

DEVELOPING SAS® APPLICATIONS USING INFORMATION ENGINEERING METHODOLOGY

Priscilla Van Grevenhof, Mayo Clinic
Roger Mueller, Mayo Clinic

ABSTRACT

The Mayo Clinic adopted IE (Information Engineering) methodology as the standard development methodology for systems in 1988. The Texas Instruments IEF (Information Engineering Facility)™ Toolset was the automated toolset chosen for application development. For institutional data the target platform is DB2.™ In the Department of Health Science Research IE methodology is also used but the target platform may not be DB2. The many research projects done at Mayo are short lived, from 1 to 3 years, and one of the tools used is SAS.® SAS as the target platform is chosen because of its flexibility and short development time. These attributes match the needs of the research community. This poster will show how IE methodology is used to develop a SAS application and how IEF objects are related to SAS objects.

INTRODUCTION

At the Mayo Clinic development of information systems for clinical research benefits from the use of system development methodology using Information Engineering (IE).^{1,2,3} The case tool selected to support the methodology is the Information Engineering Facility (IEF).™⁴ This methodology approach is platform/target environment independent.

Many of the research projects have a defined life span of several years. These research projects benefit from the use of the IE methodology but are not targeted to a production database such as DB2.™ Instead many reside in SAS tables. To the investigator, the advantages of the SAS® environment is that it is the user friendly environment that they have been using for years. Added to this is the ability to rapidly develop an

application in SAS with the look and feel of a relational model through the use of views and SQL.

Many papers have been presented at other SAS conferences discussing relational theory and database design, and it is assumed the reader has some knowledge of this subject.^{5,6,7,8}

PROJECT DEVELOPMENT

Using IEF for project development with SAS has the same benefits as any IEF target environment.

1. Better defined system
2. Identify shared functions and data
3. Develop better understanding between clients and application developers
4. Develop reusable code and models
5. Excellent documentation

This paper assumes that the project is being developed under an enterprise wide Information Strategy Plan (ISP). If this is not true, then more work is necessary in determining that the same information is not collected in more than one system, and it is more difficult to determine the scope of the project.

A Mayo research project is initiated with the development of a detailed research protocol. This is a high level description of the information needed and the activities necessary to support the project from project sponsors. This is the scope on which the Business Area Analysis is performed. This usually describes only about 25% of the information that project will require.

The project team is then identified and a project plan developed. The project team, in the initial phases, is comprised of researchers, statistician,

data coordinator, analysts, project leader, and scribe. A scribe is a person versed in IE methodology and the IEF tool set. Technical support areas are called upon when needed.

The first Joint Application Design (JAD) session is a scoping JAD. At this time the JAD project team members are instructed in the terminology and symbols that are used in data modeling and process modeling. They are given the definitions of entities, relationships, optionality and cardinality. The research protocol is used to determine that the scope of the project is not exceeded. The future needs of the project are discussed rather than the current operation.

Function analysis is done at the same time as the data analysis. The scribe documents in the IEF toolset the discussions of the JAD participants, separating the entity analysis from the function analysis. The Activity Hierarchy or process model is the resulting document of function analysis. The Entity Relationship Diagram is the result of data analysis.

After the model and activities are defined, interviews are conducted to determine the information needs of the project. This may result in a cross function JAD session to examine the data that may be in the form of shared data. The existence of the data in a shareable form does not necessarily mean that the researcher has all the information from that source. The electronic form of the data may not be inclusive of the information needed in a retrospective study. Some of the information may need to be collected from a paper copy. The medical history is one form of historical information found in paper copy form.

The deliverables from the JAD process are a data model and a process model. From the IEF toolset, the documentation would be the Entity Relationship Diagram (ERD) and the Activity Hierarchy Diagram (AHD). This is an easy form for the users to review and approve since their involvement was enlisted from the beginning. The project starts with documentation done up front.

SAS APPLICATION DEVELOPMENT

When review is complete, the Business Area Analysis tasks of the ERD and AHD are described and documented in the toolset. The discussion to follow describes how IEF objects from the PC tools are equated with their SAS counterparts.

The Entity Relationship Diagram is represented as a SAS data library. The data model is divided into subject areas which are groups of entity types of common interest. Subject areas add to the understanding of the data model. The subject areas can be described as separate SAS data libraries. If a subject area has many entity types, separate SAS data libraries may help in maintaining and understanding of the system. At Mayo, the method of using multiple SAS data libraries is in the discussion stage at this time.

An entity type is related to a SAS table and is the information that the project needs to keep. The entity contains information with common characteristics and participates in common relationships.

It may be necessary to create a subtype of an entity type. This is done so that additional information or relationships can be associated to the members of the subtype. An example of a subtype of an entity is a PATIENT entity type which has a subtype of DOMESTIC and FOREIGN. In this case the attribute SOCIAL SECURITY NUMBER is not part of the subtype FOREIGN. In the case of a subtype, the table contains all attributes in the entity type but the screen shows only the correct attributes to enter the information depending on the classifying attribute of PATIENT TYPE.

An attribute is the piece of information that is necessary for the project. It corresponds to a variable in the SAS dataset. The properties of the attribute are the type (numeric, character), length, format and informat.

