A Macro to Import Subject Data Saved in a Location with Separate Subfolders for each Subject

Laurie Smith, Cincinnati Children's Hospital Medical Center, Cincinnati, OH
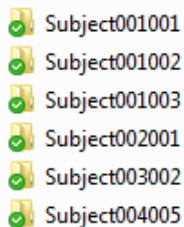
ABSTRACT

Often, in Health Sciences, subject data is saved as one file per subject. When subject data is saved in separate files, it is often difficult to import each separate file into SAS without great effort. If data is organized with the subject ID as the folder name and each subject's data in the corresponding folder, this macro will allow a programmer to read the subject ID's from the folder names, and loop though each subjects folder, importing all data within each folder.

INTRODUCTION

The methods introduced will allow a SAS® Programmer, with basic SAS® programming skills, use SAS® to import individual subject data files saved with similar files name and folder name structures.

SUBJECT DATA

Very often, data files, especially test results generated by a device, are created and saved in separate directories with similar names per subject and similarly named data files per subject. All this data usually needs to be imported into one dataset.

Display 1. Example of Subject Data Directory

IMPORT LIST OF SUBJECTS

When data is organized such that there is one folder per subject in a directory, and there is no set list of subjects, it is simple to obtain a listing of subjects to use for importing all data.

USE A FILENAME STATEMENT TO CREATE THE LIST OF SUBJECTS

A filename statement using pipe and a dir command can be used with an infile dataset to obtain a listing of the folder names from the desired directory. Using a folder naming convention that includes the subject's identifier allows a user to isolate each identifier and create a listing of all identifiers from the directory in one dataset.

```
filename dir pipe 'dir "<Source Data Location>" ';


data dirlistxx;
      infile dir lrecl=32727 truncover scanover;
      input dirtext $200.;
run;
```

```
                                 dirtext
Volume in drive C is OSDisk
Volume Serial Number is 8CCC-DBD4

Directory of C:\Users\BISE3R\OneDrive - cchmc\MWSUG2018\TestData

08/10/2018 05:49 PM   <DIR>         .
08/10/2018 05:49 PM   <DIR>         ..
08/24/2018 10:40 PM   <DIR>         Subject001001
08/24/2018 10:38 PM   <DIR>         Subject001002
08/24/2018 10:37 PM   <DIR>         Subject001003
08/24/2018 10:35 PM   <DIR>         Subject002001
08/24/2018 10:32 PM   <DIR>         Subject003002
08/24/2018 10:34 PM   <DIR>         Subject004005
       0 File(s)          0 bytes
       8 Dir(s)  281,374,351,360 bytes free
```

Display 2. dirlistxx

SUBSET SUBJECT IDS TO IMPORT DATA

Once the dataset with the folder names has been obtained, the subset of the data is created to exclude any observations that do not contain subject identifiers. A substring of each folder name isolating each identifier can be created in a separate variable. Only this new variable is retained. The count of subject identifiers is retained as a macro variable for the loop in the macro that imports the data.

```
data dirlist (drop=dirtext);
      length subjid $500.;
      set dirlistxx;
if dirtext^='' and scan(dirtext,1,'')^='Volume' and
      scan(dirtext,1,'')^='Directory' and scan(dirtext,2,'')
      ^in('File(s)','Dir(s)') and scan(dirtext,-1,'') ^in("..",".");
subjid=substr(scan(dirtext,-1,''),8);
run;


data dirlist;
      set dirlistx nobs=subjcnt;
call symput('subjcnt',put(subjcnt,best12.));
run;


proc sort data=dirlist; by subjid; run;
```

subjid
001001
001002
001003
002001
003002
004005

Display 3. dirlist

IMPORT SUBJECTS' DATA INTO ONE DATASET

USE A FILENAME STATEMENT TO CREATE A LIST OF DATA FILES

Similar to above, a listing of results data filenames can be created from each subject's folder to use for importing the data into one dataset. A filename statement using pipe and a dir command can be used with an infile dataset so create a listing of all files in each subject's folder. A dataset of this list with only filenames is then retained.

```
%macro Panel(obsnum=);
      %do i=1 %to &obsnum;
            data _null_;
                  set dirlist end=eof;
            if _n_=&i then do;
                  call symput('subjid',strip(trim(subjid)));
            end;
            run;

            %let subjData=&Source.\Subject&subjid;
            libname subjData "&subjData.";
            %let subjData=%sysfunc(quote(%qsysfunc(dequote(&subjData))));
            filename sdir&i pipe %sysfunc(quote(dir &subjData));

            data subjdirXX&i;
                  infile sdir&i lrecl=32727 truncover scanover;
                  input subjdir $200.;
            run;

            data subjdirX&i (keep=panel);
                  set subjdirXX&i;
            if subjdir^='' and scan(subjdir,1,'')^='Volume' and
                  scan(subjdir,1,'')^='Directory' and scan(subjdir,2,'')
                  ^in('File(s)','Dir(s)') and scan(subjdir,-1,'')
                  ^in('.','..');
            panel=cat(scan(subjdir,-3,'. '),'',scan(subjdir,-2,'. '));
            run;

            data subjdir&i;
                  set subjdirX&i nobs=panelcnt;
            call symput('panelcnt',put(panelcnt,best12.));
            run;
```

panel
Panel 1
Panel 2
Panel 3

Display 4. subjdir1 (data files for first subject)

