

## Multilingual data support in Dataset-XML with SAS® Clinical Data Integration

Jing Gao, SAS R&D, Beijing, China

### ABSTRACT

Dataset-XML is a CDISC XML format for exchanging clinical study data between any two entities. That is, in addition to support the transport of datasets as part of a submission to the FDA, it may also be used to facilitate other data interchange use cases. For example, the Dataset-XML data format can be used by a CRO to transmit SDTM or AdAM datasets to a sponsor organization. Dataset-XML can represent any tabular dataset including SDTM, AdAM, SEND, or non-standard legacy datasets.

With the growing trends in the globalization of Drug Development, there are increasing clinical trials conducted in various countries. So clinical trial data that comes from various countries using different languages may need to be processed. In other hand, CDISC standards are becoming more accepted outside the USA, especially, SDTM is used in many countries that use other character encodings (e.g. Shift-JIS in Japan) for submissions to local regulatory authorities. In this context, one of the advantages of the Dataset-XML format is highlighted: Dataset-XML supports all language encodings supported by XML. This requires that the related industry solutions not only support US-ASCII characters, but also support non-ASCII characters in Dataset-XML.

This presentation will introduce: 1) how to create Dataset-XML files with multiple encodings (UTF-8, ISO-8859-1, Shift-JIS, etc.) from SAS datasets using SAS Clinical Data Integration (CDI); 2) how to choose the appropriate encoding for the particular languages in Dataset-XML; 3) the SAS Macros called by CDI to create Dataset-XML; 4) Lastly, let's look into the non-ASCII characters whether are supported by the Dataset-XML Tools (OpenCDISC, XPT2DatasetXML, etc.).

### INTRODUCTION

SAS Clinical Data Integration provides an easy-to-use visual interface for transforming, managing, and verifying the creation of industry-mandated data standards such as those created by Clinical Data Interchange Standards Consortium (CDISC). The SAS solution has prebuilt transformations for CDISC models. The CDISC-Dataset-XML Creation transformation is one of prebuilt transformations and used to visually create a separate Dataset-XML file for each domain or data table in a study or submission. The CDISC-Dataset-XML Creation transformation enables you to select the output encoding for the Dataset-XML files. You can also enter any valid value for the output encoding. With this functionality, you can create the Dataset-XML files that contain non-ASCII characters, and multilingual data in Dataset-XML can be processed correctly.

## CREATING A DATASET-XML FILE WITH SAS® CLINICAL DATA INTEGRATION

### OVERVIEW OF CREATING A DATASET-XML FILE

The CDISC-Dataset-XML Creation transformation creates a separate Dataset-XML file for each domain or data table in a study or submission. Each Dataset-XML file is named based on the domain or data table name. For example, the DM domain creates the dm.xml file.

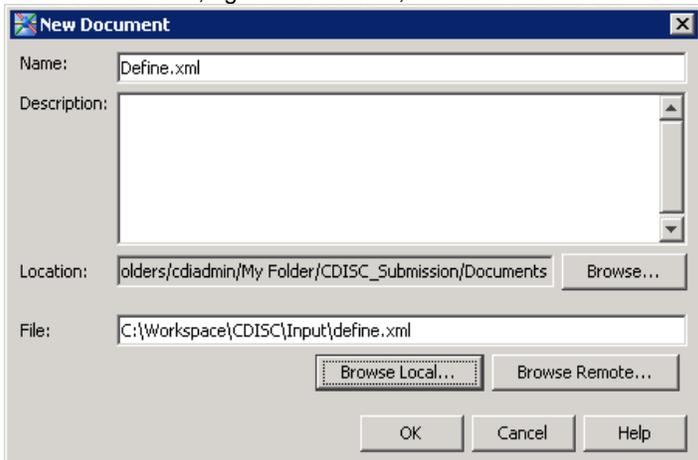
The CDISC-Dataset-XML Creation transformation enables you to specify these options:

- Select the domains or data tables for which to create Dataset-XML files.
- Create a ZIP file that contains the Dataset-XML files.
- Delete Dataset-XML files that are included in a ZIP file.
- Check the data lengths of text variables against the metadata in the define.xml input file.
- Specify a header comment to include in the Dataset-XML files.

