Designing Databases
Using a Customized SAS/AF® Frame Entry Application

Esther Kwan, Ischemia Research and Education Foundation, San Francisco, California, USA

ABSTRACT

A customized SAS/AF application was developed to facilitate the design of a large research database which has to be applicable in three software systems. This paper gives an overview of this innovative SAS application, as it is related to the project work requirements. SAS automation organized and improved the efficiency and accuracy of this database design task.

INTRODUCTION

An international multi-center clinical research project is being conducted at Ischemia Research and Education Foundation (IREF) to study heart surgery patients in over 60 medical centers. Its in-hospital data collection tool is a questionnaire or Case Report Form (CRF) that is over 100 pages in length. In addition, three other questionnaires collect follow-up information about the patients after they are discharged from the hospital.

Based on various considerations, a portable customized Microsoft Visual Basic data collection system will be used to capture the data at the study sites. The data entry screens will be linked to a Microsoft Access database. The Access data from the field sites will be transferred to and 'warehoused' in an Oracle® Data Management System at IREF. Eventually, the study data will be extracted from the Oracle database to the SAS System for reporting and statistical analyses.

These research data will flow through three software systems. A database must be designed and created for each of these systems by programmers in these respective groups. Designing a SAS database for the CRF requires designing a variable name, label, type, length, and, when applicable, a format for each data field on the 100+ pages of the data form.

Since these databases must be parallel in design to ensure the integrity of the data during the flow, the design should be straight forward, i.e., one design for all software systems. However, the database structures among these systems are not identical, which complicates the design task. For example, numeric variable type in SAS is 'numeric'; this type is called 'number' in Oracle and 'integer', 'long' or 'single' in Microsoft Access depending on the size and format of the numeric value.

The three databases must be designed concurrently to ensure parallelism. Despite documented specifications and standard operating procedures (SOP), consistency within database requires that the design work be performed by one programmer. The CRF database alone will have at least 6,000 variables partitioned into more than 50 data sets; each variable will have attributes specific for Microsoft Access, Oracle and SAS. It was apparent that designing the databases for this study would be a lengthy and tedious task for one person.

To facilitate the design of the databases for this study, a customized application was developed using SAS/AF frame entries, SAS/FSP®, BASE/SAS®, SAS/Assist® and the Netscape Web browser software. The ***Database Design System*** is a UNIX-based application developed on version 6.09, but currently running on version 6.11 of the SAS System.

PROJECT WORK REQUIREMENTS

The CRF is partitioned into short logical sections. The database is designed one section or data set at a time. For each data field in the section, a variable is assigned in the data set for that section. The attributes of this variable are designed for SAS, Oracle and Access. Concurrently, the 'coding dictionary' or format library is built for all the coded data fields in the CRF.

Upon completion of the design, each assigned variable is verified against the design specifications, such as is the naming convention for the variable correct. The variable attributes are checked for validity, e.g., the variable name assigned cannot be more than 8 characters long. The variables are listed for comparison against the data form, to ensure that complete and correct assignments have been made in the data set. The design is listed to be circulated to the Quality Assurance (QA) Unit and another member of the study team for verification.

Finally, the appropriate design information is output for the Microsoft-Access Group and the Oracle Group to create the database for their respective software system. (The database structure is not needed for the SAS System until the study data are extracted from the Oracle System.) The final design is documented for the database managers and end-users. All changes to the design made after distribution are tracked for 'audit trail'. The progress of this project is documented and its status is reported to the study team periodically. Most of these tasks, including some actual assignments of variable attributes, are automated in the *Database Design System*. The following flow diagram summarizes the procedural steps of this project:

