SAS® TOOLS FOR WORKING WITH DATASET-XML FILES

Mark Lambrecht, Principal Consultant, SAS Health and Life Sciences Global Practice
Lex Jansen, Principal Software Developer @ SAS, CDISC XML Technologies Team

What is Dataset-XML

- Alternative to SAS Version 5 Transport (XPT) format for data sets
- Based on CDISC ODM XML and Define-XML for representation of SDTM, SEND, ADaM or legacy (non-CDISC) tabular data set structures
- Capability to support CDISC data submissions to the FDA
- Based or aligned with Define-XML metadata
- Easy to transform to a data set for analysis (SAS, R, ...)

AE

STUDYID	DOMAIN	USUBJID	AESEQ	AESPID	AETERM	AEMODIFY	AEDECOD	AEBODSYS	AESEV	AESER	AEACN	AEREL	AESTDTC	AEENDTC	AESTDY	AEENDY	AEENR
CDISC01	AE	CDISC01.10...	1	1	AGITATED	AGITATION	Agitation	Psychiatric d...	MILD	N	DOSE NOT ...	POSSIBLY ...	2003-05		3		AFTER
CDISC01	AE	CDISC01.10...	2	2	ANXIETY		Anxiety	Psychiatric d...	MODERATE	N	DOSE NOT ...	POSSIBLY ...	2003-05-13		15		AFTER
CDISC01	AE	CDISC01.10...	3	3	DECREASE...		Decreased a...	Metabolism ...	MILD	N	DOSE NOT ...	POSSIBLY ...	2003-08-19	2003-09-15	113	140	
CDISC01	AE	CDISC01.10...	1	1	DIARRHEA		Diarrhoea	Gastrointesti...	MILD	N	DOSE NOT ...	NOT RELAT...	2004-01-06		84		AFTER
CDISC01	AE	CDISC01.10...	2	2	HEMORRH...		Haemorrhoids	Gastrointesti...	MODERATE	N	DOSE NOT ...	NOT RELAT...	2004-01-06		84		AFTER
CDISC01	AE	CDISC01.10...	3	3	HEADACHE		Headache	Nervous syst...	MILD	N	DOSE NOT ...	NOT RELAT...	2004-01-27		105		AFTER
CDISC01	AE	CDISC01.10...	4	4	VOMIT	VOMITING	Vomiting	Gastrointesti...	MODERATE	N	DRUG INTE...	POSSIBLY ...	2004-02-03	2004-02-03	112	112	
CDISC01	AE	CDISC01.10...	5	5	VOMIT	VOMITING	Vomiting	Gastrointesti...	SEVERE	Y	DRUG INTE...	POSSIBLY ...	2004-02-04	2004-02-09	113	118	
CDISC01	AE	CDISC01.20...	1	1	ANXIETY		Anxiety	Psychiatric d...	SEVERE	N	DOSE NOT ...	POSSIBLY ...	2003-10-16	2003-10-20	17	21	
CDISC01	AE	CDISC01.20...	2	5	LEFT KNEE ...		Arthralgia	Musculoskel...	SEVERE	N	DRUG WITH...	NOT RELAT...	2004-02-02		126		AFTER
CDISC01	AE	CDISC01.20...	3	3	CONSTIPAT...		Constipation	Gastrointesti...	MODERATE	N	DOSE NOT ...	NOT RELAT...	2003-12-25		87		AFTER
CDISC01	AE	CDISC01.20...	4	4	TIREDNESS		Fatigue	General diso...	SEVERE	N	DOSE NOT ...	POSSIBLY ...	2003-12-25		87		AFTER
CDISC01	AE	CDISC01.20...	5	2	NAUSEA IN...		Nausea	Gastrointesti...	SEVERE	N	DOSE NOT ...	POSSIBLY ...	2003-10-16	2003-10-20	17	21	
CDISC01	AE	CDISC01.20...	1	3	LIGHTHEAD...		Dizziness	Nervous syst...	MILD	N	DOSE NOT ...	NOT RELAT...	2004-02-26	2004-02-26	140	140	
CDISC01	AE	CDISC01.20...	2	1	MUSCLE S...		Muscle spas...	Musculoskel...	MILD	N	DOSE NOT ...	NOT RELAT...	2004-01-05		88		AFTER
CDISC01	AE	CDISC01.20...	3	2	PALPITATIO...		Palpitations	Cardiac diso...	MILD	N	DOSE NOT ...	NOT RELAT...	2004-01-05		88		AFTER

SAS Version 5 Transport (XPT) limitations

Limitations of SAS Version 5 Transport (XPT)