IMPORT DATA INTO ONE DATASET

A do loop is used to loop through each subject id to create a dataset of each subject's files to be imported by storing the id in a macro variable (as seen above). Once the subject's id is defined in the do loop, a

second do loop is executed to create a dataset containing filenames of files to be imported in the final dataset.

```
            %do j=1 %to &panelcnt;
                data _null_;
                    set subjdir&i;
                if _n_=&j then do;
                    call symput('panel',trim(strip(panel)));
                    call
                    symput('comppanel',trim(strip(compress(panel))));
                end;
                run;

            proc import out=subject&subjid.&comppanel.x
                    datafile="&Source.\Subject&subjid.\&panel..xlsx"
                        dbms=xlsx replace;
                    sheet="&panel";
                    getnames=yes;
                    run;

                data subject&subjid.&comppanel;
                    set subject&subjid.&comppanel.x;
                subjid="&subjid";
                panel="&panel";
                run;

                data panels;
                    set panels
                        subject&subjid.&comppanel;
                if subjid^='';
                run;
            %end;
        %end;
%mend;
```

The structure of the final dataset is determined per the structure of the data in the imported files and once do loops are executed for all ids, a final dataset that contains all results from the imported files is created. In this example all files are of XLSX format and panel number and subject identifier are included in the final dataset.

	A	B	C	D	E	F
1	Test No	Result 1	Result 2	Result 3	Result 4	Result 5
2	Test 1	0.048	0.612	1.273	8.983	0.279
3	Test 2	0.237	0.721	1.463	9.683	0.168
4	Test 3	0.388	0.501	1.683	7.913	0.356

Display 5. Source data structure

```
%Panel(obsnum=&subjcnt);
```

subjid	panel	Test_No	Result_1	Result_2	Result_3	Result_4	Result_5
001001	Panel 1	Test 1	0.025	0.589	1.25	8.96	0.256
001001	Panel 1	Test 2	0.214	0.698	1.44	9.66	0.145
001001	Panel 1	Test 3	0.365	0.478	1.66	7.89	0.333
001001	Panel 2	Test 1	0.118	0.682	1.343	9.053	0.349
001001	Panel 2	Test 2	0.307	0.791	1.533	9.753	0.238
001001	Panel 2	Test 3	0.458	0.571	1.753	7.983	0.426
001001	Panel 3	Test 1	0.228	0.792	1.453	9.163	0.459
001001	Panel 3	Test 2	0.417	0.901	1.643	9.863	0.348
001001	Panel 3	Test 3	0.568	0.681	1.863	8.093	0.536
001002	Panel 1	Test 1	0.201	0.765	1.426	9.136	0.432
001002	Panel 1	Test 2	0.39	0.874	1.616	9.836	0.321
001002	Panel 1	Test 3	0.541	0.654	1.836	8.066	0.509
001002	Panel 3	Test 1	0.193	0.757	1.418	9.128	0.424
001002	Panel 3	Test 2	0.382	0.866	1.608	9.828	0.313
001002	Panel 3	Test 3	0.533	0.646	1.828	8.058	0.501
001003	Panel 1	Test 1	0.135	0.699	1.36	9.07	0.366
001003	Panel 1	Test 2	0.324	0.808	1.55	9.77	0.255
001003	Panel 1	Test 3	0.475	0.588	1.77	8	0.443
001003	Panel 2	Test 1	0.099	0.663	1.324	9.034	0.33
001003	Panel 2	Test 2	0.288	0.772	1.514	9.734	0.219
001003	Panel 2	Test 3	0.439	0.552	1.734	7.964	0.407
002001	Panel 2	Test 1	0.128	0.692	1.353	9.063	0.359
002001	Panel 2	Test 2	0.317	0.801	1.543	9.763	0.248
002001	Panel 2	Test 3	0.468	0.581	1.763	7.993	0.436
002001	Panel 3	Test 1	0.168	0.732	1.393	9.103	0.399
002001	Panel 3	Test 2	0.357	0.841	1.583	9.803	0.288
002001	Panel 3	Test 3	0.508	0.621	1.803	8.033	0.476
003002	Panel 1	Test 1	0.071	0.635	1.296	9.006	0.302
003002	Panel 1	Test 2	0.26	0.744	1.486	9.706	0.191
003002	Panel 1	Test 3	0.411	0.524	1.706	7.936	0.379
004005	Panel 3	Test 1	0.048	0.612	1.273	8.983	0.279
004005	Panel 3	Test 2	0.237	0.721	1.463	9.683	0.168
004005	Panel 3	Test 3	0.388	0.501	1.683	7.913	0.356

Display 5. Result of macro execution (All panel data from all subjects)

CONCLUSION

In some cases, it is necessary to collect subject data one file per subject per result. Storing each subject's data using the same naming convention for the files and folders in one folder per subject allows for easier import into one dataset. The macro presented above can be used to import this data.

CONTACT INFORMATION

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

Laurie A Smith
Cincinnati Children's Hospital Medical Center
(513) 803-9001
Laurie.bishop@cchmc.org
www.cincinnatichildrens.org

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.