# Management of SAS Based Data Analysis

Wm. Ford Calhoun, James Holt
Mount Sinai School of Medicine, City University of New York

ABSTRACT

A small group of data managers and programmers in the Research Computing Unit of the Biomathematical Sciences Department of the Mount Sinai School of Medicine of the City University of New York(CUNY) supports 5 statisticians on approximately 100 projects per year. The projects range in size from large surveys of thousands of observations to small studies of a few patients containing 7 or 8 variables. The Unit is able to maintain a high level of productivity by tightly managing and controlling its data analysis function.

Components of the management system include computerized logs for: projects, programs and programmer activities; strict naming conventions for programs and files; standardized programming style, report generating programs and SAS related data management utilities. Major advantages of the system are accountability and low overhead, automation of many data management tasks and ease with which programmers can be interchanged on projects. The components of this system, their interrelationships and system advantages will be discussed.

## 1. INTRODUCTION

SAS is the best general purpose data management and analysis system available. All other things being equal, a manager of biomedical research data will prove more efficient using SAS than any other system. However, there is more to running an efficient data management shop than putting SAS into the hands of intelligent data managers. The long term quality and through-put of a data analysis operation is dependent on the organization and operating procedures of the shop. In our experience, management of personnel and resources usually does not receive the attention it deserves.

Over the past few years the Research Computing Unit has evolved an organization and procedure that appears to work well for a high volume mix of large and small projects in a Medical School environment.

Key aspects of this management system are:

(1) Fixed relationship between the client, statistician and the data manager. This assures that each member understands his responsibilities.
(2) A relational data base for all data on manager activities relating to a project and for generating management reports. This assures that all information can be interrelated for use by the data manager or management.
(3) Various software tools and programming procedures. These assure more efficient use of the data manager's time.

These points will all be discussed in more detail below. Though many aspects are unique to our shop, the concepts should have value in any SAS Data Management shop.

## 2. PERSONNEL RELATIONSHIPS

Each project is initiated with a meeting between the Statistician, Client and Data Manager. During this initial consultation, the goals of the research, the necessary data and appropriate Methodology are stated. At this time, a project log form is completed by the data manager and given to the unit's administrative assistant for entry into the data base.

The Client and/or Data Manager complete the data specification forms, which are keypunched and processed by data entry personnel. After processing, the raw data is verified by data entry personnel, and any necessary corrections made. An initial program is then given to the Data Manager containing a list of the data and summary statistics (PROCs PRINT,FREQ,UNIVARIATE). These are checked by the data manager for gross errors, and if in reasonable form, are returned to the client for final verification. The client is instructed to compile a list of changes to the original data in the form of transactions which are then keypunched and incorporated into the transaction section of the data generation program through an UPDATE.

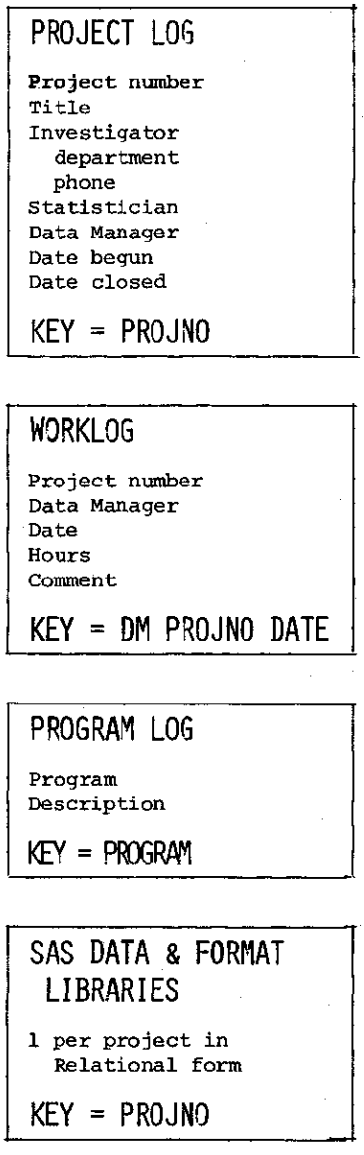
Once the data has been verified and updated, the analysis programs are run. The Statistician, Client and Data Manager meet again to discuss the results and possible future analyses.

## 3. DATA BASE

### 3.1 OVERVIEW

The diagram below illustrates the components of our system. Those components are:

(1) Project Log
(2) Activity Logs
(3) SAS Data Libraries
(4) SAS Format Libraries




Each component is a member of a relational data base, i.e. each has a unique key for merging. Project number alone or as a member of a composite is a key for all files. It is used in the activity logs, program names, SAS data and format libraries, project logs and reports.

### 3.2 PROJECT LOG

A Project Log is maintained by the Units Administrative Asssistant via a terminal in response to prompts. This file, updated as necessary, contains descriptive information on each project. A three digit project number is assigned to that project. This number identifies the project in all reports, is a key in all relational files and is essential in our naming conventions and activity logs.

The log is used as both a reference for project numbers and relevent information as well as an external file accessed for SAS generated reports and programs (SASGEN).