Technical

- Data set and Variable name length limitation (8)
- Data set and Variable label length limitation (40)
- Character variable data lengths limitation (200)
- Limited data types (Character, Numeric)
- Very limited international character support (only ASCII)

Structural

- Two-dimensional “flat” data structure for hierarchical/multi-relational “round” data
- Lack of robust information model

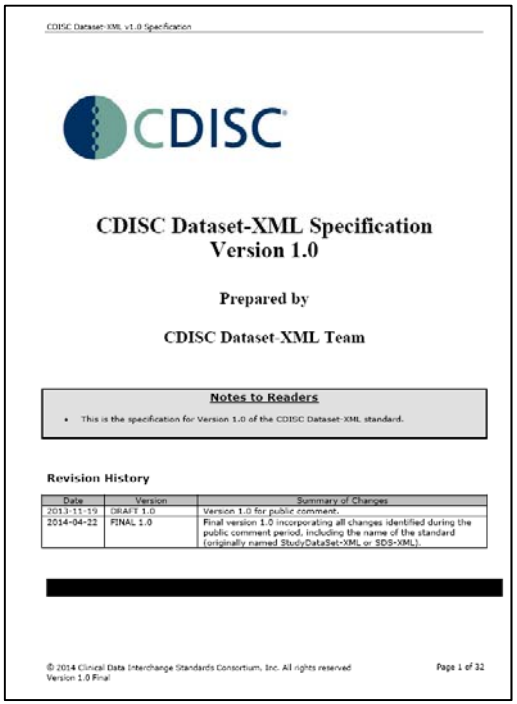
Benefits

- Open, non-proprietary standard without the field width or data set and variable naming restrictions of SAS V5 Transport files
- Supports representation of data relationships, metadata versions and audit trails
  - Note: not all of these will be available in the first release
- Data elements include references to metadata in Define-XML
- Straightforward implementation starting from tabular data in SAS
- Supports FDA goal of encouraging open source reviewer tool development
- Facilitates Validation since both data and metadata share underlying technology
- Enables re-thinking some of the length restrictions in standards

What is Dataset-XML

Status

- Final specification for version 1.0 has been released in April 2014
- Includes sample Define-XML files with associated Define-XML file and XML schema


CDISC Dataset-XML v1.0 Specification

CDISC

CDISC Dataset-XML Specification Version 1.0

Prepared by

CDISC Dataset-XML Team

Notes to Readers

- This is the specification for Version 1.0 of the CDISC Dataset-XML standard.

Revision History

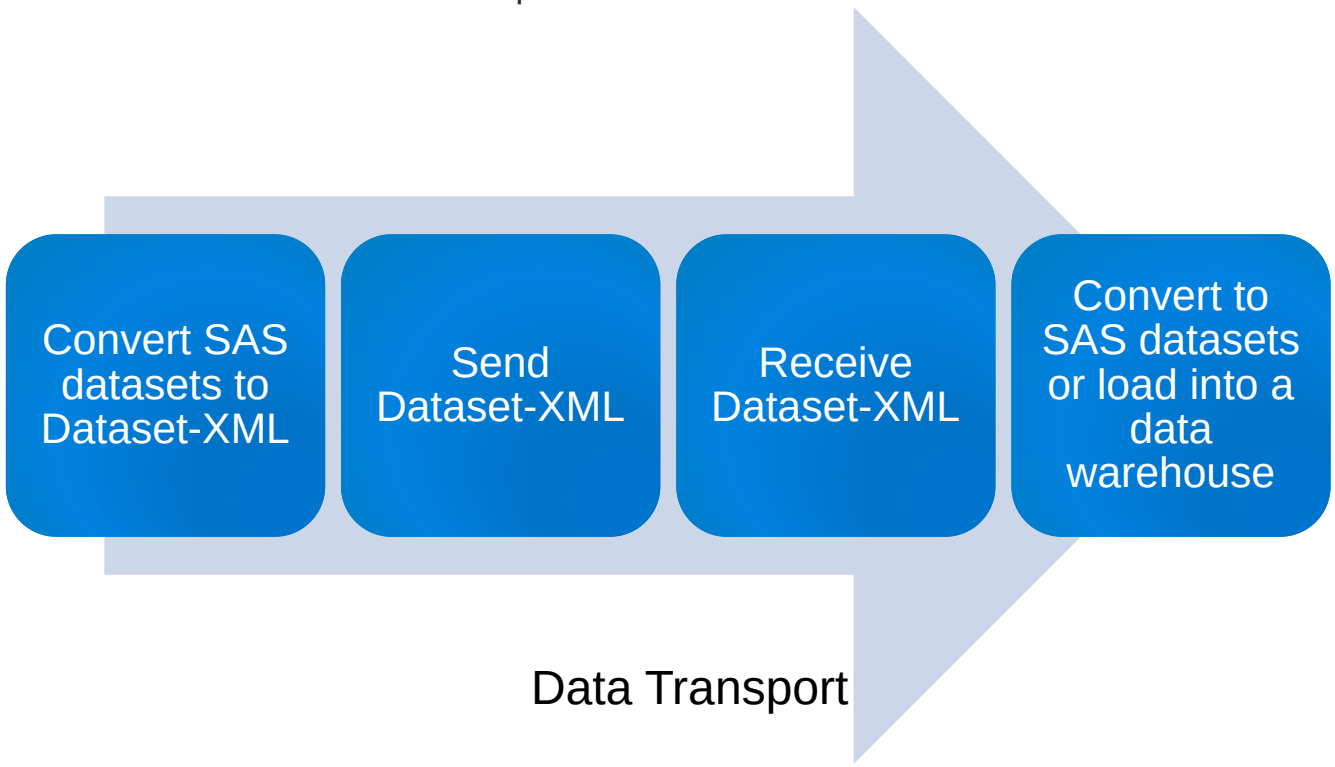
Date	Version	Summary of Changes
2013-11-19	DRAFT 1.0	Version 1.0 for public comment.
2014-04-22	FINAL 1.0	Final version 1.0 incorporating all changes identified during the public comment period, including the name of the standard (originally named StudyDataSet-XML or SDS-XML).

