PharmaSUG China 2021 - Paper DS-33

Consideration and Solutions of Garbled Codes in Multiple Languages Projects

Yi Lu, Shijia Wang, Tigermed Co., Ltd.

ABSTRACT

In studies with multiple languages or those need to convert from one language to another language, we often meet some garbled code issues when dealing with the data and documents. Garbled code is a widespread problem in the cross-device transfer of a variety of files, including data sets, codes, text file (TXT), excel file, flat file (CSV), xml file, etc. Therefore, we need to deeply understand the nature of garbled code problems and data encoding, to avoid this problem before and during the project implementation.

This paper starts from the project level and systematically summarize the possible garbled problems in CRT Package when conduct submission in China. We compiled methods and precautions for handling different file encodings to ensure the content is presented clearly in study data.

INTRODUCTION

In previous MRCT (Multi-regional Clinical Trial) projects, almost all work was done in English. However, with the release of NMPA (National Medical Products Administration) Guideline on the Submission of Clinical Trial Data, data localization has gradually become a part of data quality just like standardization. The guidelines clearly require that the key contents of the relevant application materials, including but not limited to the data set label, variable label, adverse event name, concomitant medication name, medical history name, and the key indicator value or code list in the data reviewers guide, should be in Chinese. Submission in multiple languages leads to a problem that is not particularly noticeable, but can easily occur: the garbled code problem.

Our work requires the use and processing of various files in different formats. Have you ever had such a question? Why a file which reads normal on the computer will turn to unreadable garbled characters on another device? In most cases, this is the fault of coding. When preparing CRT package, you need to fully consider the compatibility of different devices where a file is displayed.

This paper will let you know the principle of coding which allows you to understand the root cause of garbled characters. At the same time, it will give you a clear view of the garbled problems in different formats of documents which you may encounter in the projects and help you avoid them.

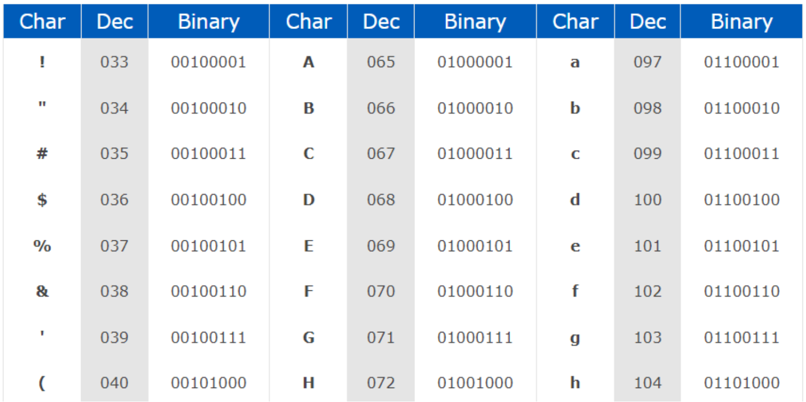
THE PRINCIPLE OF GARBLED

To solve the garbled problem, we must first understand what encoding is.

WHAT ENCODING IS

Coding is a computer term. It refers to the use of codes to represent groups of data, making it information that can be processed and analyzed by a computer. Character data is stored as a series of bytes in the computer. A byte is composed of 8 bits (binary digits). A bit which contains the value 1 or 0 is the smallest unit of data in the computer. Generally, bytes can be organized into different sequences to represent numbers between 0 and 255.

The data encoding will contain a character set, which is a list of characters that can be represented by the encoding. In order to connect byte-level data to the required characters, the coded character set maps the number represented by a byte to its corresponding character.



Char	Dec	Binary	Char	Dec	Binary	Char	Dec	Binary
!	033	00100001	A	065	01000001	a	097	01100001
"	034	00100010	B	066	01000010	b	098	01100010
#	035	00100011	C	067	01000011	c	099	01100011
$	036	00100100	D	068	01000100	d	100	01100100
%	037	00100101	E	069	01000101	e	101	01100101
&	038	00100110	F	070	01000110	f	102	01100110
'	039	00100111	G	071	01000111	g	103	01100111
(	040	00101000	H	072	01001000	h	104	01101000