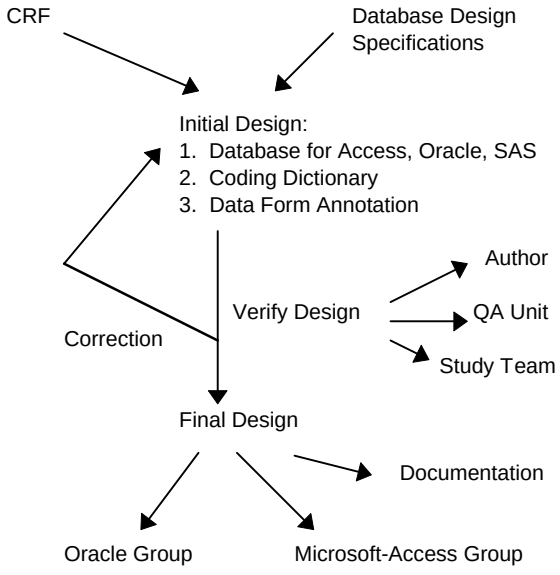

Figure 1 Database design work flow diagram

APPLICATION OVERVIEW

A SAS database is used to capture the designs of these three parallel databases. For a given study data set, such as a section of the CRF, the design or variable attribute information for the various systems are keyed into a design data set (Table 1). One design data set is created for each study data set to be created. The coding dictionary or format library information is captured in a separate data set (Table 2).

Table 1 Contents of design data set

Variable	*Type*	*Length*
Study name	C	6
Database type	C	15
Data field ID: CRF section number	C	3
Data field ID: page number	C	3
Data field ID: question number	C	5
Long data field description	C	140
Variable label	C	40
Variable name	C	8
Variable type SAS	C	1
Variable type Access	C	10
Variable length	C	3
User-defined format/code name	C	10
Variable SAS format	C	10
Data set name	C	12
Variable number	N	8
Date-time design last modified	N	8

Table 2 Contents of coding dictionary data set

Variable	*Type*	*Length*
Study name	C	6
User-defined format/code name	C	10
Data code & resolved text 1	C	42
Data code & resolved text 2	C	42
Data code & resolved text 3	C	42
Data code & resolved text 4	C	42
Data code & resolved text 5	C	42
Data code & resolved text 6	C	42
Data code & resolved text 7	C	42
Data code & resolved text 8	C	42
Data code & resolved text 9	C	42
Data code & resolved text 10	C	42
Data code & resolved text 11	C	42
Data code & resolved text 12	C	42
Data code & resolved text 13	C	42
Data code & resolved text 14	C	42
Data code & resolved text 15	C	42

For a given study data set to be designed, the variable labels and the long data field descriptions are created first as an ASCII or text file outside of the application. The *Database Design System* serves to facilitate and automate the remaining work requirements. The system main menu has the following task options:

1. Set up design data set
2. Open design data set
3. QA design data set
4. List design
5. Output design files
6. Project progress
7. View design specs
8. SAS/Assist
9. Help
10. Exit

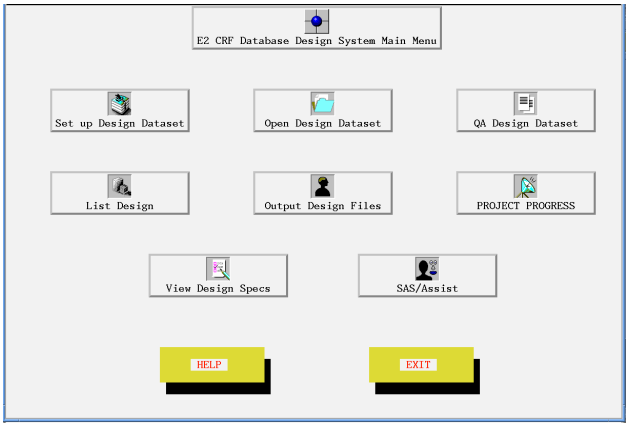

Figure 2 Database Design System Main Menu

Many of these menu options have a link to a sub-menu of additional task selections. An overview of the function of each main task option is described below.

1. Set Up Design Data Set

The design data set for a given study data set is created using an empty master design data set (Table 1) and the variable label text file.

Prior to the making of the data set, this option allows for 1) the selection of a database type (CRF or Follow-up), 2) the naming of the data set, 3) the browsing of the text file via PROC FSLIST of SAS/FSP and 4) the selection of a data set label for the design data set.

During the creation of the data set, some of the design work are initialized. For example, the assignment of the type and length for certain data fields is automated based on keywords found in the variable label. The key variables and their attributes are duplicated from the first completed data set of the database to subsequent data sets.

The newly created design data set is checked for accuracy via PROC FSBROWSE of SAS/FSP and the standard output of PROC CONTENTS.

2. Open Design Data Set

A design data set is opened in PROC FSEDIT or PROC FSVIEW of SAS/FSP to enter, edit or browse the design information (Table 1).

The coding dictionary data set may be opened to enter, edit or browse the coding information (Table 2).