Version 1.0 Final

Page 1 of 32

SAS THE POWER TO KNOW.

- Dataset-XML for Data Transport



What is Dataset-XML

Data Transport

ae.xpt	SAS Xport Transport File
blankcrf.pdf	Adobe Acrobat Document
cm.xpt	SAS Xport Transport File
complexalgorithms.pdf	Adobe Acrobat Document
da.xpt	SAS Xport Transport File
define.xml	XML Document
define2-0-0.xsl	XSLT Stylesheet
dm.xpt	SAS Xport Transport File
ds.xpt	SAS Xport Transport File
eg.xpt	SAS Xport Transport File
ex.xpt	SAS Xport Transport File
ie.xpt	SAS Xport Transport File
lb.xpt	SAS Xport Transport File
mh.xpt	SAS Xport Transport File
pe.xpt	SAS Xport Transport File

ae.xml	XML Document
blankcrf.pdf	Adobe Acrobat Document
cm.xml	XML Document
complexalgorithms.pdf	Adobe Acrobat Document
da.xml	XML Document
define.xml	XML Document
define2-0-0.xsl	XSLT Stylesheet
dm.xml	XML Document
ds.xml	XML Document
eg.xml	XML Document
ex.xml	XML Document
ie.xml	XML Document
lb.xml	XML Document
mh.xml	XML Document
pe.xml	XML Document

```
<def:leaf ID="LF.AE" xlink:href="ae.xpt">
  <def:title>ae.xpt</def:title>
</def:leaf>
```

```
<def:leaf ID="LF.AE" xlink:href="ae.xml">
  <def:title>ae.xml</def:title>
</def:leaf>
```

Dataset-XML | and Define-XML (data and metadata)

```
proc contents data=sdtm.ae varnum;
run;
```

SAS Data

The SAS System

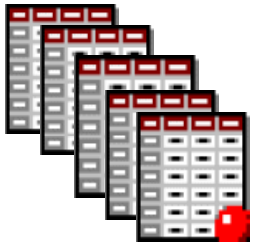
The CONTENTS Procedure

Data Set Name	SDTM.AE	Observations	106
Member Type	DATA	Variables	51
Engine	V9	Indexes	0
Created	09/26/2014 11:01:18	Observation Length	2200
Last Modified	09/26/2014 11:01:18	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	YES
Label	Adverse Events		
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Dataset-XML | and Define-XML (data and metadata)

```
proc contents data=sdtm.ae varnum;
run;
```

SAS Data



Variables in Creation Order

#	Variable	Type	Len	Label
1	STUDYID	Char	40	Study Identifier
2	DOMAIN	Char	8	Domain Abbreviation
3	USUBJID	Char	40	Unique Subject Identifier
4	AESEQ	Num	8	Sequence Number
5	AEGRPID	Char	40	Group ID
6	AEREFID	Char	40	Reference ID
7	AESPID	Char	40	Sponsor-Defined Identifier
8	AETERM	Char	200	Reported Term for the Adverse Event