### 3.3 ACTIVITY LOG

Logs are maintained by all members of the data management staff to document time spent on each project. These logs are in two parts:

(1) Worklog - Time spent
(2) Program Log - Program descriptions

An individual's worklog contains the amount of time spent on an individual project per day, as well as an optional comment for possible future reference. A Program Log is also maintained containing each program's name and a one line description. This brief description of a program in the Program Log is by no means to replace documentation within a SAS program, but rather as an index to each program's purpose.

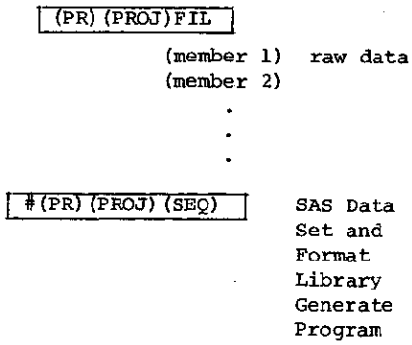
These logs are entered via a terminal in response to prompts gererated by an interactive program. To assure a relative level of accuracy, it is suggested that they be updated on a daily basis. These files serve two purposes. First, as an organizational tool for the data manager to coordinate his own time and resources. Secondly, these files are accessed by SAS report writing programs which generate reports on the unit's activity. The logs are checked for errors in both the interactive program as well as the SAS report writing program.

### 3.4 SAS DATA AND FORMAT LIBRARIES

All data is maintained using SAS data libraries with corresponding format

libraries. These libraries are conceptual components of the relational data base as the file names serve as defacto keys, allowing linkage of these libraries to information stored in other files. To serve this linkage function, the files must follow a strict naming convention.

The common naming convention is also followed for program names and raw data files. The SASGEN program generates code consistent with these conventions.




(PR) - Programmers Initials
(PROJ) - Three digit Project Number
(SEQ) - Sequence Number
(MNM) - Three Character Mnemonic

Our naming conventions are limited by the restrictions set by the implementation of the operating system at our computing facilities (CUNY). If the operating system were more flexible, a better convention would be possible.

The naming convention also allow for relative ease by which projects can be transfured or shared between data managers. However, transferring is the exception rather than the rule.

example

Programmer (JH) has a member is his PDS Library for project (003), the first for this project (000), which read the raw data JH003FIL(ORIG), generates a FORMAT Library (JH003FMT), and a SAS Data Library (JH003SAS) containing members (JH000RP0) amd (JH000RPM), containing repeated measures and repeated observations data sets. Subsequent data generating programs add members to these existing Libraries.

## 3.5 SAS GENERATED REPORTS

All of these components of our management system are tied together through SAS generated reports. These SAS report programs generate reports on the unit's activity, active research projects within the department and resources used. The unit's administrative assistant generates these programs through an interactive program.

# 4. TOOLS AND PROCEDURES

## 4.1 SASGEN

The SASGEN Program reads the data specification sheets and produces a SAS program containing corresponding INPUT statement, FORMATS, LABELS as well as SAS PROCedures to generate appropriate summary statistics. TITLE statements are generated using the Project Log File, as well as Investigators, Statistician, Data Manager and date of initiation.

## 4.2 TRANSACTIONS

All changes to the original data are implemented through transactions in the original data generating program. As a rule, the data should be completely defined by one program. A transaction data set is generated containing corrections to the original data, which is implemented with an UPDATE statement. This UPDATE is then checked for invalid corrections. Changes in the raw data using a text editor is never done. With this system, every change in the data is documented.

For additional information on SASGEN and Transactions, see Calhoun et.at.(1,2).

## 4.3 STYLE

To assure readable, reliable programs, a uniform style is necessary. We have recomended two books as a 'Style Guide' to our data managers (3,4).

In addition, bi-weekly meetings are held to review data management techniques, as well as keeping abreast of changes and enhancements in SAS. Typically, a data manager presents a program which presented him with a unique problem, and his solution to that problem. Criticisms, both positive and negative are prepared prior to this meeting, and discussed by all members of the unit.

Three topics are discussed when reviewing programs:

(1) style
(2) adequate defensive programming
(3) structure

A friendly atmosphere and open exchange is a must if this technique is to be helpful.

## 5. CONCLUSIONS

The management of a SAS based data management unit as described is a result of evolution. As operating systems change and SAS changes and new options become available, modifications are made in the system. A balance must be struck between saving time with this system, and generating work to maintain the system. This system has proved to be very maintainable, with very little overhead. Once the system is set up, most of the management functions can be maintained by a person not at all familar with computers.

More importantly, it gives the data manager a record keeping system of pertinent information concerning each of the many projects he might be involved with at any one time, and a management report on the unit's data management activity.

## REFERENCES


(1) Calhoun, W.F.,Blavatnik, L.,Dorph, D., 1981. Solutions to Two Data Management Problems: Data Definitions and Updating. SAS Users Group International Conference, 1981.

(2) Calhoun, W.F.,Holt, J.,Dorph, D.,Anderson, H., 1981 Data Management Problems with a Large Medical Survey. SAS Users Group International Conference, 1981

(3) Kernighan, B.W. and P.J. Plaugher 1978.
The Elements of Programming Style 2nd Edition., New York: McGraw Hill.

(4) Yourden, E., 1975
Techniques of Program Structure and Design,
New Jersey: Prentice-Hall.