An identifier is an attribute(s) and/or relationship(s). Each row in a table must be unique based on the identifiers. These identifiers are either foreign keys or primary keys and are

identified in the SAS table as indexes. We have used SAS datasets created from the output of proc contents to produce a SAS based data dictionary. The keys are identified in the data dictionary.^{9,10}

A relationship associates one entity type with one or more entity types. To enforce this in our systems we select the attributes where the pairing will be made. To make this easier for the user, standard views are created and appropriate joins are done using PROC SQL.^{11,12} Optionality and cardinality are also handled by the views. The logical model may have many to many relationships but the physical model resolves those relationships into a table.

The process model is developed using activity analysis. An activity acts on the data; it makes a change to the data. If there is no activity on the data, it is declared out of the scope of the project and dropped. The activity is then decomposed into elementary processes. Elementary processes leave the SAS dataset in a consistent state and is a single execution. This is the lowest level of activity and can be translated into SAS MACROS and SAS/AF[®] and SCL programs.

Many of the processes are create, update, and delete done at data entry time. The SAS complement to this is SAS/FSP[®] and SCL. To maintain integrity of the data all changes to the data are done through the procedures written to support the model.

CONCLUSION

Think of this paper as a primer on using CASE Tools to develop a system destined for the SAS environment. Much more could be written on this topic as many additional elements in the IEF product could be translated into a SAS counterpart.

The drawbacks of targeting to the SAS System are few and some can be overcome. The drawbacks are: 1) There is no way to directly build SAS tables from the ERD Model. 2) The Elementary process does not result in the ability to generate SAS code. 3) There is no formal data dictionary in SAS.

The following benefits far outweigh the drawbacks. 1) Using IE Methodology for SAS development supports the institutional goal of using a standard methodology approach. 2) Because of heavy user involvement in all stages of project development, IE methodology ensures that the resultant product is the system the user wanted. 3) With the user supplying the documentation up front during the JAD sessions and with the aid of automated tools, documentation is done more efficiently. 4) Over time the business will build an inventory of common data elements and procedures, thus reducing project development time. 5) It is possible to develop a generic model that fits most research applications. This generic modeling process requires additional study. 6) SAS is a familiar software product at Mayo since SAS72. Using SAS, therefore, does not require additional training of developers or users.

If SAS procedures were written to interface with the IEF toolset, most of the drawbacks would be eliminated and great efficiencies would be realized in project development.

ACKNOWLEDGMENTS

Special thanks to Marie Eidem and Kristine Beck for their aid in presenting this poster.

TRADEMARKS

Information Engineering Facility and IEF are trademarks of Texas Instruments Incorporated.

SAS, SAS/FSP, SAS/AF are a registered trademark of SAS Institute Inc., Cary, NC.

Database 2 is a registered trademark of IBM Corporation.

REFERENCES

1. A Guide to Information Engineering Using the IEF™: Computer-Aided Planning, Analysis, and Design. Texas Instruments Inc., 1990.

2. Martin, J. Information Engineering: Book 1--Introduction. Englewood, NJ: Prentice Hall, 1989.
3. Martin, J. Information Engineering: Book II--Planning and Analysis. Englewood, NJ: Prentice Hall, 1990.
4. Martin, J. Information Engineering: Book III--Design and Construction. Englewood NJ: Prentice Hall, 1990.
5. Baer, Darius S. DB2™ and the SAS® System - Information Delivery for the 1990s. Proceedings of the Fifteenth Annual SAS® Users Group International Conference, Cary, NC: SAS Institute, Inc., 1990.
6. Johnson, James R., Cornejo, Roger D. Interfacing Normalized Relational Database Structures with SAS® Software. Proceedings of the Fifteenth Annual SAS® Users Group International Conference, Cary, NC: SAS Institute, Inc., 1990.
7. Moorefield, Jean. Using Relational Database Technology and the SAS® system to Develop SAS® System Applications under the VMS Operating System. Proceedings of the Fifteenth Annual SAS® Users Group International Conference, Cary, NC: SAS Institute, Inc., 1990.
8. Sloan, Faith Renee. Relational Database Design Theory: An Overview. In Proceedings of MWSUG'91, edited by LeRoy Bessler, 217-222. Fox Point, Wis.: MidWest SAS Users Group, 1991.
9. Johnson, C., Arrighi, M., VanGrevenhof, P., Mueller, R., Panser, L., Lee, R., Chute, C. Some Considerations in Designing SAS®-Based Data Management Systems for Large International Epidemiologic Studies. Proceedings of the Seventeenth Annual SAS® Users Group International Conference, Cary, NC: SAS Institute, Inc., 1992.
10. Kretzman, Peter. Why Be Normal? Constructing SAS-based Tools to Facilitate Database Design. Proceedings of the Sixteenth Annual SAS® Users Group International Conference, Cary, NC: SAS Institute, Inc., 1991.
11. Gerend, Patricia L. Normalized Data: Implications for the use of SAS® Software, SQL, and Views. Proceedings of the Sixteenth Annual SAS® Users Group International Conference, Cary, NC: SAS Institute, Inc., 1991.
12. Kretzman, Peter. Old Wine, New Wineskins: Why the SQL Procedure Will Change the Way You Think. Proceedings of the Sixteenth Annual SAS® Users Group International Conference, Cary, NC: SAS Institute, Inc., 1991.

The author may be contacted at:

Priscilla Van Grevenhof
 Senior Analyst/Programmer
 Mayo Clinic
 200 First Street S.W.
 Rochester, MN 550905
 Tel:(507)284-5585
 Fax:(507)284-1516