

USING SAS TO MAINTAIN HISTORICAL DATABASES OF RESEARCH PROJECT INFORMATION

Susanne L. Daood, Mayo Clinic

The Health Sciences Research Department of the Mayo Clinic has a growing collection of more than 15,000 datasets, 30,000 programs, and 1000 programming requests acquired over the years for the purpose of analyzing clinical and research data. Managing this collection of files on the IBM mainframe has been the responsibility of the Statistical Systems unit of Information Services. In 1982, the Project Control System (PCS) was developed using SAS® to track, archive, and restore data, programs, and requests by statistical analysts. This system has proven its worth many times over.

During the last year, two events have prompted review of the PCS: 1) the adoption of a UNIX-based computing environment for statistical analysis and 2) the upgrading of SAS on the IBM/MVS operating system to Version 6. This paper will discuss the components and functionality of the present PCS, the conversion to a Version 6 based PCS, and the steps necessary to include information about UNIX based data and programs in the PCS.

® SAS is the registered trademark of SAS Institute, Inc., Cary, NC, USA.

USING SAS TO MAINTAIN HISTORICAL DATABASES OF RESEARCH PROJECT INFORMATION

Susanne Daood, Mayo Clinic
Randy Stick, Mayo Clinic

INTRODUCTION

The Mayo Clinic began collecting data for research purposes at the turn of the century and has never turned back in its efforts in this area. The value of the data collections is undisputed - providing the basis for publications, presentations and, ultimately, improvements in the quality of patient care.

The need for computer management of the research information was realized in 1980. At that time, the computing environment consisted of statisticians, epidemiologists, data analysts and programmers accessing data on an IBM® mainframe via terminals. SAS® was the analysis tool and PANVALET® and SAS were the primary receptacles for data. A large percentage of the data was abstracted from medical records and recorded on paper forms for data entry.

Today, the computing needs of statisticians and epidemiologists are vastly different from what they were in 1980. Servers, workstations and x-terminals networked together and to the IBM 3090 mainframes define the present environment. The analysis tools employed are SAS and S-PLUS.® Although much of the data is stored in SAS and PANVALET libraries, data sources range from data forms to PC databases to data extracted from Ingres and DB2® databases. The world of computing offers more options but the management of information resources is more complex and

perplexing.

Using SAS, a Project Control System (PCS) was developed by the Statistical Systems unit to manage data, programs and requests for the Department of Health Sciences Research in 1982. The evolution of this system is the topic of this paper.

ORIGINAL PROJECT CONTROL SYSTEM

It has been the experience of the programmers in the Statistical Systems unit that many of the personnel we support prefer to have operations such as management of information resources as transparent as possible. The more we can provide them with menu interfaces and behind the scenes processing the better they like it. They realize they have some responsibility but they wish to be shielded from the actual grueling details. Consequently, the development of the original Project Control System was strongly supported.

The objectives of the original system were to give users the ability to:

1. locate all the known data and programs associated with a study.
2. restore and archive data and programs on an overnight basis.
3. detect problem situations with respect to study data.
4. enter restore, archive and request information into an online application.

SAS was used as the database for storing the information about the resources and SAS was the programming tool for controlling the functions. In 1982, SAS/FSP® and SAS/AF® were not available so ISPF dialog manager was chosen for the transaction entry system. Special utilities were used to identify OS datasets associated with research studies and to read the PANVALET directories. A SAS procedure was written to process the transaction file data and the PANVALET information.

The use of the system was simple and its benefits many. Users adapted to standards requirements because the benefits of following the rules and letting the system do the rest paid off. Time spent in searching for the "right" data and in identifying the contact people for a study decreased markedly. Appendix I illustrates the type of information available about a particular study at any user's terminal. In addition to data and program information, request and accounting information was stored in other SAS databases and was available for reports and special retrievals.

The system was based on the following components:

- A) An on-line transaction system to collect data/program movement requirements.
- B) The processing of these transactions in a series of nightly batch jobs. In addition to error checking the transactions, this series of jobs read all the on-line PANVALET library directories and the OS on-line files, set up jobs to move files to and from holding files for the restore and archive functions, created jobs to read the PANVALET and SAS archive tapes for restoring necessary
- resources, generated on-line files of study information (see Appendix I) and generated reports confirming the completion or rejection of transactions.
- C) A series of weekend batch jobs to do the actual archiving of PANVALET and SAS files and to generate reports on the status of the information in the various databases.

Dependence on this system was so heavy that its continued existence was one of the requirements for making any changes in the computing environment in Health Sciences Research.

THE 1992 UPGRADE TO THE PCS

Over the past ten years the PCS has served its users well. As times changed and technology brought new hardware and software options, the need for a new system became more apparent. The goal with this set of modifications is to adapt to the current operating environment and to initiate a design for future archiving needs.

