Paper 184-28

The Use of Scripting Languages, Database Technology, and SAS/IntrNet® to Revolutionize the Research Process

Kristin Rankin, Center for the Advancement of Distance Education,
University of Illinois, Chicago, IL

ABSTRACT
The Internet has not been utilized to its fullest potential in the research arena, where the organization, quick communication, and security that it provides are crucial to the advancement of scientific knowledge. The model suggested here greatly improves the efficiency of the research process. The University of Illinois at Chicago, Center for the Advancement of Distance Education has developed a Research Survey Composer that allows you to create a data entry form that displays on a web page and stores data in database. SAS/ODBC® allows for the extraction of data from the database into SAS® software. Then, you can create a web page form that prompts the user to enter a query of the data. The SAS/IntrNet® broker is called when the web form is submitted, prompting existing SAS macros to perform analyses on the data. The Output Delivery System (ODS) can be used to format output in any way you desire, including finished tables, reports, or graphs on the web. Researchers can access these results from any place around the world without having any knowledge of SAS programming. This process can be executed on the Windows or Unix operating systems by intermediate to advanced SAS users.

INTRODUCTION
The Internet has quickly become the most widely-used and efficient tool for information transfer. Businesses and corporations rely heavily on this tool for daily communication and commerce. Despite the wide success of agencies that use the Internet, it has not been utilized to its fullest potential in the research arena, where the organization, quick communication, and security that the Internet provides are crucial to the advancement of scientific knowledge. Use of the model suggested here will alleviate some of the problems that delay the progress of many current research projects and will greatly improve the efficiency of the research process.

THE RESEARCH PROCESS
The status quo for many academic research projects involves a team of investigators who each specialize in one area and converge to share their findings. Often, the principal investigator for the study does not specialize in technology and relies on research assistants to use sophisticated statistical analysis tools, such as SAS®, to manipulate the data. Information transfer occurs by the principal investigator outlining a set of descriptive analyses to be performed by the research assistant, who runs the analyses in SAS and transfers the desired results into tables in a word processor. The research assistant eventually prints or e—mails the results back to the principal investigator for review. This time-consuming process takes place many times, as the principal investigator requests to see the calculations performed another way or wants to expand on a given analysis. This method of information transfer is very inefficient and leads to a long, inefficient process for research projects.

Many researchers do not even consider using the Internet to transfer information because they are comfortable with the status quo and are not aware of the improved efficiency and accuracy that is possible through the use of web technology, such as scripting languages, server-side databases, and SAS/IntrNet®. Not only can these technologies be used for greater efficiency, they can also create a more organized and secure structure in which to develop a research project. The following will discuss in more detail the potential for these new technologies to revolutionize the process of research, allowing for quicker dissemination of valuable results that could have a great impact on public health.

A NEW SOLUTION
The University of Illinois at Chicago, Center for the Advancement of Distance Education (CADE) has developed a Research Survey Composer using Microsoft Active Server pages (ASP) that allows for the creation of a data entry form that displays on a web page and stores data in a Microsoft SQL Server database. (This application can also be developed with the use of other scripting languages, such as PSP or PHP, or other databases, such as Oracle or Sybase.) The data from the database can then be imported into SAS and results can be displayed on the web with SAS/IntrNet.

SURVEY DESIGN
For a research team that has already collected data on paper, the hard copy of the survey can be entered into the Research Survey Composer to produce data entry forms on the Internet. The Research Survey Composer can be used for any research project, as it allows for the entry of any survey questions, lab results, or personal comments in the form of multiple choice, all-that-apply, or open-ended questions.

The Research Survey Composer also allows you to add, delete, and edit questions from a survey, or choose a template from a bank of frequently used questions or responses. Skip patterns that exist within the survey can also be programmed into the data entry forms with the Research Survey Composer. For example, if the survey indicates to skip several questions if the subject answered "no" to the previous question, the data entry form can be programmed to not display these questions. This can save much time and confusion for the person entering the data. It can also lead to cleaner data from the start, saving valuable time later in the process. Since the data entry forms on the web page can look identical to the hard copy of the data collection instrument, the result is a much clearer data entry tool, which leads to more accurate data input.

VALIDATION
In the past, poor data entry has been at fault for the slow progress of many research projects. Before a research team can begin analyses on a project, the data must go through several rounds of data cleaning to assure the accuracy of the inputted data set. It is common for large data sets to have many typographical errors, as the data entry process is not perfect. This lengthy process of data cleaning can be partially alleviated through the use of validation on data entry forms. With validation, ranges for certain values can be specified to make sure that entered values are logical. For example, in a study of children 18 and under, the data entry form can be specified to only allow for the entry of ages under 19. If the person entering the data accidentally enters "28" instead of "18", an error message will be returned upon submittal that indicates this entry is not valid. With this procedure, one can also ensure the entry of all data so that blanks do not get mistaken for missing data. If the data entry person overlooks a certain question, an error message will be returned to ensure the entry of all data. Validation can be specified in the Research Survey Composer to improve the accuracy of the data.

