

SAS[®] GLOBAL FORUM 2020



USERS PROGRAM

See All of Your Characters in SAS® Studio

Elizabeth Bales, SAS

Elizabeth Bales is the Senior Manager of the Globalization R&D team at SAS Headquarters. She joined the Internationalization team in 1997.

Exploring the Great Unknown

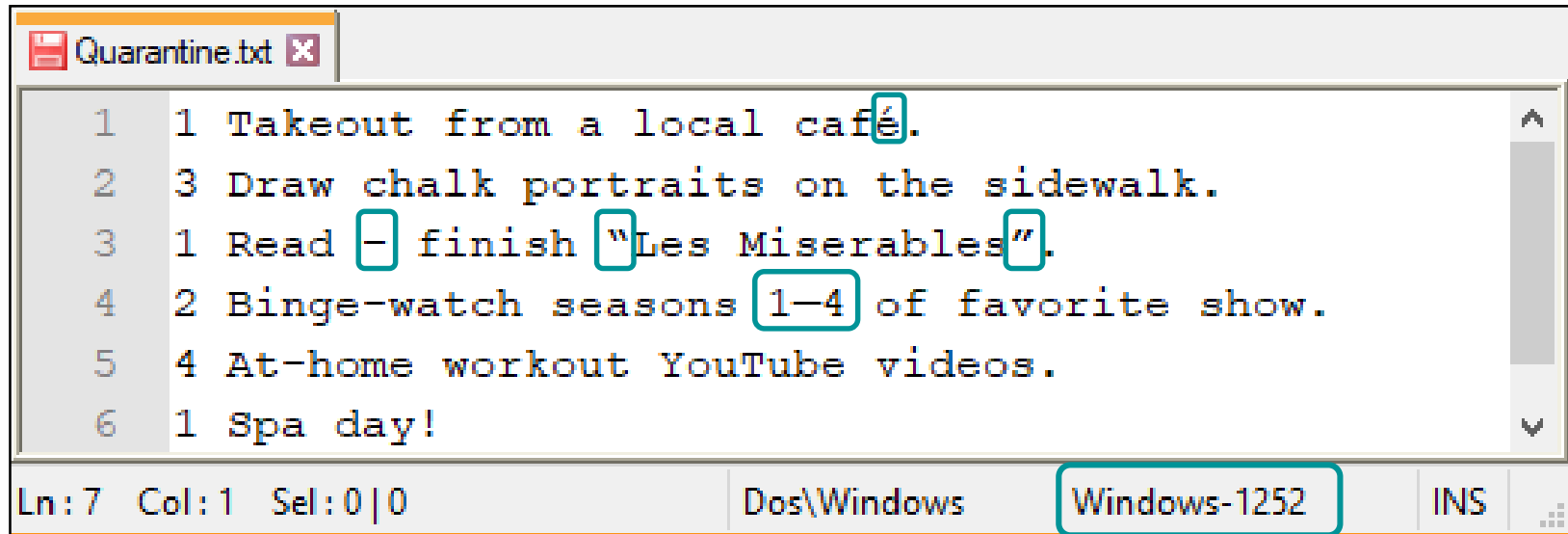


Manage Unexpected Characters

- Read data files
- Open SAS programs
- Convert to ASCII

Read Data Files

WLATIN1 data file

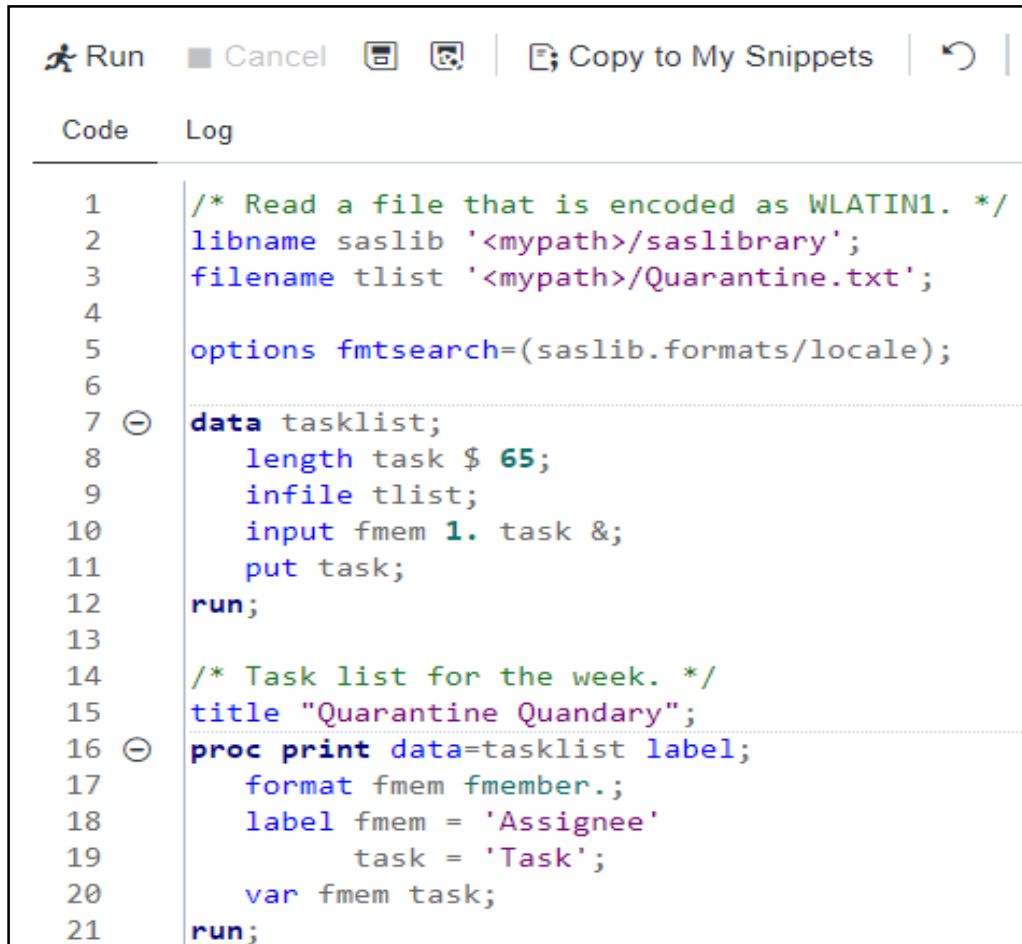


```
Quarantine.txt x
1 1 Takeout from a local café.
2 3 Draw chalk portraits on the sidewalk.
3 1 Read - finish "Les Misérables".
4 2 Binge-watch seasons 1-4 of favorite show.
5 4 At-home workout YouTube videos.
6 1 Spa day!
Ln: 7 Col: 1 Sel: 0|0 Dos\Windows Windows-1252 INS
```

WLATIN1 Character set

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
2_	(space)	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3_	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4_	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5_	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6_	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7_	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL
8_	€		,	f	„	...	†	‡	^	%o	Š	<	Œ		Ž	
9_		‘	’	“	”	•	—	~	™	š	>	œ		ž	Ÿ	
A_	NBSP	ı	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	(shy)	®	¯
B_	°	±	²	³	´	µ	¶	·	,	¹	º	»	¼	½	¾	¿
C_	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D_	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E_	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F_	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Represented as
two or three
bytes in UTF-8



The screenshot shows the SAS IDE interface. At the top, there is a toolbar with icons for 'Run' (a person running), 'Cancel' (a square), and 'Copy to My Snippets' (a document with a plus sign). Below the toolbar, there are two tabs: 'Code' and 'Log'. The 'Code' tab is active, displaying the following SAS code:

```
1  /* Read a file that is encoded as WLATIN1. */
2  libname saslib '<mypath>/saslibrary';
3  filename tlist '<mypath>/Quarantine.txt';
4
5  options fmtsearch=(saslib.formats/locale);
6
7  data tasklist;
8      length task $ 65;
9      infile tlist;
10     input fmem 1. task &;
11     put task;
12 run;
13
14 /* Task list for the week. */
15 title "Quarantine Quandary";
16 proc print data=tasklist label;
17     format fmem fmember.;
18     label fmem = 'Assignee'
19           task = 'Task';
20     var fmem task;
21 run;
```

