



Dataset-XML, a new standard for data exchange

Mikkel Traun – Novo Nordisk A/S
Lex Jansen – SAS Institute Inc., USA

Copenhagen, Denmark, June 10th 2015

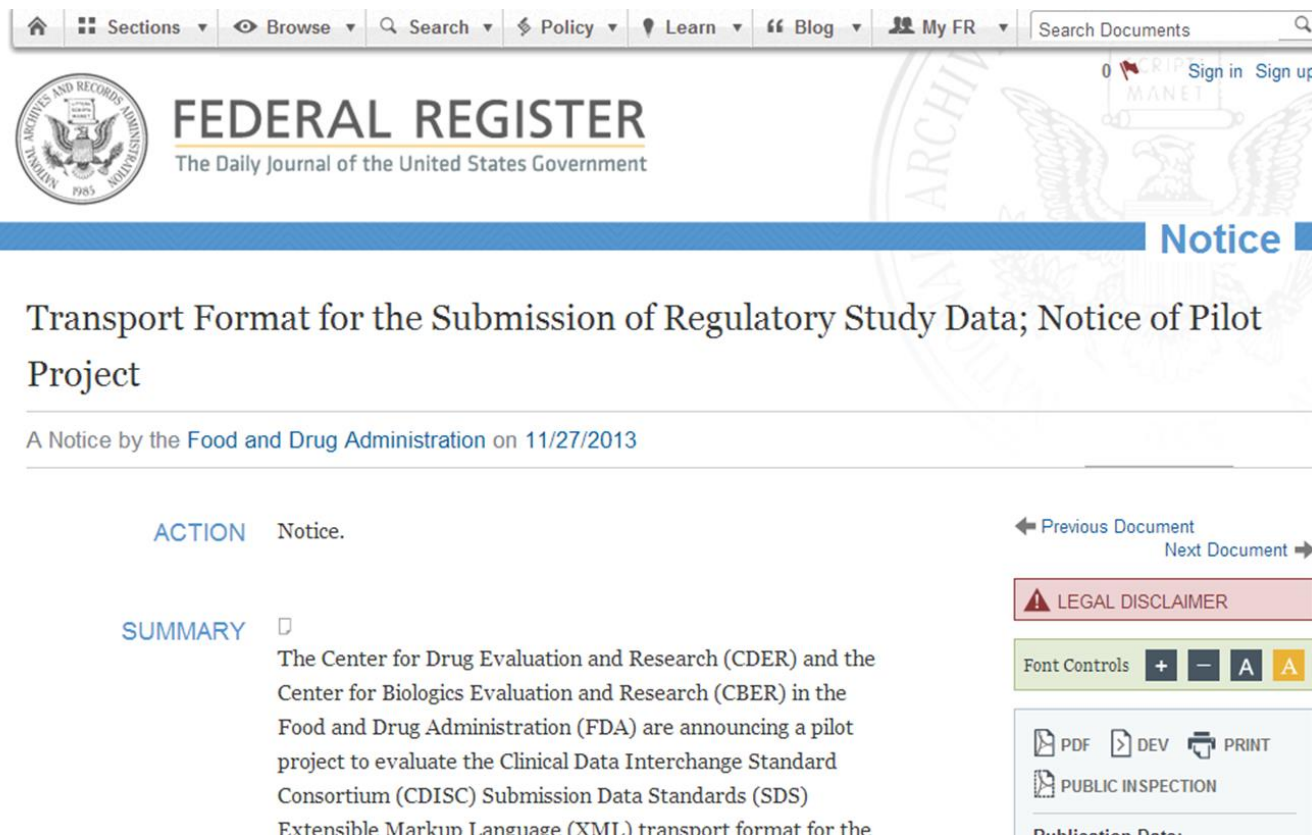


Agenda

- Introduction
 - FDA Pilot Notice
 - Novo Nordisk / SAS partnership
- Why Dataset-XML?
- What is Dataset-XML?
- Dataset-XML and Define-XML
- Tools for Dataset-XML
- FDA Pilot Lessons