Display 1. Part of the ASCII code table

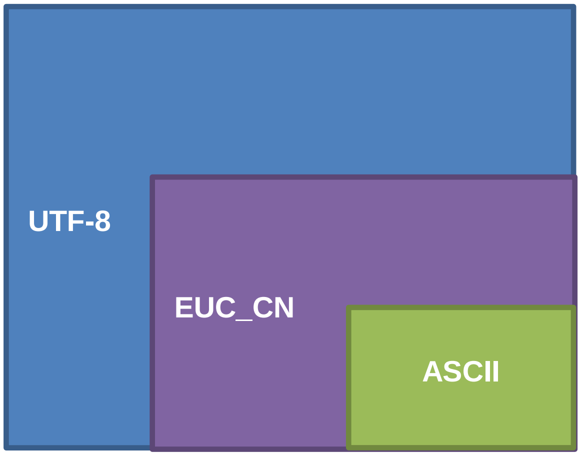
A different point for the type of data encoding is the number of bytes used to store characters. The first Western encodings, ASCII/Latin1, were all single-byte encodings, but when computers were internationalized, more and more text symbols could not be included in the single-byte coded character set. So there are also double-byte character sets (DBCS) and multi-byte character sets (MBCS). A tip is that in all coding tables, the codes from 0 to 127 are the same (ASCII). So you will find that no matter how the garbled is generated, English words can be displayed normally.

THE FORMATION OF GARBLED

There are two causes of the garbled:

1. **The code table does not have this character**

The most common cause of garbled characters is the use of mismatched code tables for decoding. For example, if you try to read Chinese with ASCII encoding, a bunch of garbled characters will appear. This is because the ASCII code table does not contain Chinese characters at all. In other words, it's similar to trying to look up Chinese words in an English dictionary.



Display 2. Comparison of the number of characters covered by different code tables

Code tables in many countries are not comprehensive enough. To solve the problem, countries around the world have cooperated to develop a language-‘Unicode’. Unicode sets a uniform and unique binary code for each character in each language. So theoretically, you don't have to worry about the aforesaid cause of garbled if using Unicode encoding.

2. **The code is truncated from the original length**

However, in real projects, using Unicode may also cause garbled characters. This is due to the "multi-byte" feature of Unicode. Multi-byte means that a character needs a longer length. If it is read by the original length, it will be truncated, which will cause garbled codes.

Here is a Chinese character “我”. You can see that utf-8 encoding requires three bytes for encoding, while EUC_CN only requires two bytes. When you convert euc_cn to utf-8 encoding without changing the length, the binary characters will be truncated to ‘11100110 10001000’ and can't be displayed normally.

Chinese character	我
UTF-8 binary	11100110 10001000 10010001
EUC_CN binary	11001110 11010010

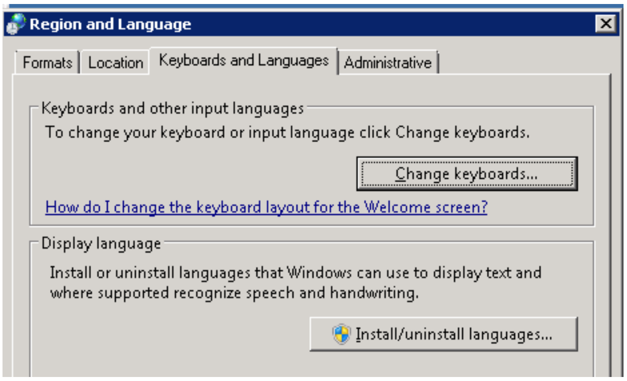
Table 1. Different binary codes of "我"

ENVIRONMENT AND CODING CHOICE

FDA and NMPA have no mandatory requirements for encoding. FDA requires submission in English, and the characters are restricted to ASCII. For projects submitted in both Chinese and English, we recommend choosing ASCII for the FDA submission. Garbled error reports can help us replace or delete non-ASCII special characters.

Based on NMPA's requirement to submit application materials related to clinical trial data in Chinese, you’d better build the project in a Chinese computer environment that can avoid a lot of garbled problems in the first place. For each file in the project, we need to select a code that can read Chinese. The most common Chinese encodings are EUC_CN (an encoding method of GB2312) and UTF-8.

If your computer users English system, please make sure you have installed the Chinese language pack on your computer. The specific operation path is ‘Control Panel→Change Display Language→Install languages’. This can help the software in your computer to read and display Chinese.


Region and Language
Formats
Location
Keyboards and Languages
Administrative
Keyboards and other input languages
Change keyboards...
How do I change the keyboard layout for the Welcome screen?
Display language
Install/uninstall languages...

To change your keyboard or input language click Change keyboards.

