

# SAS 9.3 UTF-8 Encoding Support and Related Issue Troubleshooting

Jason (Jianduan) Liang

SAS certified:

Platform Administrator, Advanced Programmer for SAS 9



# Agenda

- Introduction
- UTF-8 and other encodings
- SAS options for encoding and configuration
- Other Considerations for UTF-8 data
- Encoding issues troubleshooting techniques  
(tips)

# Introduction

- What is UTF-8?
  - A character encoding capable of encoding all possible characters
- Why UTF-8?
  - Dominant encoding of the www (86.5%)
- SAS system options for encoding
  - Encoding – instructs SAS how to read, process and store data
  - Locale - instructs SAS how to present or display currency, date and time, set timezone values

# UTF-8 and other Encodings

## ➤ ASCII (American Standard Code for Information Interchange)

- 7-bit
- 128 - character set
- Examples (code point-char-hex):

32-Space-20;

63-?-3F;

64-@-40;

65-A-41

# UTF-8 and other Encodings

- ISO 8859-1 (Latin-1) for Western European languages
- Windows-1252 (Latin-1) for Western European languages
  - 8-bit (1 byte, 256 character set)
  - Identical to ascii for the first 128 chars
  - Extended ascii chars examples:
    - 155-£-A3; 161- ©-A9
  - SAS option encoding value: wlatin1 (latin1)

# UTF-8 and other Encodings

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-																
1-																
2-		!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
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	{		}	~	
8-																
9-																
A-		ı	ç	£	¤	¥	¦	§	¨	©	ª	«	¬	-	®	¯
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

# UTF-8 and other Encodings

## ➤ Problems

- Only covers English and Western Europe languages, ISO-8859-2, ...15
- Multiple encoding is required to support national languages
- Same character encoded differently, same code point represents different chars

## ➤ Unicode

- Unicode – assign a unique code/number to every possible character of all languages
- Examples of unicode points:
  - U+0020 – Space
  - U+0041 – A
  - U+00A9 - ©
  - U+C3BF - ÿ

# UTF-8 and other Encodings

## ➤ UTF-8

- UTF-8 – implementation of encoding of unicode character set
- Variable-length
- 8 bit (byte) code unit
- Scheme:

Bits of code point	First code point	Last code point	Bytes in sequence	Byte 1	Byte 2	Byte 3	Byte 4
7	U+0000	U+007F	1	0xxxxxxx			
11	U+0080	U+07FF	2	110xxxxx	10xxxxxx		
16	U+0800	U+FFFF	3	1110xxxx	10xxxxxx	10xxxxxx	
21	U+10000	U+1FFFFF	4	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx
26	U+200000	U+3FFFFFFF	5	111110xx	10xxxxxx	10xxxxxx	10xxxxxx
31	U+4000000	U+7FFFFFFF	6	1111110x	10xxxxxx	10xxxxxx	10xxxxxx



# UTF-8 and other Encodings

## ➤ UTF-8

- Examples:
  - U+0020 – Space – 20 – 0010 0000
  - U+0041 – A – 41 – 0100 0001
  - U+00A9 - © - C2A9 – 1100 0010 1010 1001
  - U+00EB – ë – C2AB – 1100 0010 1010 1011
  - U+20AC - € - E282AC – 1110 0010 1000 0010 1010 1100
- Features of UTF-8
  - Backward compatibility with ASSCII – 1 byte
  - Generally, most characters from latin languages are 1 or 2 bytes, characters from most of Asian languages are 3 bytes
  - Uniquely decodable
  - Clear indication of code sequence length

# SAS Options for Encoding

## ➤ System option encoding

- values: utf-8, wlatin1, latin1
- only valid on SAS invocation
- good for a SAS session
- default value setup at installation (specify or decided by locale value)

## ➤ System option locale

- EN\_CA -> wlatin1

## ➤ SAS Access Interface for RDBMS (DB2)

- environment variable db2codepage=1208

# SAS Options for Encoding

- Encoding options for transcoding
  - data set option encoding=
  - libname statement option encoding=
  - encoding= option for file access
- SAS configuration and configuration changes for encoding
  - set SAS session encoding=UTF-8
  - for DB2 client, setup env db2codepage=1208
  - a couple of things could go wrong

# Other Considerations for UTF-8 Data

- characters in one encoding may not be present in another encoding
- number of bytes of a character in one encoding may differ from that in another encoding – truncation
- use k-functions vs. byte-based string functions

<b>byte-based string functions</b>	<b>k-functions</b>
length()	klength()
substr()	ksubstr()
index()	kindex()
count()	kcount()
scan()	kscan()
...	...