Dataset-XML and Define-XML

Data set name? Variable names?

```
<ClinicalData
  StudyOID="cdisc01"
  MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
  <!-- Dataset (AE) -->
  <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="1">
    <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
    <ItemData ItemOID="IT.AE.DOMAIN" Value="AE"/>
    <ItemData ItemOID="IT.USUBJID" Value="CDISC01.100008"/>
    <ItemData ItemOID="IT.AE.AESEQ" Value="1"/>
    <ItemData ItemOID="IT.AE.AESPID" Value="1"/>
    <ItemData ItemOID="IT.AE.AETERM" Value="AGITATED"/>
```

Dataset-XML and Define-XML

```
<ItemGroupDef OID="IG.AE" Domain="AE" Name="AE"
  Repeating="Yes" IsReferenceData="No"
  SASDatasetName="AE" Purpose="Tabulation"
  def:Structure="One record per adverse event per subject"
  def:Class="EVENTS" def:ArchiveLocationID="LF.AE">
  <Description>
    <TranslatedText xml:lang="en">Adverse Events</TranslatedText>
  </Description>
  <ItemRef ItemOID="IT.STUDYID" OrderNumber="1" Mandatory="Yes" Key
  <ItemRef ItemOID="IT.AE.DOMAIN" OrderNumber="2" Mandatory="Yes"/>
  <ItemRef ItemOID="IT.USUBJID" OrderNumber="3" Mandatory="Yes" Key
  <ItemRef ItemOID="IT.AE.AESEQ" OrderNumber="4" Mandatory="Yes" Me
  <ItemRef ItemOID="IT.AE.AESPID" OrderNumber="5" Mandatory="No"/>
  <ItemRef ItemOID="IT.AE.AETERM" OrderNumber="6" Mandatory="Yes"/>
```

Dataset-XML and Define-XML

```
<ItemDef OID="IT.AE.AETERM" Name="AETERM" DataType="text" Length="25"
   SASFieldName="AETERM">
  <Description>
    <TranslatedText xml:lang="en">Reported Term for the Adverse Event</Tra
  </Description>
  <def:Origin Type="CRF">
    <def:DocumentRef leafID="LF.blankcrf">
      <def:PDFPageRef PageRefs="21" Type="PhysicalRef"/>
    </def:DocumentRef>
  </def:Origin>
</ItemDef>
```

Dataset-XML and Define-XML

```
<ODM ...
  <Study OID="cdisc01">
    <GlobalVariables>
      <StudyName>CDISC01</StudyName>
      <StudyDescription>CDISC Test Study</StudyDescription>
      <ProtocolName>CDISC01</ProtocolName>
    </GlobalVariables>
    <MetaDataVersion OID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2" ... >

  <ItemGroupDef OID="IG.AE"
    Domain="AE"  Name="AE"  Repeating="Yes" IsReferenceData="No" SASDatasetName="AE"
    Purpose="Tabulation"  def:Structure="One record per adverse event per subject"
    def:Class="EVENTS"  def:ArchiveLocationID="LF.AE">
    <Description><TranslatedText xml:lang="en">Adverse Events</TranslatedText></Description>
    <ItemRef ItemOID="IT.STUDYID" OrderNumber="1" Mandatory="Yes" KeySequence="1"/>
    <ItemRef ItemOID="IT.AE.DOMAIN" OrderNumber="2" Mandatory="Yes"/>
    <ItemRef ItemOID="IT.USUBJID" OrderNumber="3" Mandatory="Yes" KeySequence="2" MethodOID="MT.USUBJID"/>
    ...
    <ItemRef ItemOID="IT.AE.AETERM" OrderNumber="6" Mandatory="No"/>
    <ItemRef ItemOID="IT.AE.AEMODIFY" OrderNumber="7" Mandatory="No"/>
    <ItemRef ItemOID="IT.AE.AEDECOD" OrderNumber="8" Mandatory="Yes" KeySequence="3"/>
    ...
    <def:leaf ID="LF.AE" xlink:href="ae.xml">
      <def:title>ae.xml</def:title>
    </def:leaf>
  </ItemGroupDef>
```

define.xml

```
<ODM ...
  <ClinicalData StudyOID="cdisc01" MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
    <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="1">
      <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
      <ItemData ItemOID="IT.AE.DOMAIN" Value="AE"/>
      <ItemData ItemOID="IT.USUBJID" Value="CDISC01.100008"/>
      ...
      <ItemData ItemOID="IT.AE.AETERM" Value="AGITATION"/>
      <ItemData ItemOID="IT.AE.AEDECOD" Value="Agitation"/>
      ...
    </ItemGroupData>
```