Install or uninstall languages that Windows can use to display text and where supported recognize speech and handwriting.

Display 3. Interface of language installment

The following table lists the advantages and disadvantages of these two common Chinese encodings. We recommend using utf-8 encoding to generate all documents like Data Reviewer’s Guide, which can minimize the system’s restrictions on file readability. When you generate a data set, both encodings are optional.

Encoding	**Advantage**	**Disadvantage**
UTF-8	More tolerant of collecting raw values	1. Chinese characters usually require a length of 3 bytes in UTF-8 encoding, and the same length can hold less content. For data sets, you will have to pay more attention to whether the limitation of a label 40

		length and a variable 200 length in the data set will be truncated. 2. You may ignore the existence of non-ASCII characters when processing English data sets.
EUC-CN	Double-byte encoding, a Chinese character only needs two bytes in length	If there are special characters in the data set that cannot be successfully decoded, UTF8 is still required for special processing

Table 2. The advantages and disadvantages of the two Chinese encodings

CONSIDERATIONS AND METHODS FOR VARIOUS FILE TRANSCODING

SAS DATA SET

For SAS data set, you will find that there are several methods for data migration on the SAS help documentation: Cross-environment data access (CEDA), Character Variable Padding (CVP) engine, and the transcoding hands-on macro.

Cross-Environment Data Access (CEDA)

Cross-Environment Data Access (CEDA) is a Base SAS feature. If not set, it’s an automatic tool to complete the encoding conversion.

When SAS directly reads a data set and put the following note without error, it proves that CEDA has taken effect and the data has been transcoded correctly.

```
NOTE: Data file TEST.DATA is in a format that is native to another host, or
the file encoding does not match the session encoding. Cross Environment Data
Access will be used, which might require additional CPU resources and might
reduce performance.
```

Output 1. CEDA Log when reading a different encoding data set

In another case, you may encounter an error or warning message. It can mean there is not enough space in one or more character columns in the data set's observation buffer to convert the data to the session encoding or that the characters can’t be represented in the session encoding.

```
ERROR: Some character data was lost during transcoding in the data set TEST.
Either the data contains characters that are not representable in the new
encoding or truncation occurred during transcoding.
```

Output 2. CEDA Error when reading a different encoding data set

If you meet that error, it means you should try the other two methods:

Character Variable Padding (CVP) engine

Chinese characters need more space in UTF-8 encoding. By using the LIBNAME statement to specify the CVP engine explicitly, the default character expansion is 1.5 times the length of the character variable. To specify a different amount of expansion, use the CVPBYTES= or CVPMULTIPLIER= option. CVPBYTES acts directly on the byte, increase to the original length. CVPMULTIPLIER sets the multiple of the length of the zoom variable

Example:

libname <your libname here> cvp <your path here> **cvpbytes**=xx;

libname <your libname here> cvp <your path here> **cvpmultiplier**=x;

Transcoding hands-on macro

SAS also provides a macro (%COPY_TO_NEW_ENCODING) to create a new version of a data set with a specified encoding. This macro will verify the maximum length needed for each character variable so that the new data set will have sufficient length to avoid any truncation (Richard D. Langston,2018).

The macro has 3 easy-to-understand parameters: the "from" data set name, the "to" data set name, and the encoding to use for the "to" data set:

%macro copy_to_new_encoding(from_dsname,to_dsname,new_encoding);

You only need to put the original data set name in [from_dsname], the newly generated data set name in [to_dsname], and the new code in [new_encoding] and then run the macro.

The difference between this macro and CVP operation is that you need to use SAS with the same encoding as the data set when using this macro. And there is no such restriction when using CVP. But using CVP will enlarge the length of the variable indiscriminately.

For the macro detail information, you could refer to the URL of the SAS document in the references.

Don't use "Encoding=ANY" option unless absolutely necessary

Per low-quality raw data, the project team is usually accustomed to using ‘Encoding=any’ to avoid CEDA runtime errors. This is a dangerous operation. Although the error in the log is avoided, in most cases, your raw data may have been changed. As we all known, all code tables were include ASCII (occupies 0-127 positions). But after the position of 127, the text represented by the same code is different. The correct approach is to find the special character that causes the error, and then decide whether to replace or keep it.

Try K-Function in multi-byte environment

The K function is a function provided by SAS to handle double-byte/multi-byte characters. Commonly used text processing K functions: klength, ksubstr, ktruncate, kindexc, kfind, kfindc, ktrim and etc.