# Other Considerations for UTF-8 Data

## ➤ k-functions example

```
data work.utf8datasample;  
set work.UTF8DATASAMPLE;  
len=length(chars_utf8);  
klen=klength(chars_utf8);  
national_chars=kcount(chars_utf8);  
idx_copyright=index(chars_utf8, '@');  
kidx_copyright=kindex(chars_utf8, '@');  
idx_bracket=index(chars_utf8, '(');  
kidx_bracket=kindex(chars_utf8, '(');  
run;
```

chars_utf8	len	klen	idx_copyright	kidx_copyright	idx_bracket	kidx_bracket	national_chars
English	7	7	0	0	0	0	0
ó6#%/£	11	6	6	4	0	0	5
中文(简体)	14	6	0	0	7	3	4

# Encoding Issues Troubleshooting

- Troubleshooting Techniques (Tips)
  - check your SAS session encoding
  - check your data set encoding
  - read binary code of source data
  - use k-functions
  - no more fixed column input
  - use EG instead of Base SAS, if you can

# Encoding Issues Troubleshooting

## ➤ Check your SAS session encoding

*proc options option=encoding;*

```
15      proc options option=encoding;run;
      SAS (r) Proprietary Software Release 9.3  TS1M2
ENCODING=UTF-8    Specifies default encoding for internal processing of data
NOTE: PROCEDURE OPTIONS used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds
```

*proc options group=languagecontrol;*

```
Group=LANGUAGECONTROL
DATESTYLE=MDY    Identify sequence of month, day and year when ANYDATE informat data is
ambiguous
URLENCODING=SESSION
                Specifies URL percent encoding for the URLENCODE and URLDECODE
functions
DBCS            Process Double Byte Character Sets.
DBCSSLANG=UNKNOWN
                Specifies the double-byte character set (DBCS) language to use
DBCSTYPE=UTF8   Specifies a double-byte character set (DBCS) encoding method
ENCODING=UTF-8  Specifies default encoding for internal processing of data
LOCALE=EN_US    Specifies the current locale for the SAS session
NONLSCOMPATMODE
                Uses the user specified encoding to process character data
NOTE: PROCEDURE OPTIONS used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds
```

# Encoding Issues Troubleshooting

## ➤ Common error messages

```
NOTE: Remote signon to IEDM2A04 commencing (SAS Release 9.03.01M2P081512).  
NOTE: Script file 'tcpunix.scr' entered.  
ERROR: The client session encoding wlatin1 is not compatible with the server session encoding  
utf-8.  
ERROR: Remote signon to IEDM2A04 canceled.
```

```
ERROR: The client session encoding utf-8 is not compatible with the server session encoding latin1.  
ERROR: Remote signon to IEDM2A04 canceled.
```

```
SAS Enterprise Guide  
An unexpected error has occurred  
A connection could not be made to the requested resource.
```

---

```
Exception information
```

---

```
[Error] The launch of the server process failed because of a SAS kernel initialization failure.
```



# Encoding Issues Troubleshooting

## ➤ Check your data set encoding

```
%let dsn=jliang1.group_info;  
%let dsid=%sysfunc(open(&dsn,i));  
%put &dsn ENCODING is:  
%sysfunc(attrc(&dsid,encoding));
```

Or

```
proc contents data=jliang1.group_info;run;
```

The CONTENTS Procedure

<b>Data Set Name</b>	JLIANG1.GROUP_INFO	<b>Observations</b>	64
<b>Member Type</b>	DATA	<b>Variables</b>	7
<b>Engine</b>	BASE	<b>Indexes</b>	0
<b>Created</b>	Wednesday, June 29, 2011 10:48:32 AM	<b>Observation Length</b>	733
<b>Last Modified</b>	Wednesday, June 29, 2011 10:48:32 AM	<b>Deleted Observations</b>	0
<b>Protection</b>		<b>Compressed</b>	NO
<b>Data Set Type</b>		<b>Sorted</b>	NO
<b>Label</b>			
<b>Data Representation</b>	HP_UX_64, RS_6000_AIX_64, SOLARIS_64, HP_IA64		
<b>Encoding</b>	latin1 Western (ISO)		