ae.xml

Dataset-XML | Subject Data Example

```
<?xml version="1.0" encoding="UTF-8"?>
<ODM
  xmlns="http://www.cdisc.org/ns/odm/v1.3" xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0"
  FileType="Snapshot" ODMVersion="1.3.2" data:DatasetXMLVersion="1.0.0"
  FileOID="www.cdisc.org.Studycdisc01-Define-XML_2.0.0(IG.AE)"
  PriorFileOID="www.cdisc.org.Studycdisc01-Define-XML_2.0.0"
  Originator="CDISC Dataset-XML Team" CreationDateTime="2014-04-01T09:31:03">
  <ClinicalData
    StudyOID="cdisc01"
    MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
    <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="1">
      <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
      <ItemData ItemOID="IT.AE.DOMAIN" Value="AE"/>
      <ItemData ItemOID="IT.USUBJID" Value="CDISC01.100008"/>
      <ItemData ItemOID="IT.AE.AESEQ" Value="1"/>
      <ItemData ItemOID="IT.AE.AESPID" Value="1"/>
      <ItemData ItemOID="IT.AE.AETERM" Value="AGITATED"/>
      <ItemData ItemOID="IT.AE.AEMODIFY" Value="AGITATION"/>
      <ItemData ItemOID="IT.AE.AEDECOD" Value="Agitation"/>
      <ItemData ItemOID="IT.AE.AEBODSYS" Value="Psychiatric disorders"/>
      <ItemData ItemOID="IT.AE.AESEV" Value="MILD"/>
      <ItemData ItemOID="IT.AE.AESER" Value="N"/>
      <ItemData ItemOID="IT.AE.AEACN" Value="DOSE NOT CHANGED"/>
      <ItemData ItemOID="IT.AE.AEREL" Value="POSSIBLY RELATED"/>
      <ItemData ItemOID="IT.AE.AESTDTC" Value="2003-05"/>
      <ItemData ItemOID="IT.AE.AESTDY" Value="3"/>
      <ItemData ItemOID="IT.AE.AEENRF" Value="AFTER"/>
    </ItemGroupData>
```

	STUDYID	DOMAIN	USUBJID	AESEQ	AESPID	AETERM	AEMODIFY	AEDECOD	AEBODSYS
1	CDISC01	AE	CDISC01.100008	1	1	AGITATED	AGITATION	Agitation	Psychiatric disorders

	AESEV	AESER	AEACN	AEREL	AESTDTC	AEENDTC	AESTDY	AEENDY	AEENRF
1	MILD	N	DOSE NOT CHANGED	POSSIBLY RELATED	2003-05		3		AFTER

SAS Tools for Dataset-XML

SAS Tools for Dataset-XML

Available Now

support.sas.com/rnd/base/cdisc/cst/index.html

CDISC Support

Base SAS

SAS® Clinical Standards Toolkit

- Introduction
- Ordering the Toolkit
- Toolkit Documentation
- Papers
  - Version 1.7
  - Version 1.6
  - Version 1.5
  - Version 1.4 and Earlier
- Discussion Forum

Introduction

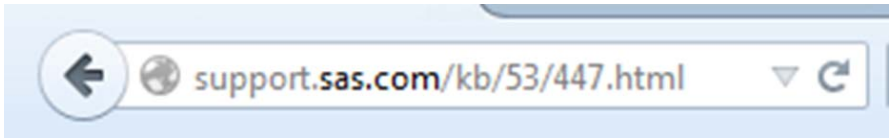
The SAS® Clinical Standards Toolkit provides support of multiple CDISC standards, including SDTM (3.1.2, 3.1.3, and 3.2), CRT-DDS (reading and creating define 1.0 XML files), Define-XML 2.0 (reading and creating define 2.0 XML files), Dataset-XML (creating Dataset-XML files from SAS data sets and creating SAS data sets from Dataset-XML files), ODM (reading and creating 1.3.0 and 1.3.1 XML files), ADaM 2.1, CDASH 1.1, SEND 3.0, and validating XML files against an XML schema file. This tool is the platform used by SAS® to support Health and Life Sciences industry data model standards