CONSISTENCY

Many researchers currently house data for large research projects on disks or CDs. If several assistants are working with the data, then several copies of the same data set may exist. To synchronize any corrections or changes made to the data, all of the parties would have to make sure each copy is updated. Researchers can alleviate this problem by housing all of the data in a database. This database can be accessible to all people on the research team, and can always house the most current copy of the data. Once in the database, the data are easily accessed for editing. The editor can search the database on a study id or a question number to access the page that needs to be edited. These changes take place prior to the creation of a SAS data set to ensure the accuracy of the data before they are imported to SAS.

SECURITY AND BACKUP

Access to the contents of the database can be controlled at many different levels of reading and writing privileges with the use of a login procedure. For example, the principal investigator may have access to confidential information about the research subjects that no one else may view. Or, a research assistant may have access to edit files within the database, while another member of the research team may only be allowed to view the data. Databases also allow for the backup of data on a regular basis to protect the information after it has been entered. This is crucial in a world with viruses and other forms of data corruption. If the server fails, there is a backup copy of the data that may have taken many years and much money to collect.

ODBC

Most importantly, SAS/ODBC® allows for the extraction of data from the database into SAS. SAS is invaluable to the research community for performing sophisticated analyses on large amounts of data. With a Transpose procedure and some minor formatting, data from the relational database can be transformed into a flat SAS file that is ready for analysis. The creation of this SAS data set can be made executable and can be scheduled to run on a regular basis, always capturing a current copy of the SAS data set. A record that is entered into the ASP page can be available in the SAS data set within minutes.

REPORTING

Once the data are captured in a permanent SAS data set, they are available for analysis. SAS macros can be programmed prior to data collection so that analyses can be run on data immediately as it becomes available. This is particularly advantageous for descriptive statistics that can be time consuming if programmed for each individual variable. Macros can write SAS programs that run a generic PROC print, PROC freq or PROC univariate of all of the variables, or can be written to execute more specific queries of the data.

To make the data available to the researchers you can create a web page form that prompts them to enter a query of the data. For example, you can download a list of variables into a drop down box, or code a text box that allows the researchers to enter a specific variable name for the query. This name corresponds with a variable name in the SAS data set and is captured as a value for the macro variable that has been programmed. Upon clicking submit, the SAS macro is called with the SAS/IntrNet broker, the macro writes a SAS program with a FILE _webout statement, and the results are sent back to your browser for display on a web page. With the use of the Output Delivery System (ODS) and procedures such as PROC Tabulate, PROC Report, and PROC Gplot, this output can be formatted in any way you desire to create finished tables, reports, or graphs on the web. The results can be displayed as html output, rich text format (RTF) output, or printer output, depending on the needs of the researcher.

ACCESS

This process gives the principal investigator and the rest of the research team the freedom to access and query their data from any place at any time. They can also produce reports of descriptive analyses at the click of a mouse without having any knowledge of the SAS programming that is involved behind the scenes. For those with SAS expertise, the data set can be made available for download from the website, if researchers decide to further customize their analyses. For this purpose, an online and searchable data dictionary can be provided to outline the variable naming techniques and formats for the statistician who will be working with the data.

The most valuable part of making a data set and descriptive statistics available on the Internet is the ability to access this information from any place around the world. This allows for ease in telecommuting, which is becoming increasingly more popular, and ease in accessing data while traveling.

FUTURE DIRECTIONS

This new research method is currently being used for a project in which data has already been collected in paper form. Therefore, the web page forms serve as a data entry tool. However, in the future, the data can be collected directly into the database when appropriate. For instance, the subjects in the study can take the survey on the web or an interviewer with a laptop can enter data directly into the data entry form during the interview, rather than collecting data on paper and entering it later. With the omission of this extra step, much time and paperwork is saved and the risk of losing hard copies of data is avoided. Also, the data are available directly after collection to be analyzed and displayed with SAS/IntrNet.

CONCLUSION

This new research method has many implications for the future of research. In a world where current knowledge is evolving so rapidly, especially in the area of human health, it is important to communicate the results of research projects quickly and accurately in order to contribute to this knowledge bank.

ACKNOWLEDGMENTS

I would like to acknowledge Colleen Monahan for all of her innovative ideas and Phillip Falkenberg for his programming expertise.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

> Kristin Rankin
> University of Illinois-Chicago
> Center for the Advancement of Distance Education
> 1603 W. Taylor M/C 923
> Chicago, IL 60612
> Work Phone: (312) 996-0179
> Fax: (312) 413-2054
> Email: krankin@uic.edu
> Web:http://www.uic.edu/sph/cade/research

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.