Example: Using KTRUNCATE to process multi-language mixed strings

When we process raw data, we often encounter situations where the length of the Standardized Medication Name (CM.CMDECOD) exceeds 200. If it is only expressed in English, we can use 200 to intercept and make multiple variables. But how to deal with the mixed Chinese and English situation? Here is a way for you to use ‘KTRUNCATE+Length’ to judge.

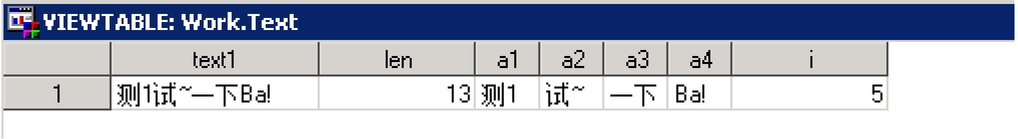
In the introduction of SAS Documentation，KTRUNCATE(argument, number, length):‘Truncates a string to a specified length in byte unit without breaking multibyte characters.’ It will cut from left to right according to the length you specify [length] from the specified position [number]. The best part is that we don’t have to worry about garbled codes.

The following is a small test. If we want to intercept a long string with a maximum length of 4, we can do so as follows. From the results of the data set, it is completely correct.

```
data a;
  text1='测1试~一下Ba!';output;
run;

data text;
      set a;
      len=0;
      length a1-a4 $4.;
      array a[*] a1-a4;
      do i = 1 to 4;
            if i=1 then a[i]=KTRUNCATE(text1,4);
            else a[i]=KTRUNCATE(text1,len+1,4);
            len=len+length(a[i]);
```

```
        end;
run;
```



	text1	len	a1	a2	a3	a4	i
1	测1试~一下Ba!	13	测1	试~	一下	Ba!	5

Output 3. A small test about %KTRUNCATE

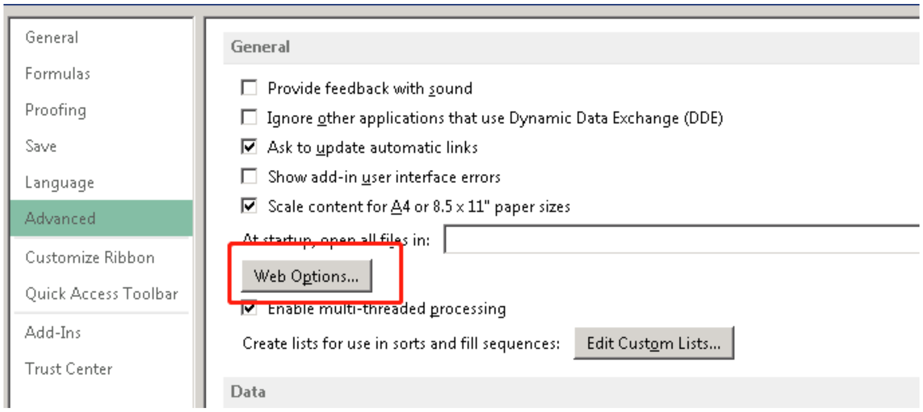
TRANSCODING FOR OTHER SUBMISSION FILE TYPES

xlsx\rtf

For xlsx and rtf files, there are two ways to modify the encoding.

The first method:

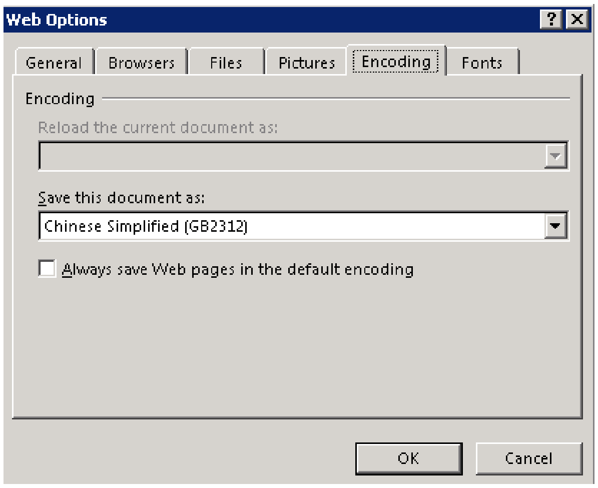
1. Open the ‘Options’ panel and select ‘Advanced’ tabs. You could see ‘Web Options…’ button.