The coding dictionary data set is linked to the design data set at the format name entry field. This linkage allows efficient selection of predefined format name and addition of new formats to the dictionary.

3. QA Design Data Set

Error check on the completed design in 1) all completed data sets, 2) one data set and 3) the coding dictionary is accomplished using SAS programming:

- Identify duplicate variable names within a data set and in all completed data sets.
- Identify missing variable name.
- Check invalid variable attributes, e.g., incorrect variable type.
- Verify the design against items in the design specifications, e.g., the length for all open-ended text fields should be '40' as specified.
- Check parallel design across the 3 software systems, e.g., if SAS variable type is numeric, Access type should be 'integer', 'single' or 'long'.
- Verify assignment of format/code name in the coding dictionary, i.e., every user-defined format name used in the database should be present in the coding dictionary.
- Identify duplicate format/code name in the coding dictionary.
- Identify missing format/code name in the coding dictionary.

The error reports are printed for documentation.

4. List Design

The design is listed and printed for manual verification during the quality assurance process.

The records (variable designs) that were changed after a specified date are listed and printed for 'audit trail'.

5. Output Design Files

The appropriate design information is output as separate ASCII or text files for the Microsoft Access and Oracle groups to create their databases.

The variable names and their applicable code names are output also as an ASCII file by data form page number. This listing is used for type-setting the 'Annotated CRF'. An Annotated CRF is a database user document, which is a blank data form listing next to each data field the corresponding variable name and applicable code name in the database.

6. Project Progress

Ongoing documentation and reporting of the progress of the project are possible with the linkage of a progress data set (Table 3) to the application.

A sub-menu option opens the progress data set in PROC FSVIEW of SAS/FSP to enter, edit or browse the project progress information. After the sequential steps of designing a study data set are completed, the current date is keyed into the progress data set via PROC FSVIEW of SAS/FSP. Another option list the project progress information for printing. This documents what has been done and when.

Table 3 Contents of project progress data set

Variable	*Type*	*Length*
Study data set ID (form section #)	C	5
Study data set description	C	50
Date completed annotated CRF	C	5
Date completed variable labels	C	5
Date set up design data set	C	5
Date completed data set design	C	5
Date checked coding dictionary	C	5
Date checked duplicate var name	C	5
Date error checked design	C	5
Date listed design	C	5
Date output design files	C	5
Date design QA'ed by author	C	5
Date distributed design to QA Unit	C	5
Date distributed design to 2nd QA	C	5

Another menu option reports the status of the database being designed. Dynamic SAS programming list the database sections completed and the current cumulative total count of variables completed in the database.

7. View Design Specs

This option allows browsing or referencing the locally stored Database Design Specification document through Netscape, a Web browser software.

8. SAS/Assist

This option provides a link to SAS/Assist for quickly performing any unanticipated management, reporting or analysis tasks of the design information.

9. Help

This option provides on-line instructions for using the application successfully.

Figure 3 at the end of this paper shows an overview of the application in flow diagram format.

In the future, the *Database Design System* may be upgraded to allow efficient and general design of SAS/Oracle databases for any research and administrative projects.

This application originated from a series of SAS codes to be submitted sequentially to perform each of the above tasks. It took less than a month for a new user of frame entry to integrate the set of programs into this application. The *Database Design System* is simple in appearance, but functional, and it provides many benefits.

APPLICATION BENEFITS

The *Database Design System* was used successfully to complete the design of a database with more than 6,500 variables partitioned into 57 data sets. Each variable has attributes specific for Microsoft Access, Oracle and SAS. Minimal errors -- none related to the compliance of the design

specifications, were identified by the quality assurance process. Annotation was completed for the 10,000+ data fields on the 100+ pages of data form with accurate variable name and code name association.

The *Database Design System* provided a centralized environment to implement this database design project systematically and efficiently. The system main menu and the sub-menus show the sequential steps and requirements for the design procedure. The system reminds the user what is required, which facilitates and simplifies staff cross-training. The system assures consistent and correct design for many aspects of the completed database. The system reports the dynamic status of the database being designed, or the end-product being delivered. More importantly, development of automation promoted thorough planning and organization during the initial phases of the project.

ACKNOWLEDGMENTS

The author wishes to thank Long Ngo, Ahvie Herskowitz, M.D. Colleen Stewart and Juliana Kipps for their editorial comments during the preparation of this paper. Dr. Ahvie Herskowitz, the Study Director, and all the members of the Study Team are acknowledged for their support in this database design project.