The current design for Mayo's research computing environment includes moving most computing to the UNIX® world. The backups for emergency recovery of data are handled on the UNIX side but, due to lack of personnel and a desire to stay out of the computer operations arena, the archiving will continue to be done on the IBM side. It will be necessary for quite some time to be dependent on the IBM for archived data and for locally written SAS procedures.

The objectives of the 1992 update to the Project Control System are as follows:

- 1) Rewrite the entire set of code in SAS Version 6.07.
- 2) Convert all the SAS archived data to SAS Version 6.07.

- 3) Convert the ISPF Dialog Manager transaction collection system to a SAS/AF application using SAS/FSP and SAS/SCL.[®]
- 4) Allow archiving and restoring of data and programs from both the UNIX and MVS environments.

The rewrites of the programs and conversion of the SAS data to Version 6.07 went smoothly and were relatively easy to accomplish. The developer of the SAS/AF application was unfamiliar with SCL but, with some help from a more experienced SCL programmer, a lot of manual reading and some experimentation a reasonable and powerful alternative to the ISPF Dialog Manager application was created. Appendix II includes the SAS/AF main menu for this application and illustrates the functions allowed.

The most interesting part of this upgrade to the PCS was the incorporation of the networked UNIX workstations and x-terminals into the IBM mainframe archiving system. The Statistical Systems unit had previously developed scripts for execution at a user's workstation to move files between the UNIX and the MVS environments. These scripts used FTP and allowed a UNIX user to move data between computing environments without logging on to the IBM mainframe. These scripts were customized to move the files needed for archiving and restoring of files from the PCS.

The primary components of the PCS in the UNIX environment are:

- 1) An INGRES[®] application was written which allows the user to enter transactions into an Ingres table. This same application presents an option to view the same listings available on the IBM side about data and programs for

each study. See Appendix II for the INGRES main menu.

- 2) After a transaction is submitted, either the data to be archived is moved to the IBM or a script command is setup to move restored data back to users UNIX directories early the next morning. If there was a problem with the movement of the file, a mail message is sent to the requestor.
- 3) The file of transactions is shipped to the mainframe late at night to be processed along with those submitted on the IBM side.
- 4) By the next morning, the requested files needing to be restored from archive tape are available in the user's directory.

FUTURE ENHANCEMENTS

The biggest disadvantage of the present PCS is the inability to locate data and programs stored on the UNIX machines. This requires some additional standards to be established and scripts to be written to extract names of files from study directories. The UNIX experts are planning to complete this update of the PCS sometime this year. At that time, users would be able to completely define the data and programs associated with a study.

The additional portion of the PCS to be completed will allow users to archive non-SAS and non-PANVALET files from the UNIX environment. There are many UNIX files created that do not fit the SAS and PANVALET molds. Examples of this are: large binary files, all the files in a particular directory, and large text files. The plan is to add this function to the PCS

by allowing transactions to be entered into the INGRES application asking for special archiving capabilities. These files would be moved to the mainframe and archived using the data storage management system used on the mainframe. The PCS will allow users to supply comments to more fully identify the files.

CONCLUSION

SAS continues to supply the tools to allow users to adjust to changing computer environments and new requirements. The ease with which SAS data and programs can be converted from Version 5.18 to 6.07 simplified the developers job. The improved application development tools, SAS/SCL, SAS/AF and SAS/FSP, and the availability of SAS 6.07 under UNIX made it logical to continue to develop a large portion of the system in SAS.

ACKNOWLEDGEMENTS

SAS, SAS/AF, SAS/FSP and SAS/SCL are registered trademarks of SAS Institute Inc., Cary, NC, US.

INGRES is a registered trademark of ASK Computer Systems, Inc.

DB2, IBM and ISPF are registered trademarks of International Business Machines Corporation.

PANVALET is a registered trademark of Pansophic Corp.

S-PLUS is a registered trademark of Statistical Sciences, Inc.

UNIX is a registered trademark of AT&T, Inc.

AUTHOR INFORMATION

Susanne L. Daood,
Lead Analyst/Programmer
Mayo Clinic
200 First Street SW
Rochester, MN 55905
Tel:(507)284-5585
EADDR:daood@mayo.edu