- ☐ Provide feedback with sound
- ☐ Ignore other applications that use Dynamic Data Exchange (DDE)
- ☑ Ask to update automatic links
- ☐ Show add-in user interface errors
- ☑ Scale content for A4 or 8.5 x 11" paper sizes
- At startup, open all files in:
- ☑ Enable multi-threaded processing
- Create lists for use in sorts and fill sequences:

Display 4. ‘Options’ panel

2. Click the button and choose the encoding you want.



Encoding

Reload the current document as:

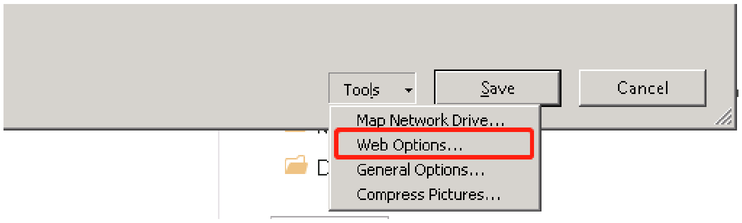
Save this document as: Chinese Simplified (GB2312)

☐ Always save Web pages in the default encoding

Display 5. ‘Web Options’ Panel

The second method:

Click ‘Save as’ button and choose ‘Web Options’ under the bottom of the panel. Follow-up operations are the same as above.

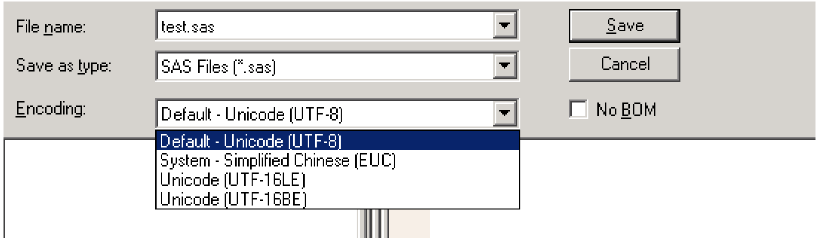


Display 6. ‘Save as’ Panel

Besides, the Microsoft Excel is a relatively tolerant software. Under most circumstances, when there is a Chinese language pack in your system, and you have loaded it in excel, then Chinese basically won't cause garbled problems.

SAS CODE and CSV\TXT

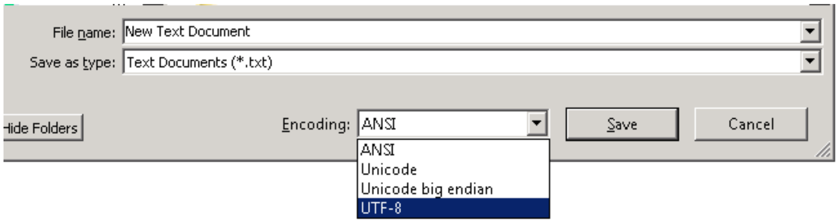
When you directly use SAS to write a program, the code file will be saved in the same encoding as the SAS session encoding. You can also use any plain text editor to write SAS code, such as Text Document, NotePad++ or PSPad. It should be noted that the encoding of the SAS code file needs to be consistent with the SAS session encoding, so that the code can be displayed normally. So if you try to switch SAS encoding in the middle of the project, remember to transcode the SAS code file. You can directly save as other encoding formats in SAS, or use the transcoding method of a text editor (mentioned below).


File name: test.sas
Save as type: SAS Files (*.sas)
Encoding: Default - Unicode (UTF-8)

Display 7. SAS Code ‘Save As’ Panel

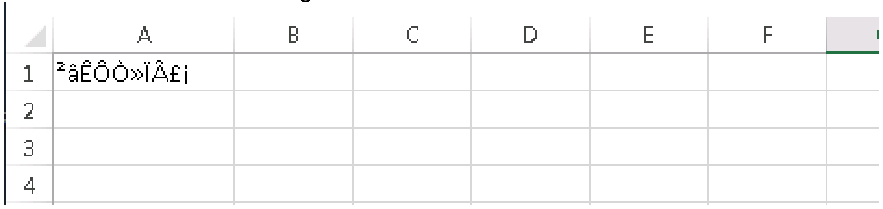
When you create a document in CSV or TXT format, the default encoding is ‘ANSI’ which is similar to ASCII. The magic of this encoding is that it is not a specific kind of encoding, and the file will use the computer's default environment to select the appropriate encoding. This also explains why the TXT file saved in the Chinese system becomes garbled when opened in the English system.

