PharmaSUG China 2015-26

A macro to re-size character variable length

Jianjun Tan, Sanofi China, Beijing, China
Eric Liao, Sanofi China, Beijing, China

ABSTRACT

FDA has a limit on the size of submitted data, i.e. one single data file no greater than 1 GB. However, there are common situations with data files greater than1GB in reality of submissions. So there are requests to resize the data files to meet the FDA's request. To facilitate and automate the resizing data files, a macro was developed as specified in FDA/PhUSE document 'Data Sizing Best Practices Recommendation', to optimize the size of dataset through managing character variable length to save wasted space in data set. This macro can automatically determine the maximum number of characters and limit the length of character variables for multiple datasets. After length limiting, the macro also will ensure that no data is truncated.

INTRODUCTION

As stated in CDISC SDTM implementation guideline(Version 3.2),'Very large transport files have become an issue for FDA to process.' The use of maximum length of 200 for character variable was mentioned as one of main contributors to this issue. 'To help rectify this situation', CDISC provided a guideline of managing the character variable length: 'Sponsors should consider the nature of the data, and apply reasonable, appropriate lengths to variables.' The FDA/PhUSE CSS Data Quality Working Group also initiated a document 'Data Sizing Best Practices Recommendation' to discuss the large transport file issue and published it on PhUSE Wiki website. In this document, the working group provided some recommendations to optimize variable lengths.

Combining the guideline from CDISC and the recommendations on PhUSE Wiki website, this macro was developed to implement the following principles:

- For character variables whose maximum length is not stated in the SDTM IG, variable length is determined by the maximum length found in data.
- For character variables whose maximum length are stated in the SDTM IG (e.g. --TESTCD), variable length is set to a pre-defined length. Such as --TESTCD variables are set to 8.
- For variables are common to multiple domains, they are allowed to have different lengths in different domains or maintain same lengths across different domains.
- When a domain is split, each variable will maintain the same length in all split datasets of the domain.

Besides the above principles, the following features are also included in this macro:

- The leading blanks in the value of character variables will be removed automatically before resizing.
- The colon (:) character can be used as wildcard character to select multiple datasets.
- When a domain has supplemental qualifier domain, no matter only the parent domain or its corresponding supplemental qualifier domain is selected or both of them are selected, the parent domain and the supplemental qualifier domain will be re-sized together.

PARAMETERS OF MACRO

The resizing macro includes both required and optional input parameters, their explanations are displayed in the following Table 1.

Parameter	**Required?**	**Explanations**	**Example**
LibIn	Yes	Specifies the directory containing the datasets to be re-sized.	%sysfunc(pathname(<*library-name*>)),U:\MACRO\TEST
LibOut	No	Specifies the directory containing the re-sized datasets. If it is null, the value of **LibIn=** will be populated.	*Similar to* ***LibIn=***

Parameter	Required?	Explanations	Example
IncDs	No	Specifies the datasets are going to be re-sized. Multiple datasets are separated by space. If it is not specified, all the datasets in **LibIn=** directory will be re-sized. Character colon(:) is used as wildcard character to select multiple datasets.	AE LB CM SUPP:
ExcDs	No	Specifies the datasets are not going to be re-sized. Multiple datasets are separated by space. Character colon (:) is used as wildcard character to select multiple datasets.	*Similar to **IncDs=***
SpltPrefix	No	Specifies the characters can be used to mark split datasets of a domain. Multiple domains are separated by space.	FA QS LB
EscapeVar	No	Specifies the variables are common to multiple domains and are going to be consistent across different domains. Multiple variables are separated by space.	STUDYID USUBJID VISIT

Table 1. Parameters for the resizing macro.

The following are some examples of calling this macro:

```
%resize(libin=U:\MACRO\TEST)
%resize(libin=%sysfunc(pathname(TESTLIB)), libout= U:\MACRO\TEST)
%resize(libin=%sysfunc(pathname(TESTLIB)), incds=EX SUPPFAAE SUPP:, excds=AD: :SUPP:,
   EscapeVar=VISIT)
```

PROGRAM FLOW IN MACRO

The program flow in this macro is consisted of the following main steps:

1. Use **LibIn=**parameter to determine which folder containing the datasets to be re-sized. Search out all SAS® data sets contained in the specified folder through SAS®TABLES dictionary table.

2. Use **IncDs=**, **ExcDs=** parameters to determine which datasets are going to be resized, and use **SpltPrefix=** parameter to determine which dataset(s) are referenced. The character Colon (:) is used as wildcard character in values of **IncDs=** and **ExcDs=** parameters.

3. If a domain has supplemental qualifier domain, make sure parent domain and supplemental qualifier domain are always selected together. If one domain is specified in **IncDs=** and **ExcDs=** at the same time, the domain will be treated as effective in **IncDs=**.

4. As re-sizing datasets and split datasets are specified, the character variables in these datasets will be retrieved from SAS®COLUMNS dictionary table. At the same time, the column information is saved into a temporary dataset for further use.

5. Use LEFT/STRIP function to remove the leading space of character variables if necessary.

6. Get maximum length of character variables and save it into the temporary dataset.

7. Considering the variables whose maximum length are stated in the SDTM IG, the different lengths of same variables in split datasets and the variables specified in **EscapeVar=** parameter, modify the temporary dataset to meet different requirements of variable length.

