Paper 009-2010

Group File Processing with SAS® Macros and the Data Step

Kevin McGowan and Matt Reich, SRA International, Durham NC

ABSTRACT

Sometimes there is a need to process a group or groups of files with one program. In some cases the number of files that need processing is not known in advance, only the location of the files and the format of the directories that contain the files are known. By using the SAS macro language along with data steps and SAS functions such as dopen, dnum, dread and d close, it is possible to process multiple files easily without a lot of complex programming. These techniques are becoming more useful with the switch of data arriving in electronic format rather than on paper. By processing a group of files with one program or data step the chance of errors and processing time can be significantly reduced. This paper will provide examples to show different techniques for processing groups of files.

INTRODUCTION

Most of the time SAS users tend to think of data as be ing processed in terms of records in a file. In some situations the data is not contained in just one file or dataset – it is spread across many different files and directories. Sometimes these files all have the same format and data and sometimes they are different in format or data. This paper will describe basic techniques to deal with a group of files by using one program rather than a set of programs.

Advantages of group file processing

The combination of built-in SAS functions and the SAS macro l anguage gives SAS programmers the ability to write programs that efficient ly process multiple files with one program. The ability to process a group of files with one flexible SAS program can come in handy in several common situations. Some examples are:

· Multiple work sites sending their data files to one central location for processing and analysis

· Collection of existing data files from various sources to be processed together for analysis , including meta-type analyses

· Quality control to compare various versions of data files to make sure they are correct or consistent

· Testing the output of programs to make sure the data or results they produce are correct or in the correct format

One issue that needs to be resolved when writing a program to process group files is how flexible the program should be. On one hand, writing programs that are very flexible is a good idea because they can be used in a wide variety of cases. The downside to writing a flexible program is that it takes longer to write, test and the added complexity may lead to hidden bugs that are not discovered during testing but show up later when the program is being used. A good approach is to design a basic group file handling program as a SAS macro that can be plugged into other programs as needed. This basic program can be modified at a later date to be more specific if that is needed.

Steps to designing a program for group file processing

There are basic steps that need to be followed before starting the writing of a program to process a group of files. These steps are:

· Determine how much data/files need to be processed – is the amount known ahead of time or can it vary each time the program is run?

· Determine the location of the files – are they in the same place every time the program is run or can they change to various locations?

· Determine the file formats – are they SAS files, text files, spreadsheets or some other file? Are the files all the same format? In many cases the format of the file can be determined from the file extension.

· Is there data or information in a separate file or as part of the file name that can be used to determine how to process each file? Possible examples are the file extension or the file header.

Once the questions above are answered then programming can start. As is the case in most programming tasks, it is better to start with a very basic program to make sure the essential elements of the program work before moving on to more complex parts of the program.

Basic SAS functions for file processing

Like most programming languages, SAS has a set of functions that are designed to work with files. These functions are used to locate files, open and close files, and count the number of files in a directory. These functions are not well known by many SAS programmers but they are the key building blocks of programs that process multiple files. The most useful file handling functions are:

Dnum – returns the number of files in a directory, this is useful when you don't know how many files you need to process

Dopen – opens a directory and returns a directory identifier, this is used when you want to read all the files in a directory

Dclose – closes a directory that was opened with dopen

Fopen – opens an external file and returns a file identifier

Fclose –closes a directory, external file or directory member

Fread – reads a record from an external file into a file data buffer

Mopen –opens a file by directory id and member name

Dread – returns the name of a directory member

%sysfunc – this macro function allows the use of a standard SAS function in a macro language programming statement.

Examples of Group File Processing

The following examples show various ways that SAS functions and the SAS macro language can be combined to write programs to handle multiple files in various ways:

In each example the source code is commented the first time it is used. Subsequent examples only have comments for code that was not used in previous examples.

Example 1: A basic use of group file handling.

```
/********************** BASIC EXAMPLE

 This example shows the basic procedures for opening a directory and reading
the filenames in that directory and then processing those files.

***********************************/

options mprint;

%macro basic1;
%let filrf=mydir;
%let rc=%sysfunc(filename(filrf,"C:\testsas\"));  /* assign dir name */
%let did=%sysfunc(dopen(&filrf));   /* open directory */
%let lstname=;  /* clear filename macro var */
%let memcount=%sysfunc(dnum(&did));  /* get # files in directory */

%if &memcount > 0 %then  /* check for blank directory  */

  %do i=1 %to &memcount;  /* start loop for files */

   %let lstname=%sysfunc(dread(&did,&i));  /* get file name to process */
   filename dr "c:\testsas\&lstname";  /* assign file name */

   data a;     /* do whatever processing is needed with file */
    infile dr;
      input b $ 1;
    a=1;
```

```
    proc print; title "&lstname"; run;

  %end;

%let rc=%sysfunc(dclose(&did));  /* close directory */
%mend basic1;
```