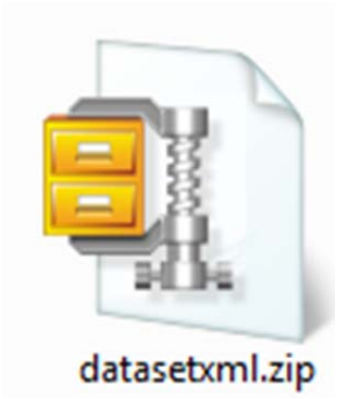
The set of new functionality provided in the recent release of SAS Clinical Standards Toolkit 1.7 includes:

- Introduction of a set of migration tools, previously offered as pre-production software, to help migrate from one version of the SAS Clinical Standards Toolkit to another.
- Support for CDISC CDASH 1.1.
- Enhanced support for creating an initial version of the six SAS source metadata data sets (source_study, source_tables, source_columns, source_codelists, source_values, and source_documents) that serve as input for creating a Define-XML 2.0 file.
- Implementation of the CDISC Dataset-XML 1.0 data standard that can be used to transport CDISC SDTM, SEND, and ADaM data sets for submission of data to the FDA. See Sample 53447: SAS Macros to support Dataset-XML v1.0.0 on support.sas.com for information about a standalone version of the macros that support the CDISC-Dataset XML 1.0 standard.
- Reduced and consolidated validation_master data sets for SDTM 3.1.2, 3.1.3, 3.2, and ADaM 2.1.
- Updated support for CDISC NCI controlled terminology.
- New framework macros have been added, including cstutilmanagecolumnsize, cstutilcomparemetadatasasdefine, cstutilsqlcolumndefinition, cstutilsqlgeneratetable, cstutilfindfixextdasciichars, cstutilregisterctsubtype, and cstutilxmlvalidate.

SAS Tools for Dataset-XML

Available Now





Sample *53447*: SAS® Macros to support Dataset-XML v1.0.0

Details | Downloads | About | Rate It

This package contains macros, XML schema files, sample data, and sample programs to support the following functionality:

- Creating Dataset-XML files from SAS data sets
- Creating SAS data sets from Dataset-XML files
- Validating Dataset-XML files against an XML schema
- Comparing original SAS data sets with SAS data sets created from Dataset-XML files

Documentation is available in this file that is part of the ZIP file: SAS-Dataset-XML-v1.0.0-support.pdf

Users are encouraged to use this new functionality. To help guide future development, post feedback on the SAS in Health Care Related Fields and Clinical Trials community:

http://communities.sas.com/community/support-communities/sas_in_health_care_rela ted_fields

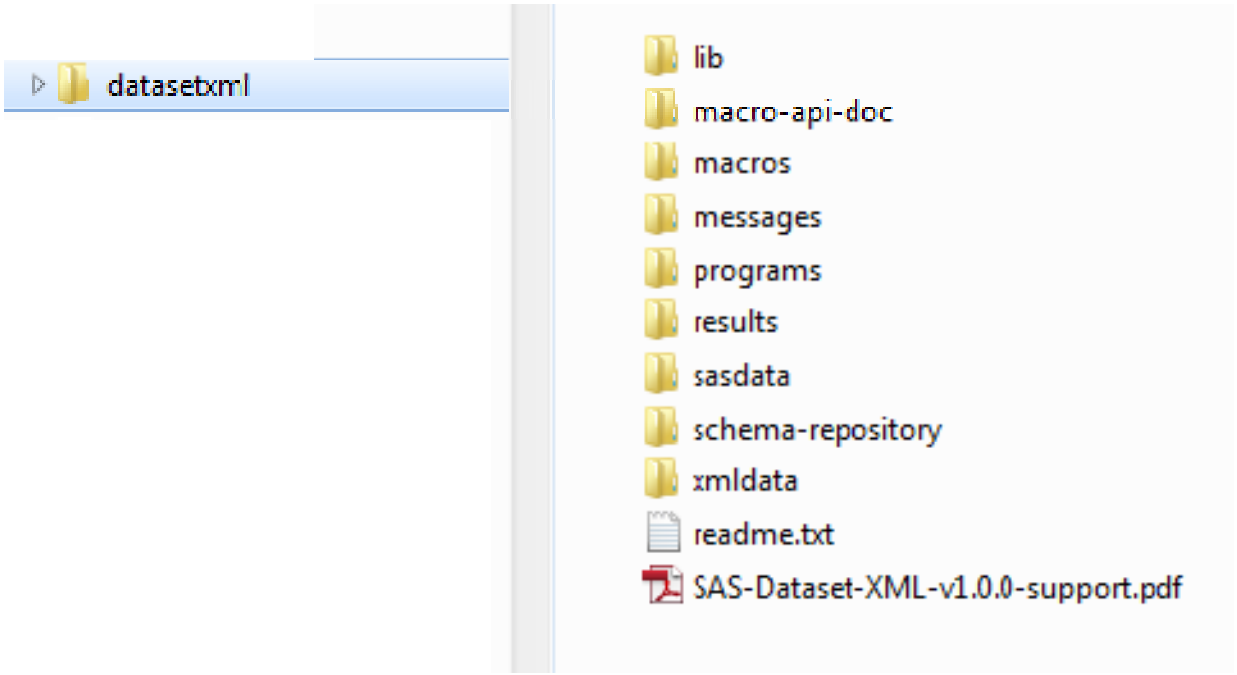
Note: These macros are standalone and do not require SAS® Clinical Standards Toolkit.

