SAS 9.3 UTF-8 Encoding Support and Related Issue Troubleshooting

Jason (Jianduan) Liang

SAS certified:

Platform Administrator, Advanced Programmer for SAS 9

CGI

Agenda

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

CGI

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

CGI

UTF-8 and other Encodings

- ASSCII (American Standard Code for Information Interchange)
  - 7-bit
  - 128 - character set
  - Examples (code point-char-hex):

    32-Space-20;

    63-?-3F;

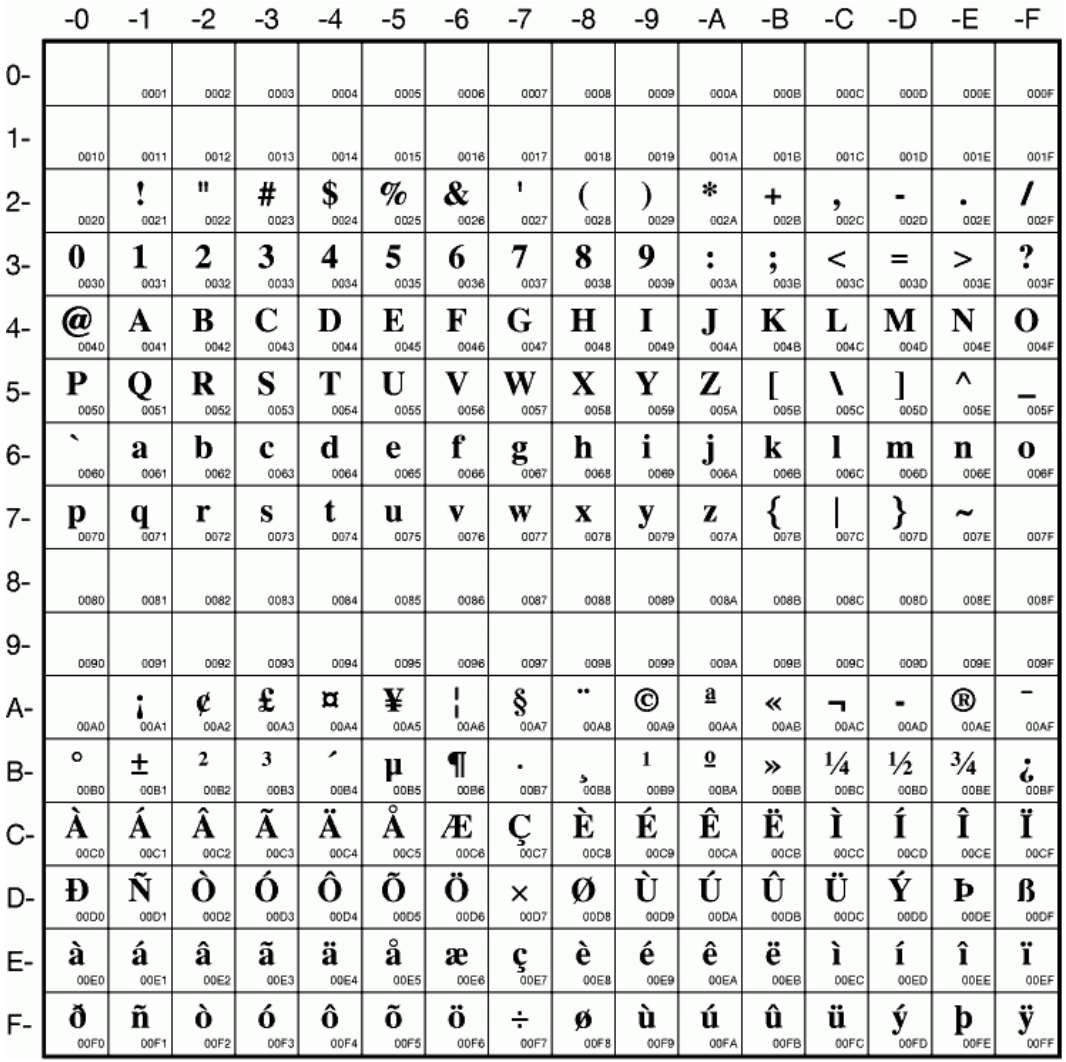
    64-@-40;