Introduction – FDA Pilot Notice *



The screenshot shows the Federal Register website interface. At the top, there is a navigation bar with links for Sections, Browse, Search, Policy, Learn, Blog, and My FR. A search bar is also present. Below the navigation bar is the Federal Register logo and the text "FEDERAL REGISTER The Daily Journal of the United States Government". A blue banner across the page reads "Notice". The main heading of the notice is "Transport Format for the Submission of Regulatory Study Data; Notice of Pilot Project". Below this, it states "A Notice by the Food and Drug Administration on 11/27/2013". The "ACTION" section is labeled "Notice." and the "SUMMARY" section begins with "The Center for Drug Evaluation and Research (CDER) and the Center for Biologics Evaluation and Research (CBER) in the Food and Drug Administration (FDA) are announcing a pilot project to evaluate the Clinical Data Interchange Standard Consortium (CDISC) Submission Data Standards (SDS) Extensible Markup Language (XML) transport format for the". On the right side of the notice, there are links for "Previous Document" and "Next Document", a "LEGAL DISCLAIMER" box, "Font Controls" with plus, minus, and font size buttons, and a box with icons for PDF, DEV, PRINT, and PUBLIC INSPECTION.

Notice

Transport Format for the Submission of Regulatory Study Data; Notice of Pilot Project

A Notice by the Food and Drug Administration on 11/27/2013

ACTION Notice.

SUMMARY 

The Center for Drug Evaluation and Research (CDER) and the Center for Biologics Evaluation and Research (CBER) in the Food and Drug Administration (FDA) are announcing a pilot project to evaluate the Clinical Data Interchange Standard Consortium (CDISC) Submission Data Standards (SDS) Extensible Markup Language (XML) transport format for the

← Previous Document
Next Document →

LEGAL DISCLAIMER

Font Controls    

 PDF  DEV  PRINT
 PUBLIC INSPECTION

Publication Date:

* <https://www.federalregister.gov/articles/2013/11/27/2013-28391/transport-format-for-the-submission-of-regulatory-study-data-notice-of-pilot-project>

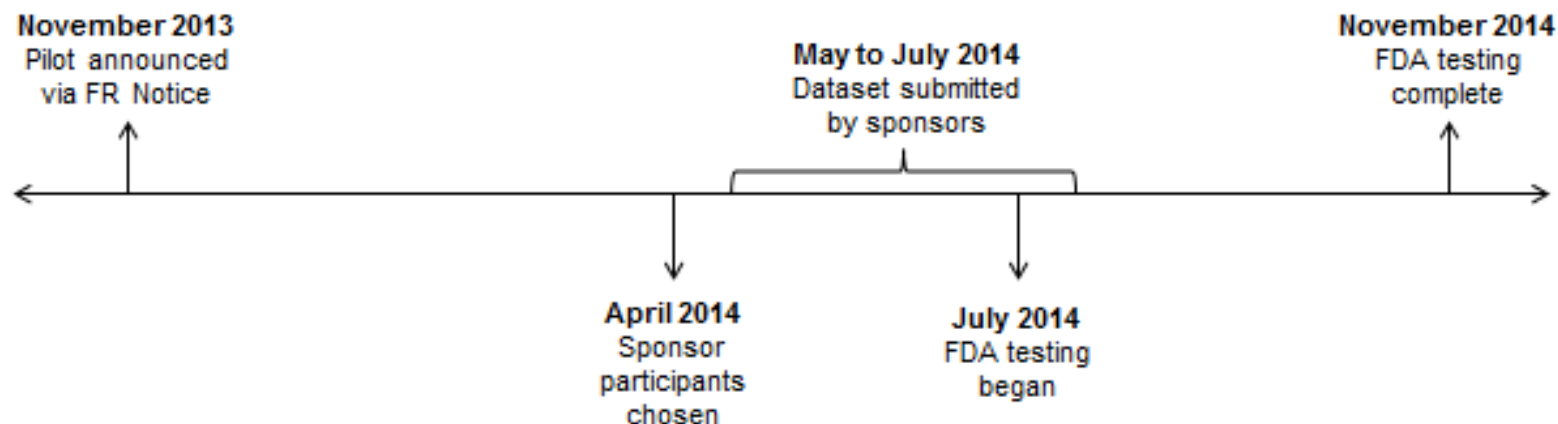
Introduction – FDA Pilot Notice *

- November 2013: FDA announces **Pilot project to evaluate CDISC Dataset-XML** transport format for submission of regulatory study data.
- Current format supported by the FDA (**SAS Transport (XPORT) version 5 format**) has been a reliable exchange format for many years, but is not an extensible modern technology.
- **Invitation to sponsors to participate** in pilot project to evaluate the Dataset-XML transport format.
- Objective: **Test the transport functionality of Dataset-XML**, which included ensuring that data integrity was maintained and that Dataset-XML format would support longer variable names, labels, and text fields.