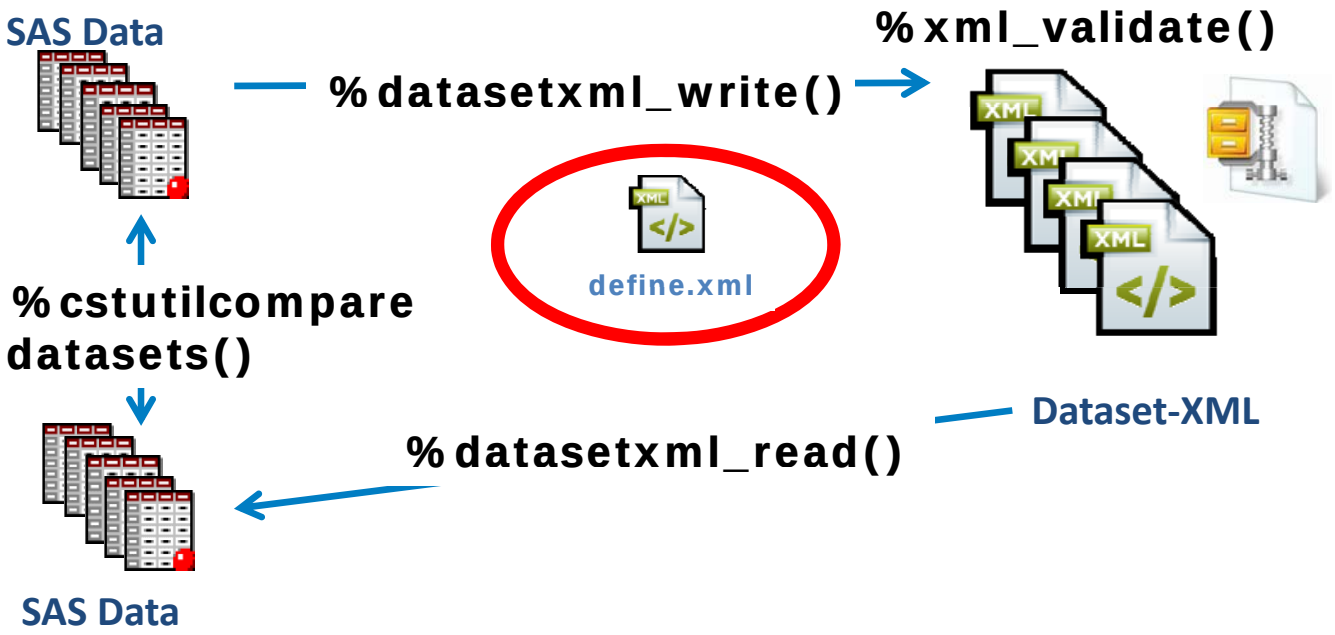
CST 1.7

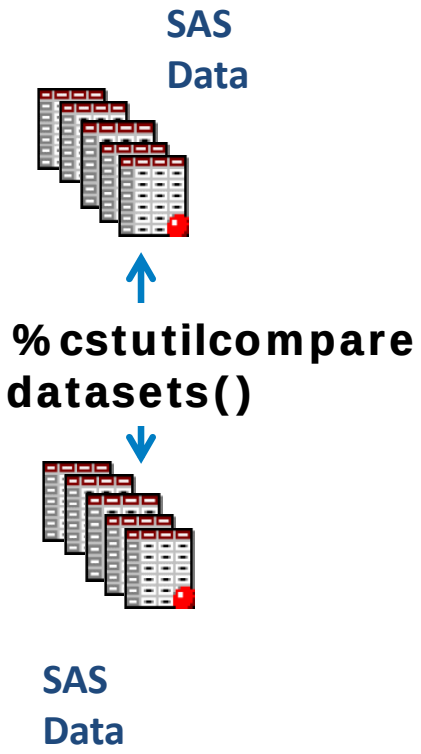
CDI 2.6

SAS Tools for Dataset-XML

Available Now







Expected differences

- Date- and time-related columns may get a different length, since they do not have a length defined in the Define-XML metadata
- Small differences in precision can be expected around the machine precision for numeric variables that represent real numbers.
- Character data that contains leading spaces or trailing spaces may lose the leading and trailing spaces.