    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 asscii 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-		0001	0002	0003	0004	0005	0006	0007	0008	0009	000A	000B	000C	000D	000E	000F
1-	0010	0011	0012	0013	0014	0015	0016	0017	0018	0019	001A	001B	001C	001D	001E	001F
2-	0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
3-	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
4-	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
5-	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
6-	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
7-	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	007F
8-	0080	0081	0082	0083	0084	0085	0086	0087	0088	0089	008A	008B	008C	008D	008E	008F
9-	0090	0091	0092	0093	0094	0095	0096	0097	0098	0099	009A	009B	009C	009D	009E	009F
A-	00A0	¡ 00A1	¢ 00A2	£ 00A3	¤ 00A4	¥ 00A5	¦ 00A6	§ 00A7	¨ 00A8	© 00A9	ª 00AA	« 00AB	¬ 00AC	- 00AD	® 00AE	¯ 00AF
B-	° 00B0	± 00B1	2 00B2	3 00B3	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	1 00B9	º 00BA	» 00BB	¼ 00BC	½ 00BD	¾ 00BE	¿ 00BF
C-	À 00C0	Á 00C1	Â 00C2	Ã 00C3	Ä 00C4	Å 00C5	Æ 00C6	Ç 00C7	È 00C8	É 00C9	Ê 00CA	Ë 00CB	Ì 00CC	Í 00CD	Î 00CE	Ï 00CF
D-	Ð 00D0	Ñ 00D1	Ò 00D2	Ó 00D3	Ô 00D4	Õ 00D5	Ö 00D6	× 00D7	Ø 00D8	Ù 00D9	Ú 00DA	Û 00DB	Ü 00DC	Ý 00DD	Þ 00DE	ß 00DF
E-	à 00E0	á 00E1	â 00E2	ã 00E3	ä 00E4	å 00E5	æ 00E6	ç 00E7	è 00E8	é 00E9	ê 00EA	ë 00EB	ì 00EC	í 00ED	î 00EE	ï 00EF
F-	ð 00F0	ñ 00F1	ò 00F2	ó 00F3	ô 00F4	õ 00F5	ö 00F6	÷ 00F7	ø 00F8	ù 00F9	ú 00FA	û 00FB	ü 00FC	ý 00FD	þ 00FE	ÿ 00FF

CGI

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

CGI

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+3FFFFFF	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

CGI

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 varible 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

CGI

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

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

CGI

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¥©¼£	11	6	6	4	0	0	5
中文(簡体)	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.
 DBCSLANG=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
```

CGI

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

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

CGI

Encoding Issues Troubleshooting

- Reading binary code of source data
  - *Binarary 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;

4120636F6465207468617420696465E7469666965732074686520696465C6976657…

Encoding Issues Troubleshooting

- Reading binary code of source data
  - SAS data sets

```
data t2;
   format CODE_TABLE_DESC $hex255.;
   set t1;
run;
```

4120636F6465207468617420696465 6E74696669657320746865206465…

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

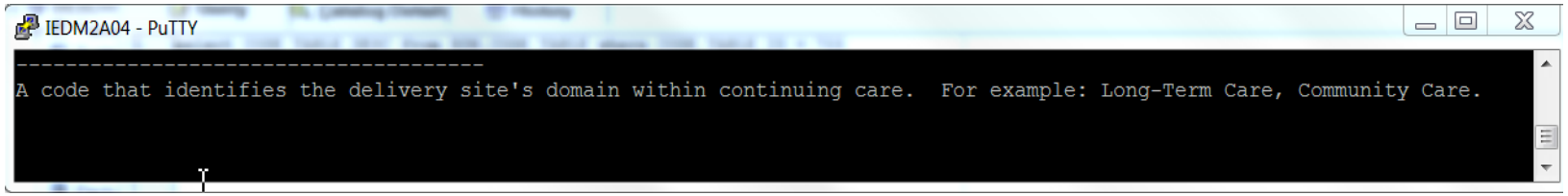
	CODE_TEXT_ID	CODE_TEXT_DESC
1	4113	Chest, spine and pelvis skeletal survey for secondary neoplasms etc.|
2	8266	Chest, spine and pelvis skeletal survey for secondary neoplasms etc.|
3	12387	Chest, spine and pelvis skeletal survey for secondary neoplasms etc.|
4	16510	Chest, spine and pelvis skeletal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

CGI

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.



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