* Note that the FDA used SDS XML, which is the original name of the standard

* <https://www.federalregister.gov/articles/2013/11/27/2013-28391/transport-format-for-the-submission-of-regulatory-study-data-notice-of-pilot-project>

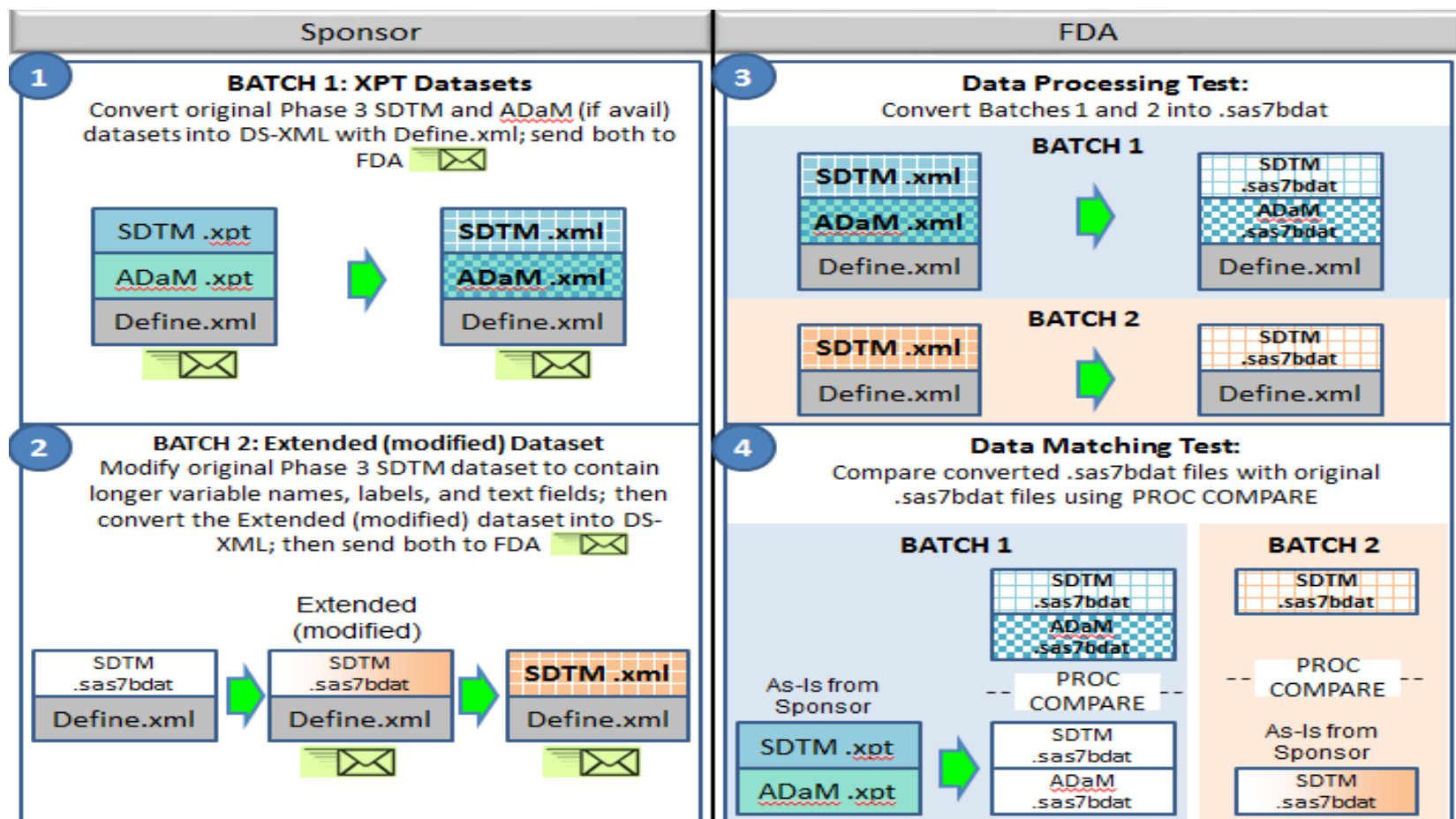
Introduction – FDA Pilot



- 6 sponsors (**including Novo Nordisk**) participated in the pilot.
- The sponsors **re-submitted a previously submitted set of Phase 3 study datasets** in the CDISC Dataset-XML format.
- Novo Nordisk and SAS entered a **partnership**.
- **SAS developed the software** that Novo Nordisk used to create and validate Dataset-XML files from the previously submitted Phase 3 study datasets.

* <https://www.federalregister.gov/articles/2013/11/27/2013-28391/transport-format-for-the-submission-of-regulatory-study-data-notice-of-pilot-project>

Introduction – FDA Pilot



Pilot Report:

<http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm380756.htm>

WHY DATASET-XML?



Dataset-XML Why?

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

What is Dataset-XML?



What is Dataset-XML

- Alternative to SAS Version 5 Transport (XPT) format for data sets
- Based on CDISC ODM 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, ...)

STUDYID	DOMAIN	USUBJID	AESQ	AESPID	AETERM	AEMODIFY	AEDECOD	AEBODSYS	AESEV	AESER	AEACN	AEREL	AESTDTC	AEENDTC	AESTDY	AEENDY	AEENR
CDISC01	AE	CDISC01.10.1	1	1	AGITATED	AGITATION	Agitation	Psychiatric d.	MILD	N	DOSE NOT	POSSIBLY ...	2003-05		3		AFTER
