volume of data in a near real-time environment. Model 204 is such a DBMS. The processing needs, however, are best accommodated by modular software that is data-driven, so that software maintenance is reduced to table revisions. The SAS System, as demonstrated in this paper, is ideal for this. Linkage between the components - data acquisition, editing, extraction and reporting - is handled through SAS Macro's that keep the processing extremely generic. A single interface to the database isolates the processing software from the structure and complexities of the data.

Integrating new technologies - where do you begin? Hardware should not be selected simply because it is compatible with the software development language. A database should not be selected only because it has an existing interface to other programming languages. Rather, the system designers should select each component on its merits (e.g., laser printer quality, data access speed, data storage capacity, software maintainability) and address the coupling of the components as a separate issue. The processing of information in a multi-stage operation can then be transformed into an information system. In this way, changes in technology can be handled with a minimal impact on the system as a whole, and each part of the system can grow with the needs of the data as well as with the users of the data.

*SAS, SAS/AF and SAS/FSEDIT are registered trademarks of SAS Institute Inc., Cary, NC, USA.

The author may be contacted at:

ORI, Inc.
601 Indiana Avenue, N.W.
Suite 1000
Washington, D.C. 20004