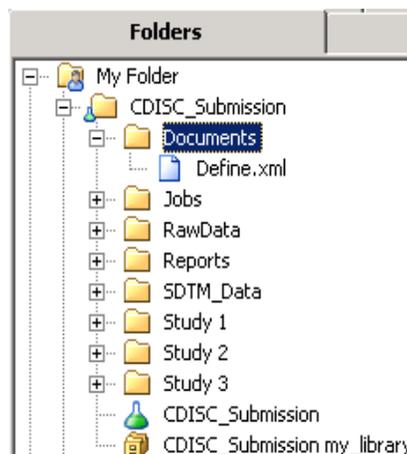
### REGISTER A NEW DOCUMENT FILE

Before you can create a job to create a Dataset-XML file, you must register a new document file. The document file must be associated with a valid define.xml file that contains definitions for all domains and data tables for which to create Dataset-XML files.

1. In the Folders tree, right-click a folder, and then select New->Document.



**Display 1. Document for the define.xml File**



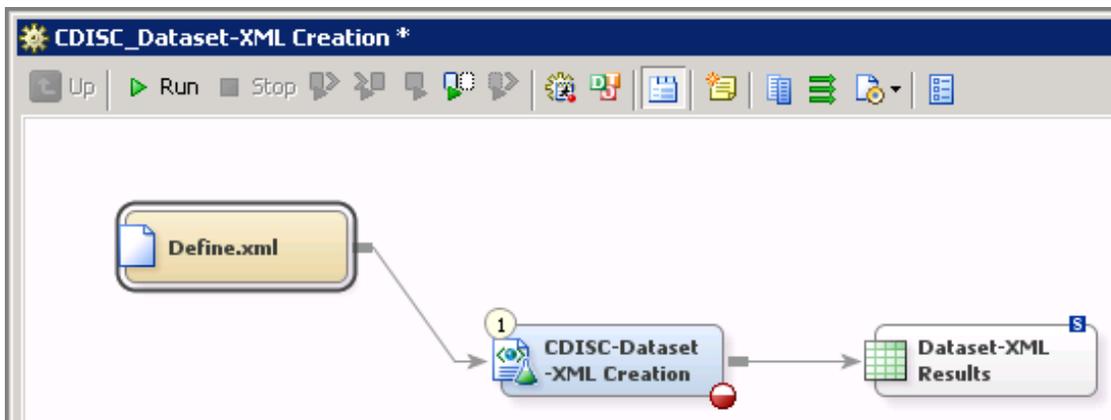
**Display 2. Document Folder**

2. Enter a name and an optional description.
3. Click Browse adjacent to the Location field, and then navigate to the location in which to store the document.
4. Click Browse adjacent to the File field, and then navigate to the location of the define.xml file.
5. Click OK.
6. Create an empty job.
7. In the Transformations tree, expand Clinical, and then drag and drop CDISC Define Creation onto the diagram.
8. Use the Transformation CDISC Define Creation to transform SDTM domains or ADaM datasets into a define.xml file.

### CREATE A DATASET-XML FILE

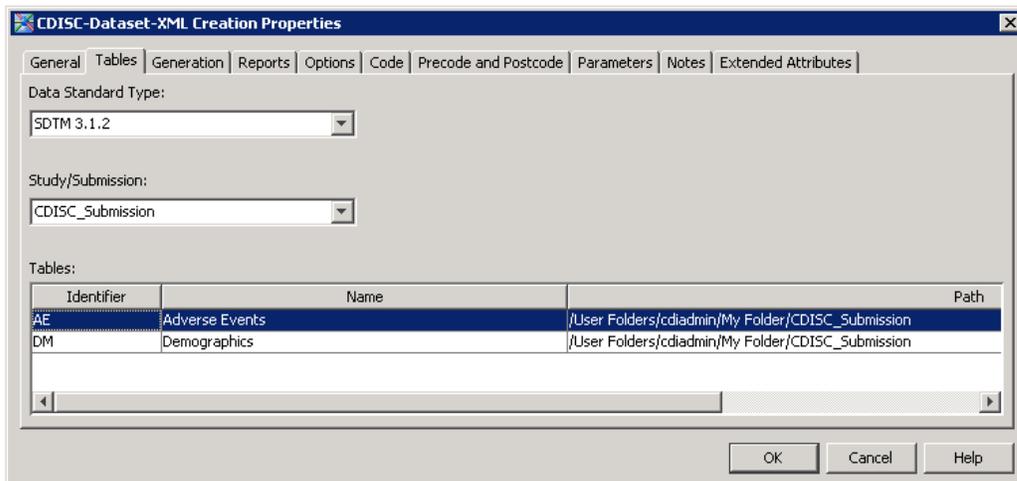
To create a Dataset-XML file, perform the following steps:

1. Create an empty job.
2. In the Transformations tree, expand Clinical, and then drag and drop CDISC-Dataset-XML Creation onto the diagram.
3. From the Folders tree, drag and drop the define.xml file onto the diagram.
4. To connect the define.xml file to the CDISC-Dataset-XML Creation transformation, drag and drop the cursor from the output port of the define.xml file to the input port of the CDISC-Dataset-XML Creation transformation.