Dataset-XML | SAS Tools - Macros

```
*****************************************************************;
* Create SDTM Dataset-XML files                                 *;
*****************************************************************;
libname sdtmdata  "&studyRootPath/sasdata/original/sdtm";
filename defxml "&studyOutputPath/xmldata/sdtm/define.xml";
libname dataxml  "&studyOutputPath/xmldata/sdtm";

%datasetxml_write(
  _cstSourceLibrary=sdtmdata,
  _cstOutputLibrary=dataxml,
  _cstSourceMetadataDefineFileRef=defxml,
  _cstCheckLengths=Y,
  _cstIndent=N,
  _cstZip=Y,
  _cstDeleteAfterZip=N
  );

libname sdtmdata clear;
filename defxml clear;
libname dataxml clear;
```

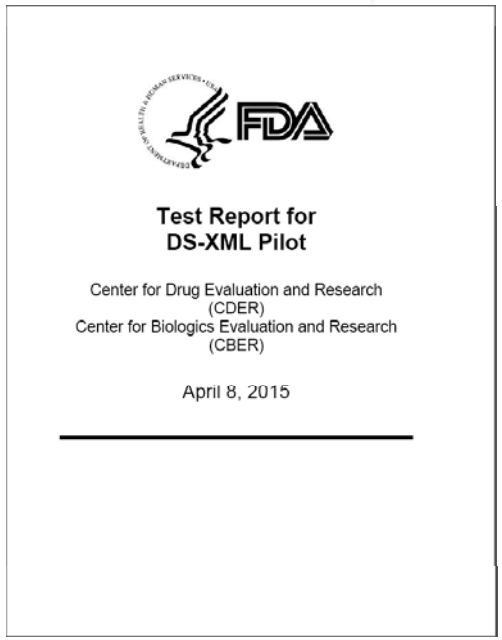
Dataset-XML | SAS Tools - Macros

```
*****************************************************************;
* Create SDTM SAS Data sets                                     *;
*****************************************************************;
libname dataxml  "&studyRootPath/xmldata/sdtm";
libname sdtmdata  "&studyRootPath/sasdata/imported/sdtm";
filename defxml  "&studyOutputPath/xmldata/sdtm/define.xml";

%datasetxml_read(
  _cstSourceDatasetXMLLibrary=dataxml,
  _cstOutputLibrary=sdtmdata,
  _cstSourceMetadataDefineFileRef=defxml,
  _cstdatetimeLength=64,
  _cstAttachFormats=Y
  );

libname dataxml clear;
libname sdtmdata clear;
filename defxml clear;
```

FDA Pilot

Dataset-XML | CDISC Standards

FDA Pilot – Conclusions



Test Report for
DS-XML Pilot

Center for Drug Evaluation and Research
(CDER)
Center for Biologics Evaluation and Research
(CBER)

April 8, 2015

- Test Report for Pilot published on April 8, 2015
- Additional testing will be needed to evaluate cost versus effectiveness as an alternate transport format
- FDA envisions conducting several pilots to evaluate new transport formats before a decision is made to support a new format

Pilot Report: http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm380756.htm

FDA Pilot – Conclusions

- Dataset-XML can transport data and maintain data integrity.
- Dataset-XML transport format can facilitate longer variable names (>8 characters), longer label name (>40 characters) and longer text field (>200 characters).
- Dataset-XML requires stricter encoding in data.
- Dataset-XML requires consistency between datasets and Define-XML.
- Based on the file size observations, Dataset-XML produced much larger file sizes than XPORT, which may impact the Electronic Submissions Gateway (ESG) and may lead to file storage issues.

THANK YOU !

QUESTIONS ?

SAS Institute Inc.
World Headquarters
SAS Campus Drive
Cary, NC 27513 USA
Tel: +1 919 677 8000
Fax: +1 919 677 4444
www.sas.com

Lex Jansen
Principal Software Developer
Health and Life Sciences
Tel: +1 919 531 9860

E-mail: lex.jansen@sas.com

THE POWER TO KNOW.