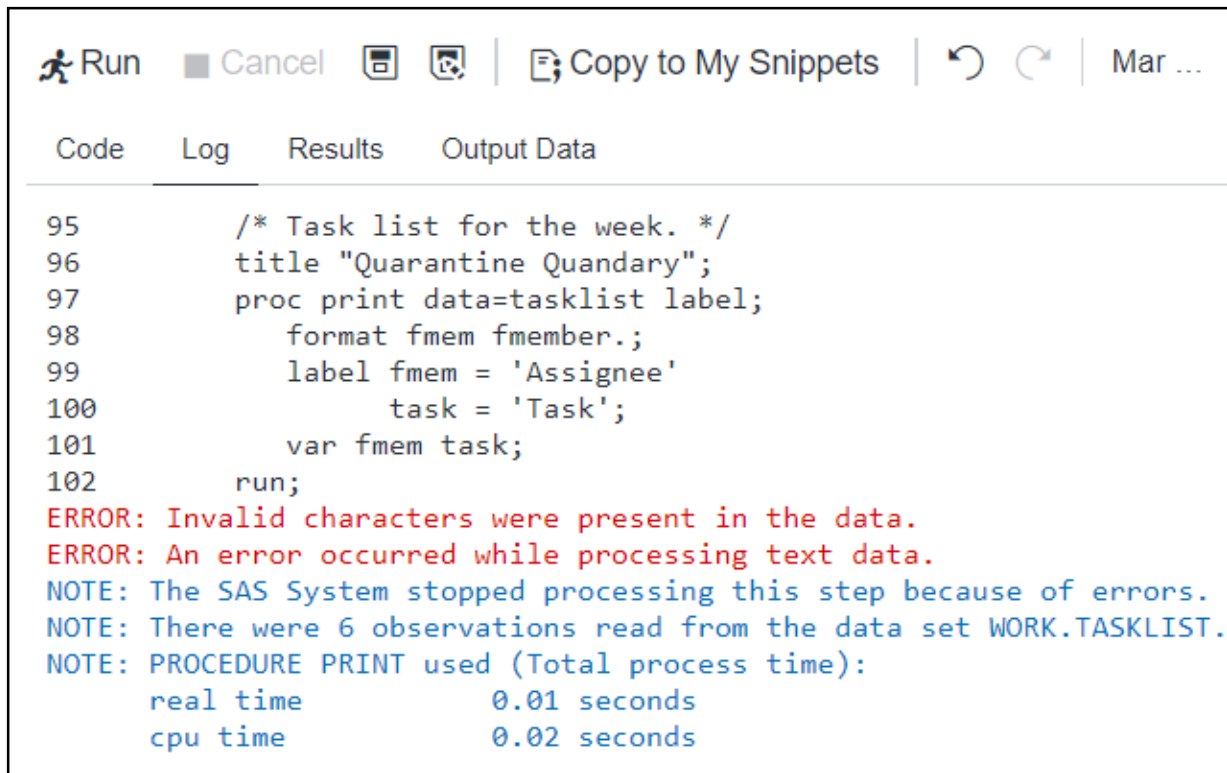

Problem: Garbage characters

```

Run Cancel Copy to My Snippets Mar ...
Code Log Results Output Data
88      data tasklist;
89          length task $ 65;
90          infile tlist;
91          input fmem 1. task &;
92          put task;
93      run;
NOTE: The infile TLIST is:
      Filename=
      Owner Name=      ., Group Name=users,
      Access Permission=-rwxr-xr-x,
      Last Modified=26Mar2020:18:15:23,
      File Size (bytes)=198
Takeout from a local caf.
Draw chalk portraits on the sidewalk.
Read finish Les Miserables.
Binge-watch seasons 14 of favorite show.
At-home workout YouTube videos.
Spa day!
NOTE: 6 records were read from the infile TLIST.
      The minimum record length was 10.
      The maximum record length was 43.
NOTE: The data set WORK.TASKLIST has 6 observations and 2 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

```

Why did Proc Print fail?



The screenshot shows the SAS Studio interface. At the top, there is a toolbar with icons for running, canceling, and copying code, along with a text input field containing 'Mar ...'. Below the toolbar, there are four tabs: 'Code', 'Log', 'Results', and 'Output Data'. The 'Log' tab is currently selected, displaying the following text:

```
95      /* Task list for the week. */
96      title "Quarantine Quandary";
97      proc print data=tasklist label;
98          format fmem fmember.;
99          label fmem = 'Assignee'
100              task = 'Task';
101          var fmem task;
102      run;
```

Below the code, the log output is displayed in red and blue text:

```
ERROR: Invalid characters were present in the data.
ERROR: An error occurred while processing text data.
NOTE: The SAS System stopped processing this step because of errors.
NOTE: There were 6 observations read from the data set WORK.TASKLIST.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.01 seconds
      cpu time           0.02 seconds
```