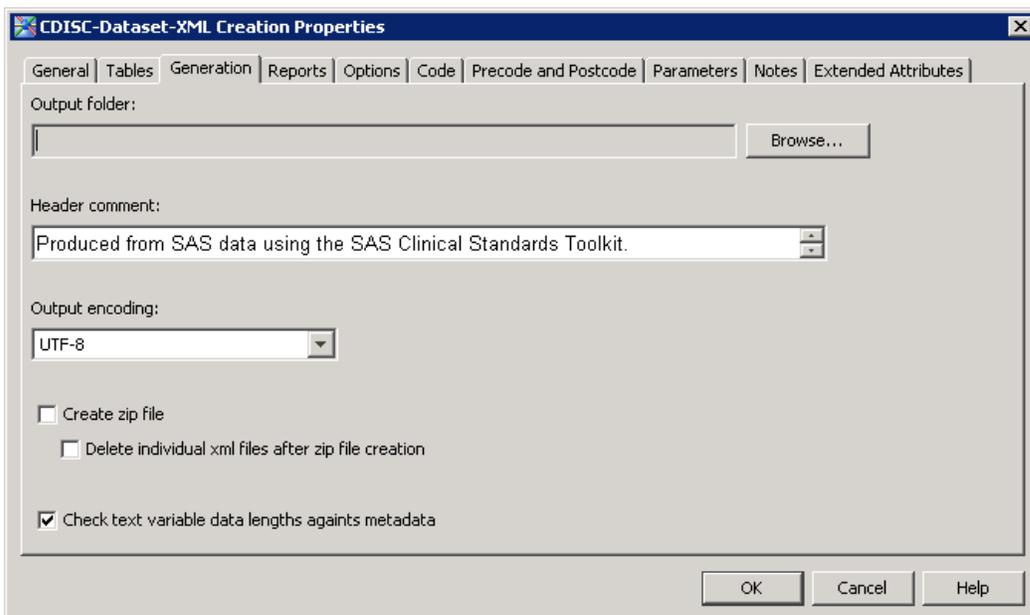
**Display 3. The CDISC-Dataset-XML Creation Job**

5. In the diagram, double-click CDISC-Dataset-XML Creation. The CDISC-Dataset-XML Creation Properties dialog box appears.
6. Click the Tables tab.



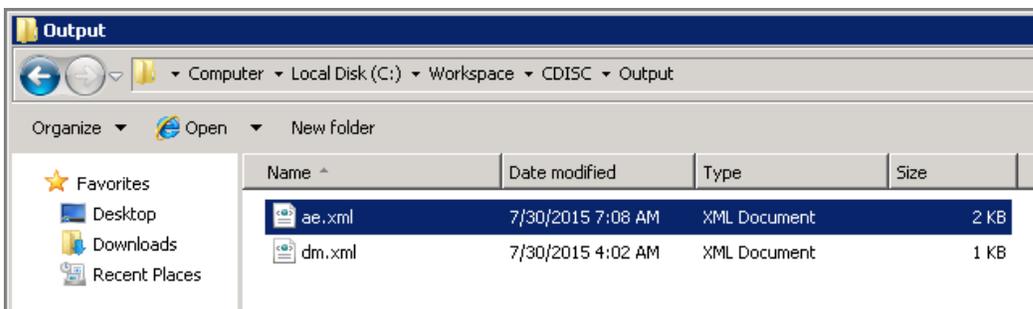
**Display 4. The Tables Tab**

7. From the Data Standard Type drop-down list, select the data standard version.
8. From the Study/Submission drop-down list, select the study or submission.
9. From the Tables list, select the domains or data tables for which to create Dataset-XML files.
10. Click the Generation tab.