So when you use TXT or CSV format to save information, please remember to save as an utf-8 encoded file. The method of modifying the CSV encoding is the same as modifying the TXT file. First, you need to open the CSV file with Text Document and then re-save it with the correct encoding.


File name: New Text Document
Save as type: Text Documents (*.txt)
Encoding: ANSI

Display 8. Using ‘Save As’ and choose the encoding for TXT

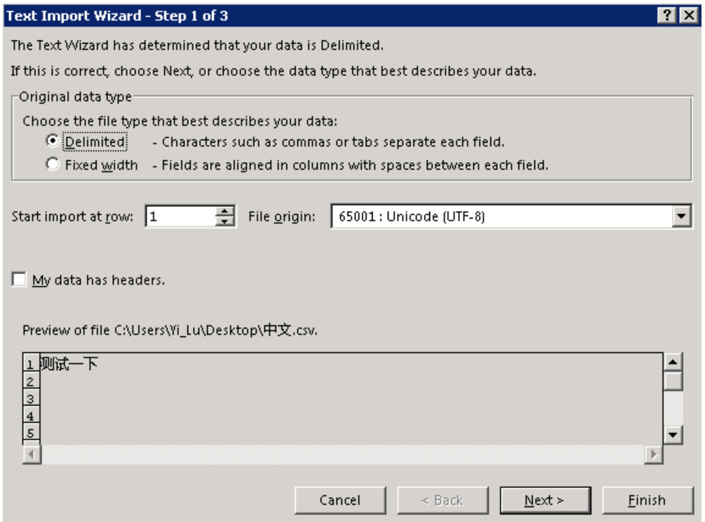
You can also use excel to read CSV files with different encodings. But even if it is an Utf-8 encoded CSV, it will be displayed as garbled characters when opened directly with excel. This is because excel will use the default ANSI encoding to decode the CSV files.



	A	B	C	D	E	F
1	²âÊÔÒ»ÏÂ£¡					
2						
3						
4						

Display 9. Garbled characters in CSV

In this case, please do not save the garbled file, close it, and reopen an empty excel. Select the [DATA] tab on the operation panel and then choose [From the Text].



Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1 File origin: 65001 : Unicode (UTF-8)

□ My data has headers.

Preview of file C:\Users\Yi_Lu\Desktop\中文.csv.

1 测试一下

Display 10. How to use Excel to read CSV

Define.xml

XML is a markup language used to mark electronic documents to make them structured. The XML file format is a plain text format.

If you find some garbled symbols in define.xml, you can use Notepad/ Text Document to open define.xml. The encoding setting is on the first line of the code. Change the encoding to the safest encoding---'UTF-8'.

```
<?xml version="1.0" encoding="UTF-8"?>
```

Output 4. The first row in define.xml

If the code is still garbled after the modification, perhaps the file encoding of the xml itself is incorrect. Please use the above method to modify the TXT file and save the xml file as utf-8 encoding.

CONCLUSION

Garbled codes are very common and easy to deal with in our daily work. The important thing is that you can understand the encoding and find out the cause of the garbled code. Here are some useful advices to avoid garbled code. Firstly, when you need to complete the submission in China, a Chinese computer environment is a good way to help you complete your work. Secondly, remember to save your documents as utf-8 encoding format. Thirdly, when you deal with data set, try not to use 'encoding=any. Please remember to convert the variable length if necessary and use the k function for multi-byte encoding.

REFERENCES

- Michael Stackhouse, Lavanya Pogula, 2018, UTF What? A Guide for Handling SAS Transcoding Errors with UTF-8 Encoded Data, https://www.pharmasug.org/proceedings/2018/BB/PharmaSUG-2018-BB08.pdf
- SAS Documentation for SAS 9.4. “SAS® 9.4 National Language Support (NLS): Reference Guide, Fifth Edition”. Available at https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/nlsref/titlepage.htm
- SAS Documentation for SAS 9.4 KTRUNCATE Function. Available at

  https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/nlsref/p0kslx8j9r3bw8n1niwz6h3k1mod.htm#p1404m82vyb43nn1bam6jeoa80mj
- Richard D. Langston, SAS Institute Inc. 2018, A Macro for Ensuring Data Integrity When Converting SAS® Data Sets, Available at

  https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1778-2018.pdf

CONTACT INFORMATION

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

Yi Lu
Tigermed Co., Ltd.
yi.lu@tigermedgrp.com

Shijia Wang
Tigermed Co., Ltd.
shijia.wang@tigermedgrp.com