Example 2: Group file handling with decisions on how to handle the files. This example is useful for the case where the processing of the file depends on the type of file. Once the file type is found, a decision is made on what to do with the file.

```
/************************* FILE TYPE EXAMPLE

   This example adds a feature that checks the file extensions as each file
is read.  The extension is then used to determine how to process the file.

  *********************************/
 options mprint symbolgen;
 %macro filetype;
%let filrf=mydir;
%let rc=%sysfunc(filename(filrf,"C:\testsas\"));
%let did=%sysfunc(dopen(&filrf));
%let fname=;
%let memcount=%sysfunc(dnum(&did));
%if &memcount > 0 %then

  %do i=1 %to &memcount;

   %let fname=%sysfunc(dread(&did,&i));

   %let iw=%index(&fname,.);  /*  find start of extension */
   %let exts=%substr(&fname,&iw);   /* find file extension */

   filename dr "c:\testsas\&fname.";

    %if &exts = .sas %then %do; /* check file extension */

   data a;     /* process file based on extension */
    infile dr;
      input b $ 1;
    type="SAS file";
    proc print; title "Sas file "; run;
    %end;

       %else %do;  /* the other way to process the file  */

   data a;
    infile dr;
      input b $ 1;
    type="Text file";
    proc print; title "Reg file "; run;

      %end;

    %end;
```

```
  %end;

%let rc=%sysfunc(dclose(&did));

 %mend filetype;
```

Example 3: Group file handling by using data about the file to determine how to process the file. This case uses information about the file name in order to de termine the processing of the file.

```
/********************** FILE LOCATION EXAMPLE

   This example uses the name of the file that is being processed to
determine where the output data should be stored.

 *******************************/
 options mprint symbolgen;
 %macro fileloc;
%let filrf=mydir;
%let rc=%sysfunc(filename(filrf,"C:\testsas\"));
%let did=%sysfunc(dopen(&filrf));
%let fname=;
%let memcount=%sysfunc(dnum(&did));
%if &memcount > 0 %then

  %do i=1 %to &memcount;

   %let fname=%sysfunc(dread(&did,&i));

   %let iw=%index(&fname,.);

  %let loc=%substr(&fname,1,%eval(&iw-1));  /* get first part of file name */

   filename dr "c:\testsas\&fname";

   filename outfile "c:\tsas\&loc\&fname";  /* use file name to determine
directory for output file */
   run;

    data a;
    infile dr;
      input b $ 1;

     file outfile;  /*  write output where it needs to go */
     put  b;

     proc print;
        title "Sas file "; run;

    %end;


%let rc=%sysfunc(dclose(&did));

 %mend fileloc;
```

Example 4: Another example using file information to determine how to process the file .

```
/*******************************CHOOSE PROC EXAMPLE

   This example uses the file extension to determine which proc to run on the
data.

 ************************************/

  options mprint symbolgen;
 %macro chproc;
%let filrf=mydir;
%let rc=%sysfunc(filename(filrf,"C:\testsas\"));
%let did=%sysfunc(dopen(&filrf));
%let fname=;
%let memcount=%sysfunc(dnum(&did));
%if &memcount > 0 %then

  %do i=1 %to &memcount;

   %let fname=%sysfunc(dread(&did,&i));

   %let iw=%index(&fname,.);
   %let exts=%substr(&fname,&iw);

   filename dr "c:\testsas\&fname.";

    %if &exts = .txt %then %do;  /* choose proc based on file extension */

   data a;
    infile dr;
      input b $ 1;
    type="text file";
    proc freq;
      title "Sas file &fname &exts"; run;
    %end;

       %else %do;

   data a;
    infile dr;
      input b  1;
    type="Non Text file";
    proc means;
      title "Non text file &fname &exts"; run;

      %end;

    %end;
  %end;
%let rc=%sysfunc(dclose(&did));

 %mend chproc;
```