**Display 5. The Generation tab**

11. Click Browse adjacent to Output folder, and then navigate to an output folder.
12. Enter a header comment, and then select the **output encoding**.
13. Click OK, and then click Run.
14. Verify that there are no errors. The Dataset-XML files are created in the output folder.



Display 6. List of the dataset-xml files created

## CHOOSE THE APPROPRIATE ENCODING FOR THE DATASET-XML

The CDISC-Dataset-XML Creation transformation enables you to select the output encoding for the Dataset-XML files: US-ASCII, ISO-8859-1 and UTF-8. You can also enter a value for the output encoding, such as Shift-JIS. The value must be a valid encoding.

Here are some examples.

### OUTPUT ENCODING: US-ASCII

US-ASCII is a 7-bit character encoding that every single byte represents a unique character. It includes 128 characters: 33 control characters and 95 printable characters.

	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	\$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	
D	CR	GS	-	=	M	]	m	}
E	S0	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

Figure 1. US-ASCII code chart

### OUTPUT ENCODING: ISO-8859-1

ISO-8859-1 is an 8-bit extension to the US-ASCII encoding, also called Latin-1. The ISO-8859-1 is generally intended for most Western European languages.

ISO-8859-1 can be used in the following European languages (to name a few):

**Danish**                      **UK English**                      **German**                      **Italian**  
**Portuguese**                      **Spanish**                      **Swedish**

ISO-8859-1 is a superset of US-ASCII, the first 128 characters of ISO-8859-1 have the same code points with US-ASCII.

ISO-8859-1																
	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	NUL	SOH	STX	ETX	EOT	ENO	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1x	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2x	SP	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL
8x	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9x	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
Ax	NBSP	ı	ç	£	¤	¥	ı	§	¨	©	ª	«	¬	®	¯	
Bx	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
Cx	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Dx	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
Ex	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
Fx	ø	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Figure 2. ISO-8859-1 code chart

### OUTPUT ENCODING: UTF-8

UTF-8 is a character encoding capable of encoding all possible characters, yet is backwards compatible with US-ASCII.

Here is an example, the SAS dataset has 3 rows of multilingual data.

#	STUDYID	DOMAIN	USUBJID	AETERM	AEBODSYS	AESEV
1	CDISC01 ...	AE	1001 ...	腹泻 ...	肠胃疾病 ...	轻微的 ...
2	CDISC01 ...	AE	1002 ...	HEADACHE ...	Nervous system disorders ...	MILD ...
3	CDISC01 ...	AE	1003 ...	頭が痛い ...	神経系統病気 ...	軽い ...

Display 7. The SAS dataset AE.sas7bdat

As we can see from the Display 7 above, the AE dataset contains Chinese characters, US English characters and Japanese characters, in this case, which encoding should be used to handle the multilingual data? UTF-8. Using UTF-8 as the encoding, the Dataset-XML can contain multilingual data without losing any data.

```

<?xml version="1.0" encoding="UTF-8" ?>
<!-- Produced from SAS data using the SAS Clinical Standard's Toolkit. -->
- <ODM xmlns="http://www.cdisc.org/ns/odm/v1.3" xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0" ODMVersion="1.3.2"
  data:DatasetXMLVersion="1.0.0">
- <ClinicalData StudyOID="STUDY1" MetadataVersionOID="MDV.Export_Dataset_XML">
- <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="1">
  <ItemData ItemOID="IT.AE.STUDYID" Value="CDISC01" />
  <ItemData ItemOID="IT.AE.DOMAIN" Value="AE" />
  <ItemData ItemOID="IT.AE.USUBJID" Value="1001" />
  <ItemData ItemOID="IT.AE.AETERM" Value="震泻" />
  <ItemData ItemOID="IT.AE.AEBODSYS" Value="腸胃疾病" />
  <ItemData ItemOID="IT.AE.AESEV" Value="軽微的" />
</ItemGroupData>
- <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="2">
  <ItemData ItemOID="IT.AE.STUDYID" Value="CDISC01" />
  <ItemData ItemOID="IT.AE.DOMAIN" Value="AE" />
  <ItemData ItemOID="IT.AE.USUBJID" Value="1002" />
  <ItemData ItemOID="IT.AE.AETERM" Value="HEADACHE" />
  <ItemData ItemOID="IT.AE.AEBODSYS" Value="Nervous system disorders" />
  <ItemData ItemOID="IT.AE.AESEV" Value="MILD" />
</ItemGroupData>
- <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="3">
  <ItemData ItemOID="IT.AE.STUDYID" Value="CDISC01" />
  <ItemData ItemOID="IT.AE.DOMAIN" Value="AE" />
  <ItemData ItemOID="IT.AE.USUBJID" Value="1003" />
  <ItemData ItemOID="IT.AE.AETERM" Value="頭が痛い" />
  <ItemData ItemOID="IT.AE.AEBODSYS" Value="神経系統病氣" />
  <ItemData ItemOID="IT.AE.AESEV" Value="軽い" />
</ItemGroupData>
</ClinicalData>
</ODM>

```

Display 8. The AE.xml created from the SAS dataset AE.sas7bdat.