CDISC01	AE	CDISC01.10.2	2	2	ANXIETY		Anxiety	Psychiatric d.	MODERATE	N	DOSE NOT	POSSIBLY ...	2003-05-13		15		AFTER
CDISC01	AE	CDISC01.10.3	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	1	DIARRHEA		Diarrhoea	Gastrointesti.	MILD	N	DOSE NOT	NOT RELAT.	2004-01-06		84		AFTER
CDISC01	AE	CDISC01.10.2	2	2	HEMORRHA		Haemorrhoids	Gastrointesti.	MODERATE	N	DOSE NOT	NOT RELAT.	2004-01-06		84		AFTER
CDISC01	AE	CDISC01.10.3	3	3	HEADACHE		Headache	Nervous syst.	MILD	N	DOSE NOT	NOT RELAT.	2004-01-27		105		AFTER
CDISC01	AE	CDISC01.10.4	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	5	VOMIT	VOMITING	Vomiting	Gastrointesti.	SEVERE	Y	DRUG INTE	POSSIBLY ...	2004-02-04	2004-02-09	113	118	
CDISC01	AE	CDISC01.20.1	1	1	ANXIETY		Anxiety	Psychiatric d.	SEVERE	N	DOSE NOT	POSSIBLY ...	2003-10-16	2003-10-20	17	21	
CDISC01	AE	CDISC01.20.2	2	5	LEFT KNEE		Arthralgia	Musculoskel.	SEVERE	N	DRUG WITH	NOT RELAT.	2004-02-02		126		AFTER
CDISC01	AE	CDISC01.20.3	3	3	CONSTIPAT		Constipation	Gastrointesti.	MODERATE	N	DOSE NOT	NOT RELAT.	2003-12-25		87		AFTER
CDISC01	AE	CDISC01.20.4	4	4	TIREDFNESS		Fatigue	General diso.	SEVERE	N	DOSE NOT	POSSIBLY ...	2003-12-25		87		AFTER
CDISC01	AE	CDISC01.20.5	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	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	1	MUSCLE S.		Muscle spas.	Musculoskel.	MILD	N	DOSE NOT	NOT RELAT.	2004-01-05		88		AFTER
CDISC01	AE	CDISC01.20.3	2	2	PALPITATIO		Palpitations	Cardiac diso.	MILD	N	DOSE NOT	NOT RELAT.	2004-01-05		88		AFTER