Check the Data Set Encoding

Quarantine Quandary

The CONTENTS Procedure

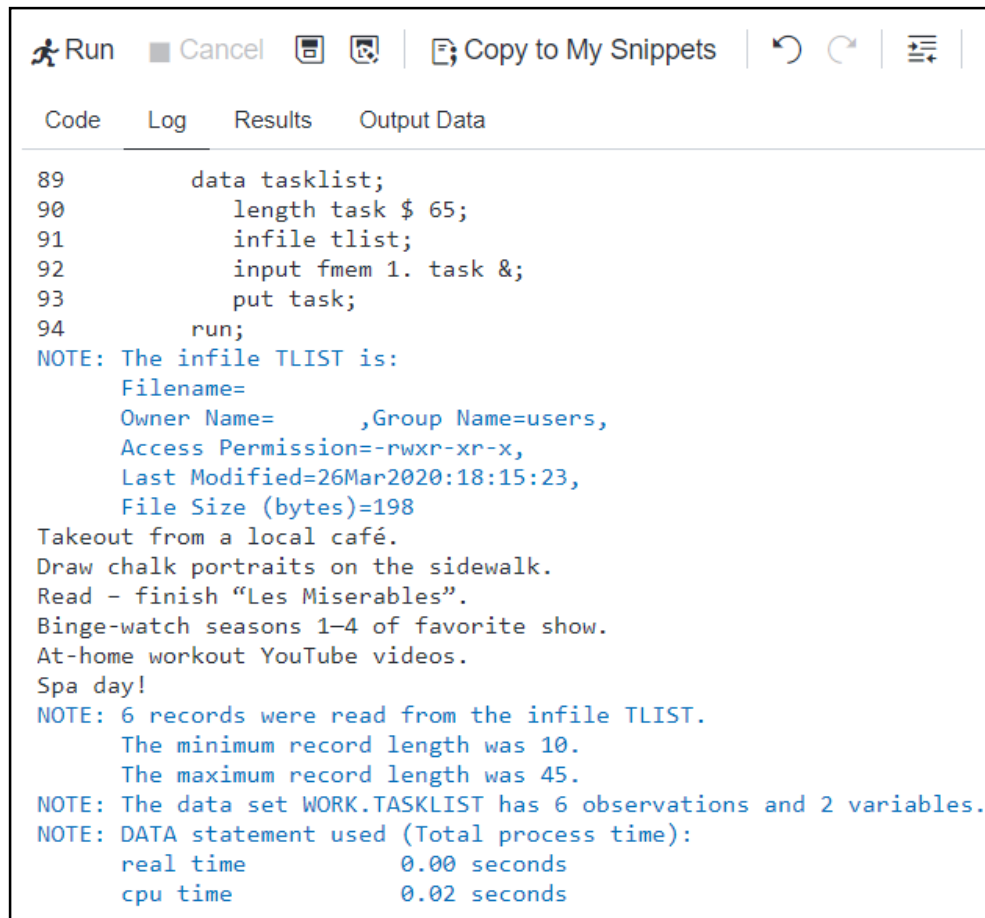
Obs	Assignee	Task
1	1	

Data Set Name	WORK.TASKLIST	Observations	6
Member Type	DATA	Variables	2
Engine	V9	Indexes	0
Created	03/27/2020 09:53:50	Observation Length	80
Last Modified	03/27/2020 09:53:50	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64, LINUX_POWER_64		
Encoding	utf-8 Unicode (UTF-8)		

Possible Solutions

- Proc Datasets with CORRECTENCODING
 - Existing data set
 - Cannot access original data
- FILENAME ENCODING= option

```
filename tlist '<mypath>/Quarantine.txt'  
encoding=wlatin1;
```



The screenshot shows a SAS IDE window with a toolbar at the top containing icons for Run, Cancel, Copy to My Snippets, and other functions. Below the toolbar is a tabbed interface with 'Code', 'Log', 'Results', and 'Output Data' tabs. The 'Code' tab is active, displaying a SAS DATA step. The 'Log' tab is also visible, showing the execution output. The code defines a dataset 'tasklist' with a length of 65, reads from an infile 'tlist', and outputs the data. The log output shows a note about the infile TLIST, a list of tasks, and summary statistics for the data set.

```
89      data tasklist;
90          length task $ 65;
91          infile tlist;
92          input fmem 1. task &;
93          put task;
94      run;
```

NOTE: The infile TLIST is:
Filename=
Owner Name= ,Group Name=users,
Access Permission=-rwxr-xr-x,
Last Modified=26Mar2020:18:15:23,
File Size (bytes)=198

Takeout from a local café.
Draw chalk portraits on the sidewalk.
Read - finish "Les Miserables".
Binge-watch seasons 1-4 of favorite show.
At-home workout YouTube videos.
Spa day!