### OUTPUT ENCODING: SHIFT-JIS

Shift-JIS is a character encoding for the Japanese language. It is also a superset of US-ASCII except for the backslash and tilde.

Shift-JIS	US-ASCII	Code Point (hexadecimal)
¥	\	5C
-	~	7E

Table 1. The same code point represents different character in Shift-JIS and US-ASCII.

### XML ENCODING

UTF-8 is the default for documents without encoding information:

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

The actual character encoding in the XML files Must Match the encoding declaration of the XML files

### WHICH ENCODING SHOULD BE USED?

When using CDI to create Dataset-XML files, it is very important to select the correct encoding for the Dataset-XML files. If you select the wrong encoding, the garbled characters will appear in the Dataset-XML files.

Which encoding should be used? It depends on which languages the SAS dataset contains.

Why don't we consider the encoding of the SAS dataset? That is because the CDI will convert the character data from one encoding to another encoding when the encoding of data in the original location is different from the encoding of the data's destination. So we just need to know whether the data in SAS dataset is supported by the Dataset-XML encoding you select.

To avoid the loss of the character data during the transcoding, we must select the correct encoding.

Table 2 lists some examples that the languages that are supported by the character encodings.

Languages	Character encodings			
	US-ASCII	ISO-8859-1	UTF-8	Shift-JIS
US English	✓	✓	✓	✓
Western European	✗	✓	✓	✗
Japanese	✗	✗	✓	✓

**Table 2. Encodings and Languages**

Note: ISO-8859-1 does not support all the Western languages, such as, the euro sign €.

Here are some examples.

1. US English

If the SAS dataset only contains the US English language, you don't need to worry about the encoding. It is ok to use the encodings such as US-ASCII, ISO-8859-1, UTF-8, etc.

2. Western European

If the SAS dataset contains Western European languages (such as UK English, German, and Spanish), US-ASCII cannot be used, ISO-8859-1 and UTF-8 can be used.

The DM domain contains two German characters: fröhlich and weiß.

#	STUDYID	DOMAIN	USUBJID	INVNAM	SEX	RACE
1	CDISC01	DM	1008	fröhlich	F	weiß

**Display 9. The SAS Dataset DM.sas7bdat**

As we can see from Display 10 below, the two German characters were transcoded incorrectly when using US-ASCII as the encoding.

```
<?xml version="1.0" encoding="US-ASCII" ?>
<!-- Produced from SAS data using the SAS Clinical Standard's Toolkit. -->
- <ODM xmlns="http://www.cdisc.org/ns/odm/v1.3" xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0"
  data:DatasetXMLVersion="1.0.0">
- <ClinicalData StudyOID="STUDY1" MetaDataVersionOID="MDV.Export_Dataset_XML">
- <ItemGroupData ItemGroupOID="IG.DM" data:ItemGroupDataSeq="1">
  <ItemData ItemOID="IT.DM.STUDYID" Value="CDISC01" />
  <ItemData ItemOID="IT.DM.DOMAIN" Value="DM" />
  <ItemData ItemOID="IT.DM.USUBJID" Value="1008" />
  <ItemData ItemOID="IT.DM.INVNAM" Value="frC6hlich" />
  <ItemData ItemOID="IT.DM.SEX" Value="F" />
  <ItemData ItemOID="IT.DM.RACE" Value="weiC" />
</ItemGroupData>
</ClinicalData>
</ODM>
```