APPENDIX I

Example of Data and Programs for a Study

13449 WHISNANT	JDS	I13449FZ	SAS ARCHIVE	3134	JDS	S I 03/21/92	FREEZE ON 3/20/92 BEFORE ADDING NEW CASES
		I13449HP	SAS ARCHIVE	38	MMH	S I 05/15/91	EST. POP OF HI PROB OF EMBOLUS
		I13449HT	SAS ARCHIVE	38	MMH	S I 05/15/91	USE OF 80-84 % OF HT TO SET. POP OF HYPERT.
		I13449IH	SAS ARCHIVE	38	MMH	S I 05/15/91	EST. POP OF IHD
		I13449JB	SAS ARCHIVE	1524	JYB	S I 05/24/91	CONTROLS -- FROZEN 10/17/90
		I13449MH	SAS ARCHIVE	1382	MMH	S I 01/13/90	STROK IN ROCHESTER, 60 TO 84
		I13449PY	SAS ARCHIVE	714	CYS	S I 02/21/91	P=CYS PERSON YEARS N=1380
SR.JDS.S13449.SAS				STAT00	SAS	JDS D	W A 06/12/92 SAS PRIVATE WORK DATABASE
SR.SLD.S13449.PDS				STAT02	PDS	SLD P	S A 08/30/88 SUPPORT OF ON-LINE APPLICATION
SR.SLD.S13449.SAS				STAT02	SAS	SLD B	S A 08/30/88 DATA, SCREENS AND PROGRAMS FOR ON-LINE APP
SR.SLD.S13449.SAS01(0)				025962	SAS	SLD D	S I 07/21/92

13449	T13449CS1	PANLIB ARCHIVES	T I	82	CYS	06/25/91	P=CYS STROKE DATA PROC PERSON YEARS
	T13449HO1	PANLIB ARCHIVES	T I	121	MMH	11/06/90 04/15/91	* GET CONTROL PCT OF HI RISK FACTORS
	T13449HO10	PANLIB ARCHIVES	T I	277	MMH	01/09/91 04/15/91	* GLIM FOR DR. INGALL
	T13449HO11	PANLIB ARCHIVES	T I	580	MMH	01/10/91 04/15/91	* EST. POPULATION BY % OF CONTROLS
	T13449HO12	PANLIB ARCHIVES	T I	277	MMH	01/16/91 04/15/91	* GLOT
	T13449HO2	PANLIB ARCHIVES	T I	377	MMH	11/07/90 01/23/91	* ADJ. POPULATION BASED ON PCT OF HI RISK FACTOR
	T13449HO3	PANLIB ARCHIVES	T I	618	MMH	11/07/90 08/19/91	* INCIDENCE USING ADJ. POPULATION
	T13449HO4	PANLIB ARCHIVES	T I	940	MMH	11/13/90 08/19/91	* EST. PREV. RATES FOR HI PROB OF EMBOLI
	T13449HO5	PANLIB ARCHIVES	T I	37	MMH	11/13/90 01/23/91	* CASES FROM MED INDEX
	T13449HO6	PANLIB ARCHIVES	T I	264	MMH	11/28/90 01/23/91	* PCT OF CONTROLS(25 YEAR PERIOD)
	T13449HO7	PANLIB ARCHIVES	T I	187	MMH	12/06/90 08/19/91	* 3 YR RUNNING AVERAGES FOR CONTROLS
	T13449HO8	PANLIB ARCHIVES	T I	234	MMH	12/13/90 01/23/91	* 12/13/90 REQUEST: 25 YRS PCT OF HI PROB
	T13449HO9	PANLIB ARCHIVES	T I	88	MMH	12/14/90 04/15/91	* FIND UNMATCHED CONTROLS & CASES
	T13449JB1	PANLIB ARCHIVES	T I	203	JYB	09/10/90 05/23/91	* P=JYB I=INGALL FREQS FOR VARIOUS CARDIAC DISEAS
	T13449JB2	PANLIB ARCHIVES	T I	244	JYB	09/17/90 05/23/91	* P=JYB REQ=9/17/90 CARDIAC DIS VS 3YR RUNNING IN
	T13449JB3	PANLIB ARCHIVES	T I	139	JYB	10/16/90 05/23/91	* P=JYB REQ=10/16/90 FREQS OF HEART DIS FOR CONTR
	T13449JB4	PANLIB ARCHIVES	T I	131	JYB	10/22/90 05/23/91	* P=JYB REQ=10/19/90 CONTROLS-HEART DIS VS 3YR IN
	T13449JB5	PANLIB ARCHIVES	T I	166	JYB	10/24/90 05/23/91	* P=JYB REQ=10/23/90 CASES - MORE 3YR RUNNING INT
	T13449JB6	PANLIB ARCHIVES	T I	168	JYB	10/24/90 05/23/91	* P=JYB REQ=10/23/90 CONTRLS - MORE 3YR RUNNING I
	T13449JB7	PANLIB ARCHIVES	T I	177	JYB	10/29/90 05/23/91	* P=JYB REQ=10/26/90 ALL PTS-3YR RUNNING ON HYPRT
							* P=JYB REQ=12/5/90 HEART DES FREQS- NEW AGE GRPS

APPENDIX II

Main Menus for SAS/AF and INGRES Application

PROJECT CONTROL SYSTEM	
OPTION	
1	Use the Transaction File
2	Use the Request File
3	View PCS Listings
4	HELP!!!!
5	Update OS Data Documentation