NOTE: 6 records were read from the infile TLIST.
The minimum record length was 10.
The maximum record length was 45.

NOTE: The data set WORK.TASKLIST has 6 observations and 2 variables.

NOTE: DATA statement used (Total process time):
real time 0.00 seconds
cpu time 0.02 seconds

```

Run Cancel Copy to My Snippets
Code Log Results Output Data
95
96      /* Task list for the week. */
97      title "Quarantine Quandary";
98      proc print data=tasklist label;
99          format fmem fmember.;
100         label fmem = 'Assignee'
101              task = 'Task';
102         var fmem task;
103     run;
NOTE: There were 6 observations read from the data set WORK.TASKLIST.
NOTE: The PROCEDURE PRINT printed page 1.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds

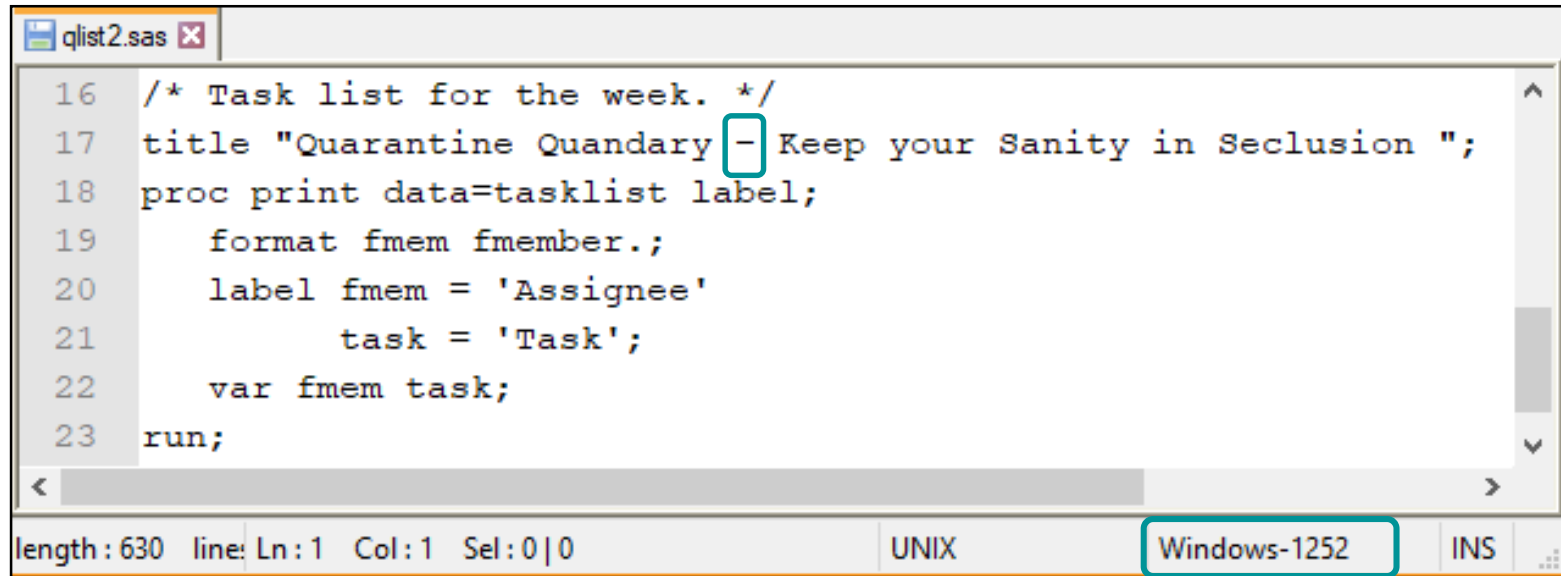
```

Quarantine Quandary

Obs	Assignee	Task
1	1	Takeout from a local café.
2	3	Draw chalk portraits on the sidewalk.
3	1	Read – finish “Les Miserables”.
4	2	Binge-watch seasons 1—4 of favorite show.
5	4	At-home workout YouTube videos.
6	1	Spa day!