What is Dataset-XML

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
- Harmonized with BRIDG, CDISC Controlled Terminology
- 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



[ABOUT CDISC](#) | [STANDARDS & INNOVATIONS](#) | [RESOURCES](#) | [NEWS](#) | [EDUCATION & EVENTS](#) | [MEMBERSHIP](#) | [MEMBERS ONLY](#)

Dataset-XML

New Dataset-XML Standard v1.0 Now Available

The CDISC XML Technologies team is pleased to announce the release of the Dataset-XML v1.0 specification for production use. Dataset-XML, which was released for comment under the name "StudyDataSet-XML" but was renamed to avoid confusion with the CDISC SDS team, is a new standard used to exchange study datasets in an XML format. The purpose of Dataset-XML is to support the interchange of tabular data for clinical research applications using ODM-based XML technologies. The Dataset-XML model is based on the CDISC Operational Data Model (ODM) standard and should follow the metadata structure defined in the CDISC Define-XML standard.

CDISC Dataset-XML v1.0 Specification



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).

© 2014 Clinical Data Interchange Standards Consortium, Inc. All rights reserved.
Version 1.0 Final

Page 1 of 32

Dataset-XML Tools

<http://wiki.cdisc.org/display/PUB/CDISC+Dataset-XML+Resources>



PUBLIC

CDISC Dataset-XML Resources

 [Watch](#)

 [Share](#)

 [Tools](#) ▾

👤 3 Added by Sam Hume, last edited by Sam Hume on Jul 17, 2014 ([view change](#)) [show comment](#)

Introduction

This page contains information on many of the tools currently available or under development for use with Dataset-XML files. Dataset-XML is a new CDISC XML Technologies standard developed as a data transport format for representing SDTM, SEND, ADaM, and legacy datasets. Dataset-XML functions as an alternative to SAS Version 5 Transport (XPT) for the transmission of datasets. More information can be found on this standard at the [CDISC web site](#).