**Display 10. The DM.xml with US-ASCII encoding**

The two German characters are displayed correctly when using ISO-8859-1 as the encoding.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!-- Produced from SAS data using the SAS Clinical Standard's Toolkit. -->
- <ODM xmlns="http://www.cdisc.org/ns/odm/v1.3" xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0"
  data:DatasetXMLVersion="1.0.0">
- <ClinicalData StudyOID="STUDY1" MetaDataVersionOID="MDV.Export_Dataset_XML">
  - <ItemGroupData ItemGroupOID="IG.DM" data:ItemGroupDataSeq="1">
    <ItemData ItemOID="IT.DM.STUDYID" Value="CDISC01" />
    <ItemData ItemOID="IT.DM.DOMAIN" Value="DM" />
    <ItemData ItemOID="IT.DM.USUBJID" Value="1008" />
    <ItemData ItemOID="IT.DM.INVNAM" Value="fröhlich" />
    <ItemData ItemOID="IT.DM.SEX" Value="F" />
    <ItemData ItemOID="IT.DM.RACE" Value="weiß" />
  </ItemGroupData>
</ClinicalData>
</ODM>
```

Display 11. The DM.xml with ISO-8859-1 encoding

### 3. Japanese

If the SAS dataset contains Japanese language, US-ASCII and ISO-8859-1 cannot be used, Shift-JIS and UTF-8 can be used.

The AE domain contains three Japanese characters: "頭が痛い", "神経系統病気" and "軽い".

#	STUDYID	DOMAIN	USUBJID	AETERM	AEBODSYS	AESEV
1	CDISC01	AE	1003	頭が痛い	神経系統病気	軽い

Display 12. The SAS Dataset AE.sas7bdat

As we can see from Display 13 below, the three Japanese characters were transcoded to garbled characters when using US-ASCII as the encoding.

```
<?xml version="1.0" encoding="US-ASCII" ?>
<!-- Produced from SAS data using the SAS Clinical Standard's Toolkit. -->
- <ODM xmlns="http://www.cdisc.org/ns/odm/v1.3" xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0"
  data:DatasetXMLVersion="1.0.0">
- <ClinicalData StudyOID="STUDY1" MetaDataVersionOID="MDV.Export_Dataset_XML">
  - <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="1">
    <ItemData ItemOID="IT.AE.STUDYID" Value="CDISC01" />
    <ItemData ItemOID="IT.AE.DOMAIN" Value="AE" />
    <ItemData ItemOID="IT.AE.USUBJID" Value="1003" />
    <ItemData ItemOID="IT.AE.AETERM" Value="i -c g| + c J" />
    <ItemData ItemOID="IT.AE.AEBODSYS" Value="g% g5 g3;g51g| | f0| " />
    <ItemData ItemOID="IT.AE.AESEV" Value="h;=c J" />
  </ItemGroupData>
</ClinicalData>
</ODM>
```

Display 13. The AE.xml with US-ASCII encoding

The three Japanese characters are displayed correctly when using Shift-JIS as the encoding.

```

<?xml version="1.0" encoding="Shift-JIS" ?>
<!-- Produced from SAS data using the SAS Clinical Standard's Toolkit. -->
- <ODM xmlns="http://www.cdisc.org/ns/odm/v1.3" xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0"
  data:DatasetXMLVersion="1.0.0">
- <ClinicalData StudyOID="STUDY1" MetaDataVersionOID="MDV.Export_Dataset_XML">
- <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="1">
  <ItemData ItemOID="IT.AE.STUDYID" Value="CDISC01" />
  <ItemData ItemOID="IT.AE.DOMAIN" Value="AE" />
  <ItemData ItemOID="IT.AE.USUBJID" Value="1003" />
  <ItemData ItemOID="IT.AE.AETERM" Value="頭が痛い" />
  <ItemData ItemOID="IT.AE.AEBODSYS" Value="神経系統病気" />
  <ItemData ItemOID="IT.AE.AESEV" Value="軽い" />
</ItemGroupData>
</ClinicalData>
</ODM>

```

Display 14. The AE.xml with Shift-JIS encoding

In a word, if you are not sure about the languages in the SAS dataset, UTF-8 is an advisable choice.

## THE SAS MACROS CALLED BY CDI TO CREATE DATASET-XML

### DATASET-XML AND DEFINE-XML

Dataset-XML defines a standard format for transporting tabular data set data in XML. The Define-XML file that describes the SAS data sets must contain metadata information about all SAS data sets and all variables to be converted. The Dataset-XML files by themselves do not have any information about the SAS data sets (name and label) or the SAS variables (name, label, data type, length, and display format). Each Dataset-XML file contains data for a single data set, but a single Define-XML file describes all the data sets included in the folder.

### %DATASETXML\_WRITE MACRO

The CDI calls the %datasetxml\_write macro to create Dataset-XML files from a library of SAS data sets:

```

/* Define the libname statements for the SAS data sets, the input define file and the
output location */
libname srcdata "&studyRootPath/data";
filename srcmeta "&studyRootPath/sourcexml/define.xml";
libname xmldata "&studyOutputPath/sourcexml";

/* Call the write macro */
%datasetxml_write(
  _cstSourceDataSets=srcdata.DM,
  _cstOutputLibrary=xmldata,
  _cstSourceMetadataDefineFileRef=srcmeta,
  _cstZip=Y,
  _cstDeleteAfterZip=N,
  _cstCheckLengths=Y,
  _cstOutputEncoding=UTF-8,
  _cstHeaderComment=%nrquote(Produced from SAS data using the SAS Clinical Standards
Toolkit.);

```

**Parameters:**

<code>_cstSourceDataSets</code>	A list of source data sets to convert.
<code>_cstOutputLibrary</code>	The libref of the output data folder/library in which to create the dataset-XML files.
<code>_cstSourceMetadataDefineFileRef</code>	The libref of the source metadata folder/library.
<code>_cstZip</code>	Zip the Dataset-XML file to a zip file in the same folder and with the same name as the Define-XML file. (Default: N)
<code>_cstDeleteAfterZip</code>	Delete the Dataset-XML file after it is zipped (Default: N)
<code>_cstCheckLengths</code>	The actual value lengths of variables with <code>DataType=text</code> are checked against the lengths as defined in the metadata. If the lengths as defined in the metadata are too short, a warning is written to the log file. (Default: N)
<code>_cstOutputEncoding</code>	The XML encoding to use for the Dataset-XML files to create (Default=UTF-8)
<code>_cstHeaderComment</code>	The short comment that is added to the top of the Dataset-XML file to produce. Default: Produced from SAS data using the SAS Clinical Standards Toolkit

The sas codes used in `%datasetxml_write` macro to specify the encoding of Dataset-XML

```
file _xml&_cstRandom encoding="&_cstOutputEncoding" &_cstLRECL;  
  