Open SAS Programs

WLATIN1 SAS Program



```
16  /* Task list for the week. */
17  title "Quarantine Quandary - Keep your Sanity in Seclusion ";
18  proc print data=tasklist label;
19      format fmem fmember.;
20      label fmem = 'Assignee'
21            task = 'Task';
22      var fmem task;
23  run;
```

length: 630 line: Ln: 1 Col: 1 Sel: 0|0 UNIX Windows-1252 INS

AutoFormatted characters

- Keyboard punctuation ➔ typographic characters
 - Hyphen or minus (-) ➔ En dash (—)
 - Apostrophe (') ➔ Quotation marks (' ')
 - Quotation mark (") ➔ Double quotation marks (" ")

```

15      /* Task list for the week. */
16      title "Quarantine Quandary ♦ Keep your Sanity in Seclusion ";
17  ⊖ proc print data=tasklist label;
18          format fmem fmember.;
19          label fmem = 'Assignee'
20                task = 'Task';
21          var fmem task;
22      run;

```

Quarantine Quandary ♦ Keep your Sanity in Seclusion

Obs	Assignee	Task
1	1	Takeout from a local café.
2	3	Draw chalk portraits on the sidewalk.
3	1	Read – finish “Les Miserables”.
4	2	Binge-watch seasons 1—4 of favorite show.
5	4	At-home workout YouTube videos.
6	1	Spa day!

Possible Solutions

- Change Default text encoding
 - Applies to *all* SAS and text files

Default text encoding: ⓘ

UTF-8 ▼



Open



- ▶ My Favorites
- ▶ Folder Shortcuts
 - ▶ mapstuff
 - ▶ SGF2020
 - ▶ data
 - ▶ saslibrary
 - ▶ utf8librarv