Dataset-XML Tool Summary

Name	Description	Provided By	Links
------	-------------	-------------	-------

Dataset-XML Tools

<http://wiki.cdisc.org/display/PUB/CDISC+Dataset-XML+Resources>

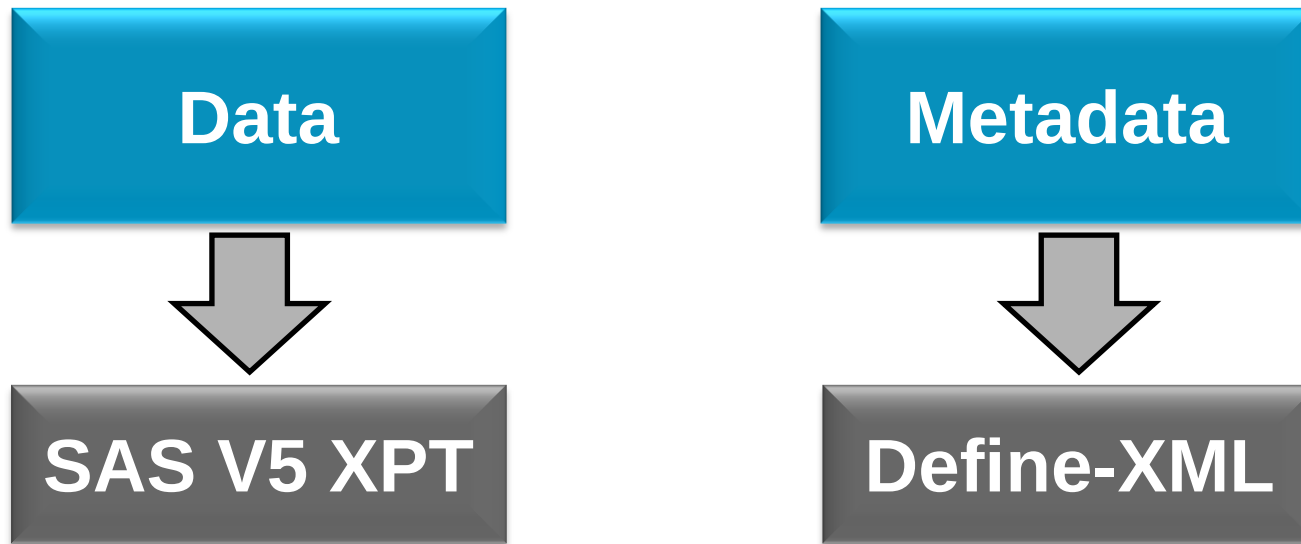
Dataset-XML Tool Summary

Name	Description	Provided By	Links
XPT2DatasetXML	- Transforms XPT datasets into Dataset-XML datasets - Freely available	XML4Pharma	Available under the Smart SDS-XML View project on source forge
Smart Dataset-XML Viewer	- Similar to the SAS Viewer, but with additional functionality - Supports working with Define-XML + Dataset-XML files - Supports SDTM, SEND, and ADaM data - Basic validation - Open source	Univ. Appl. Sciences FH Joanneum Graz - eHealth	- The application and tutorial is available under the Smart SDS-XML View project on source forge Youtube video on the Smart Dataset-XML Viewer
EZ Convert	- Converts Dataset-XML files into SAS datasets - Supports Define-XML Version 1 or Version 2 - Open Source	@Sally Cassells	EZConvert Demonstration video Beta version of EZConvert
SAS Clinical Standards Toolkit	- 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. - 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: - 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 - These macros are standalone and do not require SAS® Clinical Standards Toolkit.	SAS Institute Inc.	SAS Clinical Standards Toolkit SAS Macros to support Dataset-XML v1.0.0
OpenCDISC v1.5	OpenCDISC v1.5 works with Dataset-XML files and Define-XML v2.0	OpenCDISC	OpenCDISC.org
R4CDISC	R4CDISC package includes functions for reading Dataset-XML and Define-XML files.	Ippei Akiya	CRAN project page with downloads Reference manual

What is Dataset-XML

Data and Metadata

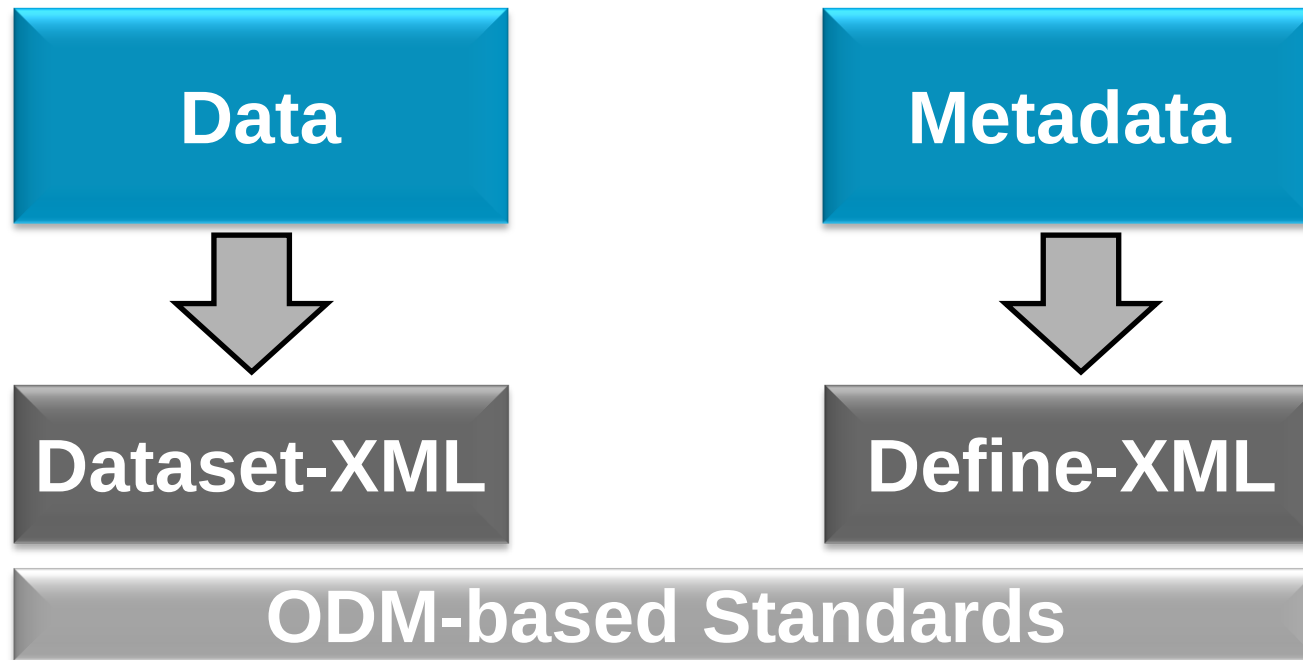
- Data and Metadata in Submissions Today