# Encoding Issues Troubleshooting

## ➤ Reading binary code of source data

- *Binary code of db2 data column*

```
proc sql;
```

```
  connect to db2 as mydb (database=biebdwd user=jliang1  
  password=<mypass>);
```

```
  create table t3 as
```

```
  select CODE_TABLE_DESC from connection to mydb  
  (select HEX(CODE_TABLE_DESC) as CODE_TABLE_DESC  
  from JL.CODE_TABLE where CODE_TABLE_ID=703);
```

```
  disconnect from mydb;quit;
```

```
4120636F64652074686174206964656E746966696573207468652  
064656C6976657...
```

# Encoding Issues Troubleshooting

## ➤ Reading binary code of source data

- SAS data sets

```
data t2;
```

```
format CODE_TABLE_DESC $hex255.;
```

```
set t1;
```

```
run;
```

```
4120636F64652074686174206964656E7469666965  
73207468652064656C6976657...
```

# Encoding Issues Troubleshooting

## ➤ Common errors

CODE_TABLE_ID	CODE_TABLE_DESC
703	A code that identifies the delivery site's domain within continuing care. For example: ♦Long-Term Care♦, ♦Community Care♦.

- Query binary code found: c293 and c294
- db2codepage value was not setup properly

CODE_TABLE_ID	CODE_TABLE_DESC
703	A code that identifies the delivery site's domain within continuing care. For example: "Long-Term Care", "Community Care".

# Encoding Issues Troubleshooting

## ➤ Troubleshooting Techniques (Tips)

Obs	CODE_TEXT_ID	CODE_TEXT_DESC
1	4113	Chest, spine and pelvis ?skeletal survey for secondary neoplasms etc.
2	8266	Chest, spine and pelvis ?skeletal survey for secondary neoplasms etc.
3	12387	Chest, spine and pelvis ?skeletal survey for secondary neoplasms etc.
4	16510	Chest, spine and pelvis ?skeletal survey for secondary neoplasms etc.
5	21002	Chest, spine and pelvis
6	20635	Chest, spine and pelvis ?skeletal survey for secondary neoplasms etc.

VIEWTABLE: Jliang1.Query\_for\_code\_text\_ct\_0000 (QUERY\_FOR\_CODE\_TEXT\_CT)

	CODE_TEXT_ID	CODE_TEXT_DESC
1	4113	Chest, spine and pelvis →skeletal survey for secondary neoplasms etc.
2	8266	Chest, spine and pelvis →skeletal survey for secondary neoplasms etc.
3	12387	Chest, spine and pelvis →skeletal survey for secondary neoplasms etc.
4	16510	Chest, spine and pelvis →skeletal survey for secondary neoplasms etc.
5	20635	Chest, spine and pelvis →skeletal survey for secondary neoplasms etc.
6	21002	Chest, spine and pelvis

	CODE_TEXT_ID	CODE_TEXT_DESC
1	4113	Chest, spine and pelvis skeletal survey for secondary neoplasms etc.
2	8266	Chest, spine and pelvis skeletal survey for secondary neoplasms etc.
3	12387	Chest, spine and pelvis skeletal survey for secondary neoplasms etc.
4	16510	Chest, spine and pelvis skeletal survey for secondary neoplasms etc.
5	20635	Chest, spine and pelvis skeletal survey for secondary neoplasms etc.
6	21002	Chest, spine and pelvis

# Encoding Issues Troubleshooting

## ➤ Example – Base SAS, aix, winsql,EG.

CODE_TABLE_DESC
A code that identifies the delivery site's domain within continuing care. For example: Long-Term Care, Community Care.

A screenshot of a PuTTY terminal window titled "IEDM2A04 - PuTTY". The terminal displays the text "A code that identifies the delivery site's domain within continuing care. For example: Long-Term Care, Community Care." on a black background with white text. The text is displayed as a single line.

CODE_TABLE_DESC
A code that identifies the delivery site's domain within continuing care. For example: Long-Term Care, Community Care.

CODE_TABLE_DESC
A code that identifies the delivery site's domain within continuing care. For example: "Long-Term Care", "Community Care".

# Reference

## **SAS ®9.3 National Language Support (NLS): Reference Guide:**

<http://support.sas.com/documentation/cdl/en/nlsref/63072/HTML/default/viewer.htm#titlepage.htm>