Utility Macros Used in the Process for Database Translations

Yuguang Zhao, sanofi-aventis research, Malvern, PA 19355
Jingchun Yu, sanofi-aventis research, Malvern, PA 19355

ABSTRACT

One challenge facing global pharmaceutical companies is that clinical trial database sometimes contains non-English text and needs to be translated into English when the database is submitted to the FDA for regulatory submissions. Two utility macros are developed to facilitate the process. One macro will search all the datasets in the database and perform PROC FREQ on all character variables and set the output datasets from the PROC FREQ procedure together to build a master list of all unique character strings for translations. After the master list is translated, a second macro will take the translated master list and merge the translated text back into the database by looping through all the character variables in all the datasets. The translated master list could also be used in the process for actual case report form (CRF) translations to ensure consistency between the CRF and database.

INTRODUCTION

Global pharmaceutical companies conduct clinical trials in different countries due to different reasons such as to fulfill local regulatory requirements and to expedite patient accrual. When clinical trials are conducted in countries where the native language is not English, the clinical trial data might be collected in other languages other than English. Sometimes the case report forms might be designed in English and the intention is to collected the data in English, however, non-English texts are inevitably recorded in the case report forms due to convenience. As a result, pharmaceutical companies end up with case report forms and/or databases containing non-English text. If such case report forms and database are used for regulatory submissions to the FDA, it is mandatory to have non-English translated into English to facilitate the review process.

In order to translate the database and facilitate the translation of case report forms and maintain consistency between case report forms and the database for translation, two utility macros are developed. One macro will search all the datasets in the database and perform PROC FREQ on all character variables and set the output datasets from the PROC FREQ procedure together to build a master list of all unique character strings. Then the master list is sent for translation. After the master list is translated, a second macro will take the translated master list and merge the translated texts back into the database by looping through all the character variables in all the datasets.

PROGRAM FLOW

FIRST MACRO FREQOUT.SAS

Macro FREQOUT.SAS is designed to perform PROC FREQ on all character variables in all datasets in a specified directory and then set the PROC FREQ output datasets together to build a master list of unique strings.

MACRO CALL:

%FREQOUT(LIB=)

One parameter, LIB, needs to be defined before the macro call. LIB is where the datasets of interests are stored.

The macro, first, creates a list containing all dataset names from LIB area and converts each dataset name into a macro variable, DATA1, DATA2, ..., and DATAn where n is the total number of datasets. The total number of datasets is also put into a macro variable TOTAL in order to control a do loop.

The macro, secondly, uses a do loop from 1 to TOTAL to go through dataset 1 to dataset n by feeding the processes with macro variables, DATA1, DATA2, ..., and DATAn. Within the do loop, first, it creates a list containing all character variables and converts the variable names into macro variables, VAR1, VAR2, ..., and VARM where m is the total number of character variables in the dataset. The total number of character variables is also put into a macro variable VARTOT. Within the do loop, there are two macros: %GETFREQ and %SETDATA.
%GETFREQ:

This macro uses a do loop from 1 to VARTOT to go through variable 1 to variable m by feeding procedure with macro VAR1, VAR2, ..., VARm. PROC FREQ with out option produces the frequencies of character variables by taking off the formats and output the result to datasets, FREQ1, FREQ2, ..., and FREQm where FREQ1 is the frequency output for the first variable, FREQ2 is for the second variable and so on for all the character variables in the data set.

%SETDATA:

%SETDATA sets FREQ1 to FREQm together into one data set FINAL&TOTAL with a do loop.

After %SETDATA macro, FINAL&TOTAL is appended to the dataset FINAL and then the dataset FINAL can be used to create the master list of unique character strings used for translations.

SECOND MACRO FREQIN.SAS

Macro FREQIN.SAS is designed to merge the translated text back to the character variables by using the master list of unique strings.

MACRO CALL:

%FREQIN(LIB=)

FREQIN has also one parameter, LIB, to be defined before the macro call. LIB is where the datasets of interests are stored.

Similarly to FREQOUT, FREQIN macro, first, creates a list containing all dataset names from LIB area and converts each dataset name into a macro variable, DATA1, DATA2, ..., and DATAn where n is the total number of datasets. The total number of datasets are also put into a macro variable TOTAL in order to control a do loop.

Then the macro uses a do loop from 1 to TOTAL to go through dataset 1 to dataset n by feeding the processes with macro variables, DATA1, DATA2, ..., and DATAn. Within the do loop, first, it creates a list containing all character variables and converts the variable names into macro variables, VAR1, VAR2, ..., and VARm where m is the total number of character variables in the dataset. The total number of character variables is also put into a macro variable VARTOT. Within the do loop, a macro called SortMergeData will sort the datasets by each variable and match merge the translated text to the datasets by using the master list of unique strings as the key.

%SORTMERGEDATA:

The macro uses a do loop, first, sort the dataset by variables, VAR1, VAR2, ..., and VARm where m is the total number of character variables in the dataset. Then the dataset is merged with the master list to get the translated text onto the dataset. Finally, the translated text is incorporated into the dataset.

DISCUSSIONS

The use of the master list of unique strings is two fold. One is that the master list of unique strings could be put into Excel in one column and then sent for translations. The translated text will be put into another column in the Excel. Then the Excel is converted to SAS dataset to be used for merging back to the database. Secondly, the master list could be used for translating non-English text on the CRFs. The master list of unique strings reflects what has been entered in the database from the CRFs, therefore, using the master list for translating the CRFs will ensure the translated text for CRFs is consistent with the translation for the database. In addition, any unique string is only translated once and the string may appear on different CRF pages. The translator could cut and paste the unique string to increase efficiency and keep consistency among different CRF pages.

CONCLUSION

The macros described in this paper are efficient utility tool in the process for translating non-English text into English. One macro will search all the datasets in the database and perform PROC FREQ on all character variables and set the output datasets from the PROC FREQ procedure together to build a master list of all unique character strings for translations. After the master list is translated, a second macro will take the translated master list and merge the translated text back into the database by looping through all the character variables in all the datasets.
ACKNOWLEDGEMENTS

The authors would like to thank our colleagues at sanofi-aventis who have contributed ideas in the past.

CONTACTING THE AUTHORS

Yuguang Zhao
Sr Manager, Statistical Systems
sanofi-aventis research
9 Great Valley Parkway
Malvern, PA 19355
yuguang.zhao@sanofi-aventis.com

Jingchun Yu
Principal Statistical Systems Analyst
sanofi-aventis research
9 Great Valley Parkway
Malvern, PA 19355
jingchun.yu@sanofi-aventis.com

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.