- qlist_withencoding.sas
- qlist.sas
- qlist2.sas
- qlist3.sas
- toASCII.sas

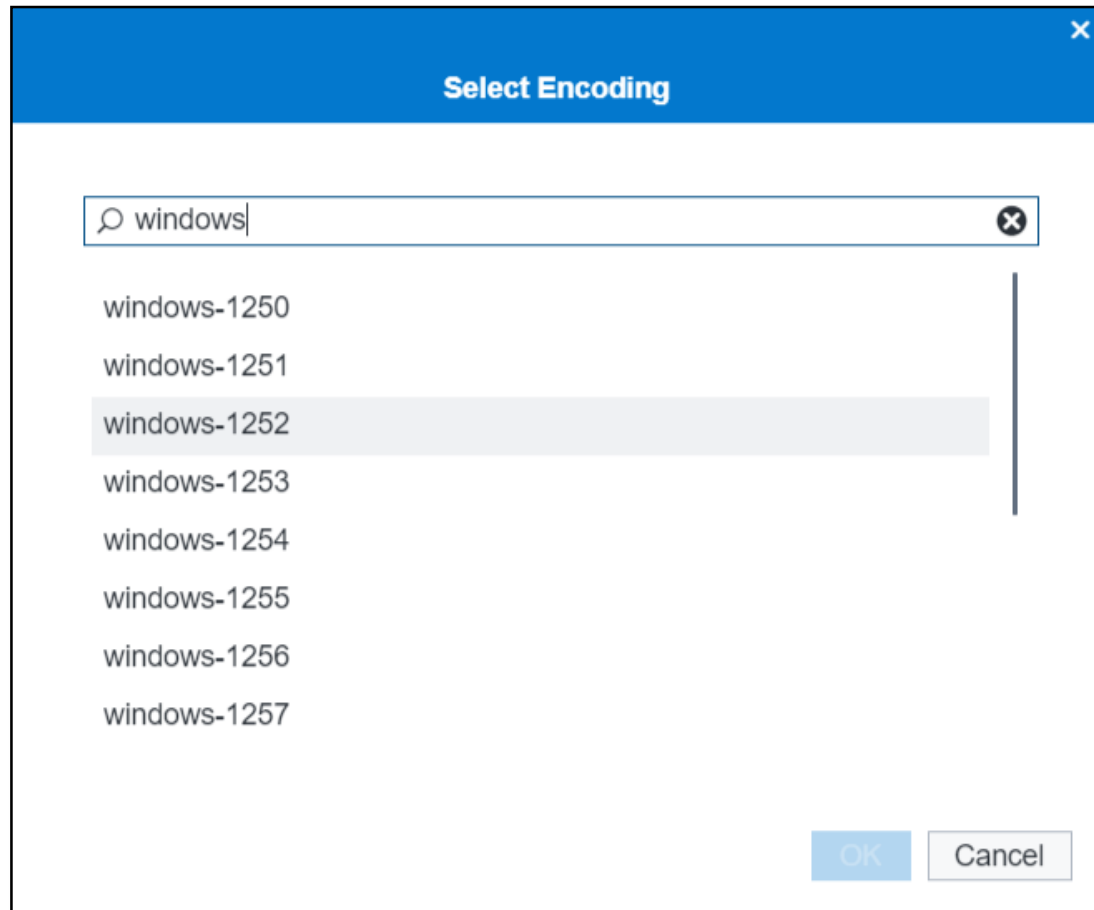
Name:
qlist2.sas

Type:
All Files (*.*)

☒ Open with encoding

Open

Cancel



```

16      /* Task list for the week. */
17      title "Quarantine Quandary - Keep your Sanity in Seclusion ";
18  ⊖  proc print data=tasklist label;
19          format fmem fmember.;
20          label fmem = 'Assignee'
21                  task = 'Task';
22          var fmem task;
23      run;

```

Quarantine Quandary – Keep your Sanity in Seclusion

Obs	Assignee	Task
1	1	Takeout from a local café.
2	3	Draw chalk portraits on the sidewalk.
3	1	Read – finish “Les Miserables”.
4	2	Binge-watch seasons 1—4 of favorite show.
5	4	At-home workout YouTube videos.
6	1	Spa day!

Convert to ASCII

SAS Features

- KPROPDATA
 - Select type of conversion
 - Optionally specify input and output encoding
- BASECHAR
 - Converts accented characters to ASCII form

KPROPDATA

Remove or convert unprintable characters

- Input string
- Output type for unprintable characters
 - UESC – Replace with Unicode escape (\unnnn)
 - BLANK – Replace Space
 - TRIM – Removes the character
 - PUNC – Converts typographic characters to ASCII
- Input encoding
- Output encoding

BASECHAR

Convert characters to base type

- Converts characters to the “base” form without accents or other combining characters

Ä → A

SAS Functions to Help

```
/* Convert typographic characters to ASCII */  
data tasklist;  
  length task $ 45;  
  input task &;  
  basechar_task = basechar(task);  
  punc_task = kpropdata(task, "PUNC");  
  datalines;  
Finish "Les Miserables"  
Binge-watch seasons 1-4  
Get takeout from a café  
run;  
  
title "Convert to ASCII";  
proc print data=tasklist label;  
  label task = "Original text"  
         basechar_task = "After BASECHAR"  
         punc_task = "After KPROPDATA PUNC";  
run;
```

Convert to ASCII

Obs	Original text	After BASECHAR	After KPROPDATA PUNC
1	Finish “Les Miserables”	Finish “Les Miserables”	Finish "Les Miserables"
2	Binge-watch seasons 1—4	Binge-watch seasons 1—4	Binge-watch seasons 1-4
3	Get takeout from a café	Get takeout from a café	Get takeout from a café

A decorative vertical strip on the left side of the slide, featuring a complex, repeating geometric pattern of interlocking lines in a light teal color. The rest of the slide has a dark teal background with a subtle, larger-scale geometric pattern.

Resources

Papers and Documentation

SAS Communities GitHub

- Look for slides and other content
 - <https://github.com/sascommunities/sas-global-forum-2020>

SAS Global Forum 2020 Papers

- [The SAS Encoding Journey: One Byte at a Time](#)
- [Your Data Will Go On: Practice for Character Data Migration](#)

Documentation

- [SAS® 9.4 National Language Support \(NLS\): Reference Guide](#)
 - [Migrating Data from WLATIN1 to UTF-8](#)
 - [Encodings and Their Aliases and Encoding Character Set Compatibility](#)

Papers

- [A Macro for Ensuring Data Integrity When Converting SAS® Data Sets](#)
- [Internationalization 101: Give Some International Flavor to Your SAS® Applications](#)
- [Prevent Culture Shock with SAS® National Language Support Formats](#)
- [SAS® and UTF-8: Ultimately the Finest](#)
- [Understanding SAS® K functions](#)

Thank you!

Contact Information
Elizabeth.Bales@sas.com



SAS[®] GLOBAL FORUM 2020