%if %sysevalf(%superq(_cstOutputEncoding)=, boolean)=0 %then %do;  
    put '<?xml version="1.0" encoding="' "&_cstOutputEncoding" '"?>';  
%end;  
%else %do;  
    put '<?xml version="1.0"?>';  
%end;
```

About the encoding option in file statement:

**ENCODING='encoding-value':** Specifies the encoding to use when writing to the output file. The value for ENCODING= indicates that the output file has a different encoding from the current SAS session encoding.

When you write data to the output files, SAS transcodes the data from the SAS session encoding to the specified encoding.

If you do not specify the encoding in file statement, SAS uses the current SAS session encoding as default.

## DATASET-XML TOOLS

Display 15 is the summary for the tools that can be used to work with the Dataset-XML files.

Dataset-XML Tool Summary			
Name	Description	Provided By	Links
XPT2DatasetXML	<ul style="list-style-type: none"> <li>Transforms XPT datasets into Dataset-XML datasets</li> <li>Freely available</li> </ul>	XML4Pharma	<ul style="list-style-type: none"> <li>Available under the <a href="#">Smart SDS-XML View project on source forge</a></li> </ul>
Smart Dataset-XML Viewer	<ul style="list-style-type: none"> <li>Similar to the SAS Viewer, but with additional functionality</li> <li>Supports working with Define-XML + Dataset-XML files</li> <li>Supports SDTM, SEND, and ADaM data</li> <li>Basic validation</li> <li>Open source</li> </ul>	Univ. Appl. Sciences FH Joanneum Graz - eHealth	<ul style="list-style-type: none"> <li>The application and tutorial is available under the <a href="#">Smart SDS-XML View project on source forge</a></li> <li><a href="#">Youtube video on the Smart Dataset-XML Viewer</a></li> </ul>
EZ Convert	<ul style="list-style-type: none"> <li>Converts Dataset-XML files into SAS datasets</li> <li>Supports Define-XML Version 1 or Version 2</li> <li>Open Source</li> </ul>	<a href="#">@Sally Cassells</a>	<ul style="list-style-type: none"> <li><a href="#">EZConvert Demonstration video</a></li> <li><a href="#">Beta version of EZConvert</a></li> </ul>
SAS Clinical Standards Toolkit	<ul style="list-style-type: none"> <li>Dataset-XML support (writing/reading/validation) will be part of the next release of SAS® Clinical Standards Toolkit. Updated information will be published at the SAS web site.</li> <li>Support for Dataset-XML is available as a pre-production package that contains SAS macros, XML schema files, sample data, and sample programs to support the following functionality:                             <ul style="list-style-type: none"> <li>Creating Dataset-XML files from SAS data sets</li> <li>Creating SAS data sets from Dataset-XML files</li> <li>Validating Dataset-XML files against an XML schema</li> <li>Comparing original SAS data sets with SAS data sets created from Dataset-XML files</li> </ul> </li> <li>These macros are standalone and do not require SAS® Clinical Standards Toolkit.</li> </ul>	SAS Institute Inc.	<ul style="list-style-type: none"> <li><a href="#">SAS Clinical Standards Toolkit</a></li> <li><a href="#">SAS Macros to support Dataset-XML v1.0.0</a></li> </ul>
OpenCDISC v1.5	<ul style="list-style-type: none"> <li>OpenCDISC v1.5 works with Dataset-XML files and Define-XML v2.0</li> </ul>	OpenCDISC	<ul style="list-style-type: none"> <li><a href="#">OpenCDISC.org</a></li> </ul>
R4CDISC	<ul style="list-style-type: none"> <li>R4CDISC package includes functions for reading Dataset-XML and Define-XML files.</li> </ul>	Ippei Akiya	<ul style="list-style-type: none"> <li><a href="#">CRAN project page with downloads</a></li> <li><a href="#">Reference manual</a></li> </ul>

Display 15. The Dataset-XML Tool Summary

## THE LIMITATIONS OF THE OTHER DATASET-XML TOOLS

We can use XPT2DatasetXML and OpenCDISC to create Dataset-XML files from SAS XPT files, however, as we know, the XPT file supports only single byte data, so we still cannot handle multilingual data using these tools.

## SAS ALSO PROVIDES STANDALONE SAS MACROS TO CREATE DATASET-XML FOR FREE

There is a standalone version of the macros that support the CDISC-Dataset XML 1.0 standard. With the standalone SAS macros, we can also create Dataset-XML files that contain multilingual data.

Documentation is available in this file that is part of the ZIP file: SAS-Dataset-XML-v1.0.0-support.pdf (<http://support.sas.com/kb/53/447.html>)

Note: These macros are standalone and do not require SAS® Clinical Standards Toolkit.

## CONCLUSION

Dataset-XML functions as an alternative to SAS Version 5 Transport (XPT) for the transmission of datasets, it removes the SAS XPORT format limitations. For example, the XPT only supports US-ASCII characters, but the Dataset-XML does not have such a limitation, it supports all language encodings supported by XML. With SAS Clinical Data Integration, it is easy and efficient to handle multilingual data in the Dataset-XML files. You just need to know whether the output encoding cover the characters in the SAS datasets. CDI makes it possible to submit the clinical data in non-ASCII characters, such as in Japanese to PMDA.

## REFERENCES

New Dataset-XML Standard v1.0

<http://www.cdisc.org/dataset-xml>

SAS Clinical Data Integration 2.6: User's Guide

<http://support.sas.com/documentation>

SAS® Macros to support Dataset-XML v1.0.0

<http://support.sas.com/kb/53/447.html>

Lex Jansen (2015). SAS® Tools for Working with Dataset-XML files

[www.lexjansen.com/pharmasug/2015/SS/PharmaSUG-2015-SS09-SAS.pdf](http://www.lexjansen.com/pharmasug/2015/SS/PharmaSUG-2015-SS09-SAS.pdf)

## ACKNOWLEDGMENTS

I would like to thank all of my colleagues who reviewed this paper and gave me valuable comments. Special thanks to Jungle Cheng and Han Liu for providing opportunities to study the CDISC related knowledge.

## CONTACT INFORMATION

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

Name: Jing Gao  
Enterprise: SAS Research and Development (Beijing) Co., Ltd.  
Address: Motorola Plaza, No. 1 Wang Jing East Road  
City, State ZIP: Beijing, 100102  
Work Phone: (8610) 83193355-3462  
Fax: (8610) 6310-9130  
E-mail: [Jing.gao@sas.com](mailto:Jing.gao@sas.com)  
Web: [www.sas.com](http://www.sas.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.

Other brand and product names are trademarks of their respective companies.