Example 5: File processing while keeping track of the number of files processed and the names of the files processed. Sometimes you need to know how many files of each type you process and this code shows how to do that.

```
/***************************** COUNT AND STORE NAMES EXAMPLE

   This example keeps a count of each file type it process and then at the
end it prints the filenames for each type.

 *************************/

 %macro countn;

%let filrf=mydir;
%let rc=%sysfunc(filename(filrf,"C:\testsas\"));
%let did=%sysfunc(dopen(&filrf));
%let fname=;
%let memcount=%sysfunc(dnum(&did));
%if &memcount > 0 %then

 %let datcount=0;
 %let sascount=0;

  %do i=1 %to &memcount;

   %let fname=%sysfunc(dread(&did,&i));

   %let iw=%index(&fname,.);
   %let exts=%substr(&fname,&iw);

   filename dr "c:\testsas\&fname.";

    %if &exts = .dat %then %do;

   %let datcount=%eval(&datcount+1);   /* this counts files by extension */

    data tdat;   /* this stores file name in a SAS dataset */
         fname="&fname";
         output;
       run;

      proc append base=alldat data=tdat;  /* file name is appended to overall
list */

      run;

   data a;
    infile dr;
      input b $ 1;
    type="dat file";
    proc freq;
      title "Dat file &datcount"; run;
    %end;

       %else %do;
```

```
    %let sascount=%eval(&sascount+1);

   data tsas;
         fname="&fname";
         output;
       run;

      proc append base=allsas data=tsas;

      run;

   data a;
    infile dr;
      input b $  1;
    type="Non dat file";
    proc freq;
      title "Non dat file &sascount"; run;

      %end;

    %end;

  %end;

%let rc=%sysfunc(dclose(&did));

  proc print data=alldat;  /* print list of dat files  */
   title "List of all dat files"
   run;

  proc print data=allsas;  /* print list of SAS files  */
   title "List of all SAS files"
   run;


 %mend countn;
```

Example 6: Searching for a certain file type. This example is useful for cases where a directory contains many types of files but only one type is used. All the files are looked at but only the files with the extension of .dat are opened and read in with the data step.

```
/***************************** This example is looking for 1 type of file
   When that type is found (extension=.dat) the data is read in,   otherwise
the file is skipped and the program moves on to the next file

 *************************/


%macro filetype(dose,kloop,dose2,bnum);
%let filrf=mydir;

%do var=1 %to &kloop;  /* start loop over directories */

 %let rc=%sysfunc(filename(filrf,"M:\CHR\Statistics\Projects\NIEHS-
NTP_Task1\Work\HTS\C elegans\TC320_6dose\48h Growth\TP&bnum._&var.\&dose."));
```

```
 %let did=%sysfunc(dopen(&filrf));
 %let fname=;
 %let memcount=%sysfunc(dnum(&did));

 %if &memcount > 0 %then %do;
   %do i=1 %to &memcount;
    %let fname=%sysfunc(dread(&did,&i));
    %let iw=%index(&fname,.dat); /* find start of extension */
      %let exts=%substr(&fname,&iw); /* find file extension */
     filename dr "M:\CHR\Statistics\Projects\NIEHS-NTP_Task1\Work\HTS\C
elegans\TC320_6dose\48h Growth\TP&bnum._&var.\&dose.\&fname.";

    %if &exts = .dat %then %do; /* check file extension to see if it's .dat
*/

       %let ftype=t0; /* set t0 as default */

       %let i0=%index(&fname,t0); /* find t0 in the filename */
       %let i48=%index(&fname,t48); /* find t48 in the filename*/

       %if &i48 > 0 %then %do;
           %let ftype=t48;
      %end;

      data drx.a&bnum._&var._&dose2._&ftype.;
        /* process file with .dat extension, create SAS dataset */
       infile dr dsd dlm='09'x firstobs=2;
        input plate row col $ clog $ status tof ext green yellow red;
             if col='A' then col='1';
             if col='B' then col='2';
             if col='C' then col='3';
             if col='D' then col='4';
             if col='E' then col='5';
             if col='F' then col='6';
             if col='G' then col='7';
             if col='H' then col='8';

             drop clog plate;
             run;
       %end;  /* end check memcount */
     %end;  /* end file loop */
    %end; /* end dir loop */
   %let rc2=%sysfunc(dclose(&did));
 %end;
 %mend filetype ;
```

CONCLUSION

More and more data are being delivered in electronic format instead of paper format. This data needs to be processed and analyzed in an efficient mann er with minimum errors. Using SAS code that processes a group of files at the same time allows users to get data in and out of their systems quicker while reducing errors. The use of built -in SAS functions and the SAS macro language gives programmers fle xibility to write code that can handle a wide range of data in both scientific and business applications.

ACKNOWLEDGEMENTS

Thanks to Nicole Creech of SRA for editing help

CONTACT INFORMATION

Kevin McGowan
Matt Reich
SRA International
2605 Meridian Parkway
Durham, NC 27713
(919) 313-7554
kevin_mcgowan@sra.com
matt_reich@sra.com
http://www.sra.com