The author may be contacted at her new address:

Esther Kwan
IBAH Resource Biometrics Clinical Software
5801 Christie Avenue, Suite 355
Emeryville, CA 94608
USA

Phone number: (510) 597-7200
FAX number: (510) 420-0651

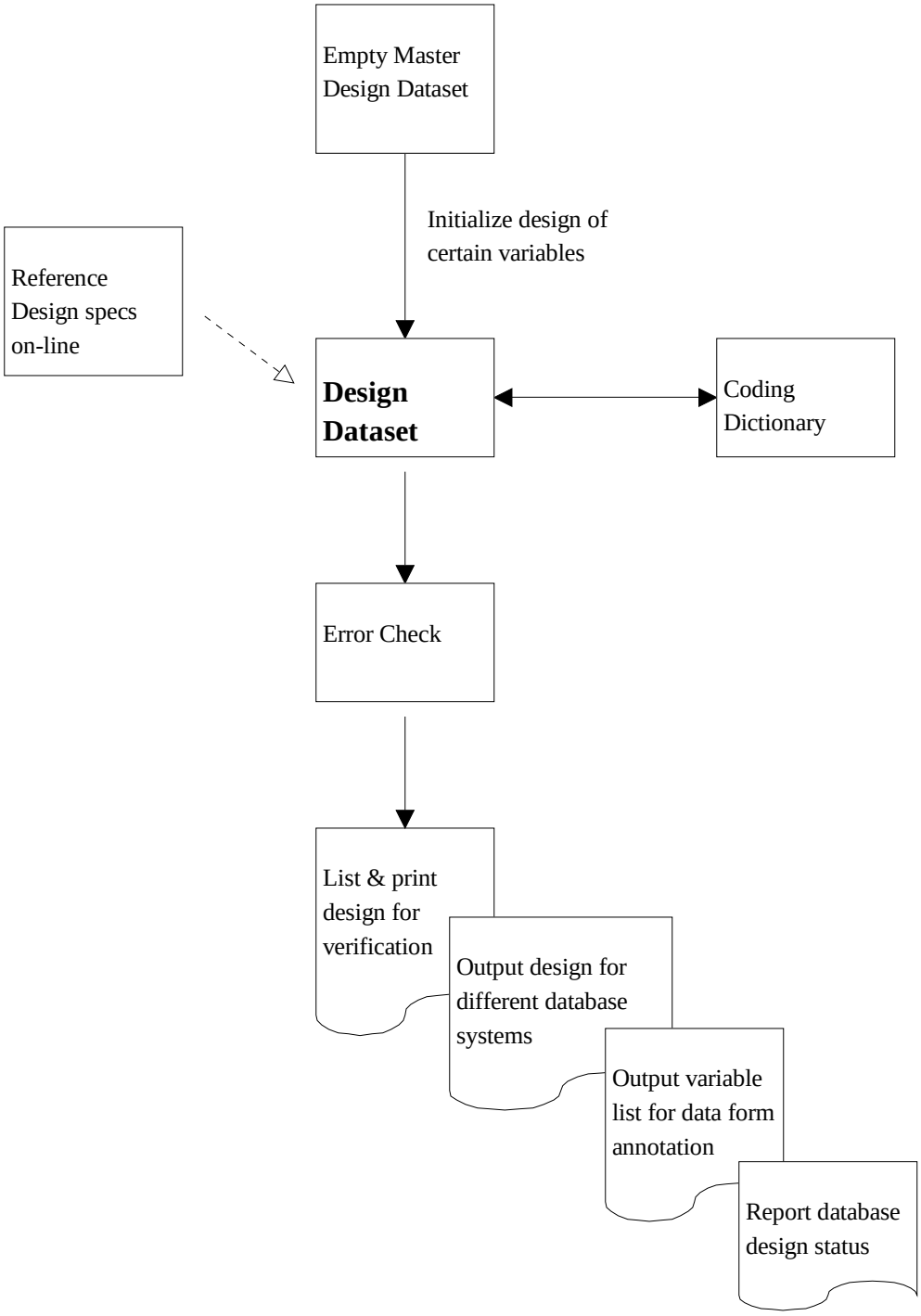

Figure 3 Database Design System Overview