8. Summarize the variables which are to be re-sized.

9. If **LibOut=** parameter is specified and different from **LibIn=** parameter, copy the specified datasets into corresponding folder. Otherwise, copy the specified datasets into WORK library for further comparison.

10. Use ALTER TABLE statement and MODIFY clause of SQL procedure to modify the datasets and variables.

11. Compare the re-sized datasets with their corresponding original datasets respectively to make sure no data change after re-sizing.

KEY FEATURES

WILDCARD CHARACTER

The wildcard character is developed based on LIKE condition of SQL procedure. The use of wildcard character is similar to the % character in LIKE condition. Once **IncDs=** or **ExcDs=** parameter is input, the following template will be used to match. The characters without wildcard character will be matched in “*<Dataset-name(s)>*” part, and the characters with wildcard character will be matched in “%*<Dataset-Name-Part>*%” part.

```
Proc sql noprint;
    create table <temporary-dataset> as
    select distinct
           memname, memtype
    from dictionary.TABLES
    where strip(upcase(libname))=%upcase("<Specified-library>")
           and ^(nobs le 0 or nvar le 0) and index(memname, '_') le 0
           and (strip(upcase(memname)) in ("<Dataset-name(s)>")
                 or (strip(upcase(memname)) like '%<Dataset-Name-Part>%')
                  )
    order by memname;
quit;
```

For example, the codes based on input**IncDs=**EX SUPPFAAERELREC SUPP: will be resolved as below:

```
proc sql noprint;
    create table <temporary-dataset> as
    select distinct
           memname, memtype
    from dictionary.TABLES
    where strip(upcase(libname))=%upcase("<Specified-library>")
           and ^(nobs le 0 or nvar le 0) and index(memname, '_') le 0
           and (strip(upcase(memname)) in ("EX","SUPPFAAE","RELREC")
                 or (strip(upcase(memname)) like 'SUPP%')
                  )
    order by memname;
quit;
```

PARENT AND SUPPLEMENTAL DOMAIN

According to SDTM IG, the name of supplemental qualifier domain starts with SUPP characters. The observations in SUPP-- datasets are one-to-one match to parent records. Therefore, in this resizing macro, once a domain is input in **IncDs=** or **ExcDs=** parameter, it will be checked if it is a supplemental qualifier domain, or if it has corresponding SUPP-- dataset. As long as it is contained, its associated domain will be contained too.

The variables of SUPP-- datasets are common to all supplemental qualifier domains. These variables can be re-sized separately in different SUPP-- datasets; their length can also be maintained across different data sets if specify SUPP in **SpltPrefix=** parameter.

USING ALTER TABLE STATEMENT

By using ALTER TABLE statement of SQL procedure, we can modify the length of character variables directly and maintain other properties of the dataset consistent, without worrying some operations on the dataset, such as losing the labels of dataset or variables, changing the orders of variables, or adding any unnecessary formats for numeric variables, etc...

```
proc sql noprint;
    alter table <dataset>
           modify <character-variable-1> CHAR(<maximum-length-of-variable-1>)
                 ,<character-variable-2> CHAR(<maximum-length-of-variable-2>)
                 , etc..;
quit;
```

CONCLUSION

The %RESIZE macro is developed to re-size the length of character variables in multiple datasets. The selection of multiple datasets by wildcard character is one of key features of this macro. Another key feature of this macro is selecting the parent domain accompany with its corresponding supplemental qualifier domain. These two features rely on the capability of dealing with characters by SAS®. With traditional SAS® techniques (DATA step, SQL procedure and MACRO facility), batch processing with characters and multiple datasets is feasible and easily achieved.

REFERENCES

[1] CDISC Submission Data Standards Team. “CDISC SDTM Implementation Guide (Version 3.2)”. 26 November 2013.

[2] FDA/PhUSE CSS Data Quality Working Group. “Data Sizing Best Practices Recommendation”. PhUSE Wiki. Last modified on 10 March 2014. Available at http://www.phusewiki.org/wiki/index.php?title=Data_Sizing_Best_Practices_Recommendation.

[3] Sandra VanPelt Nguyen. “Reducing Variable Lengths for Submission Dataset Size Reduction”. PharmaSUG 2014, Paper CC37. Available at http://www.pharmasug.org/proceedings/2014/CC/PharmaSUG-2014-CC37.pdf

ACKNOWLEDGMENTS

Thank you to You Li. This paper wouldn’t be written without his encouragement. Thanks to Haoyu Li for providing valuable feedback on using this macro.

CONTACT INFORMATION

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

Name: Jianjun Tan
Enterprise: Sanofi China, B&P
Address: 5th Floor. HP Building, No. 112Jiangguo Road, Chaoyang District
City, StateZIP: Beijing, China
Work Phone: 010-65634787
E-mail:jianjun.tan@sanofi.com

Name: Eric Liao
Enterprise: Sanofi China, B&P
Address: 5th Floor. HP Building, No.112Jiangguo Road, Chaoyang District
City,StateZIP: Beijing, China
Work Phone: 010-65634894
E-mail:eric.liao@sanofi.com