What is Dataset-XML

Data and Metadata

- Data and Metadata in Submissions Tomorrow



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

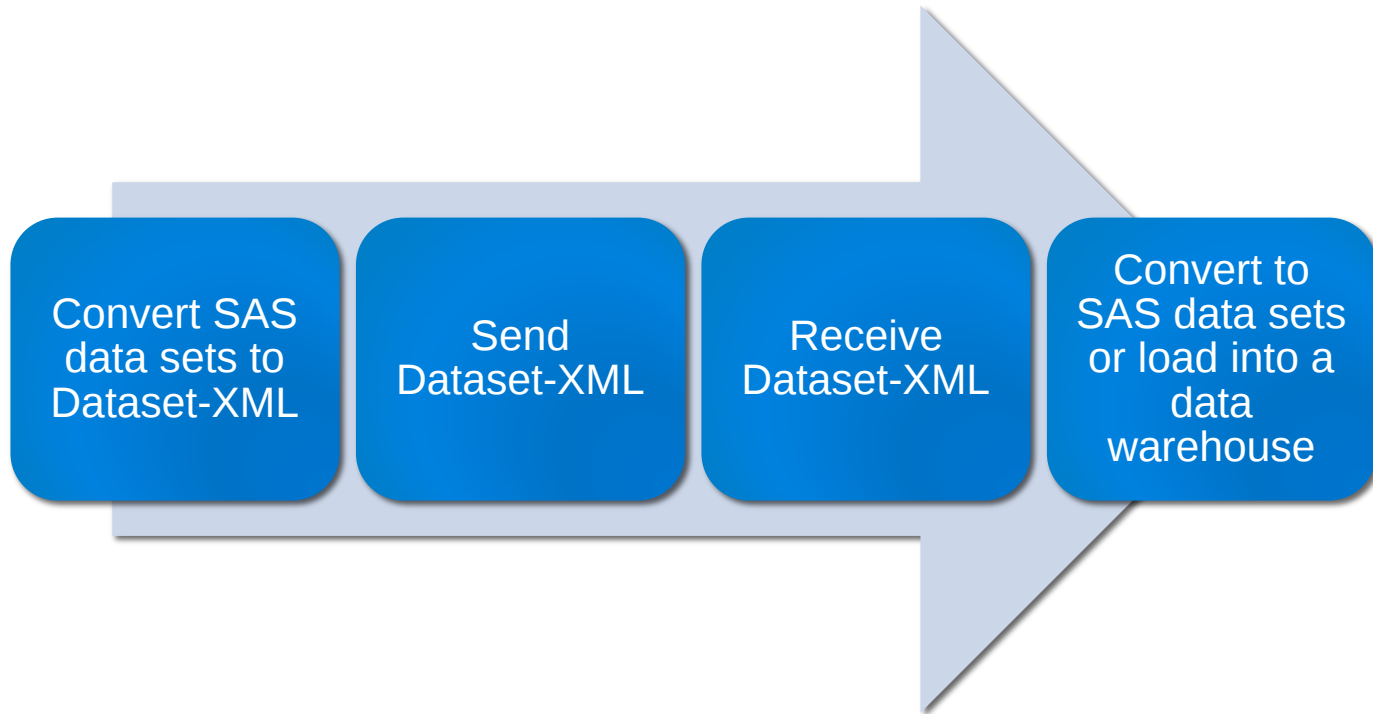
 ae.xml	XML Document
 blankcrf.pdf	Adobe Acrobat Document
 cm.xml	XML Document
 complexalgorithms.pdf	Adobe Acrobat Document
 da.xml	XML Document
 define.xml	XML Document
 define2-0-0.xsl	XSLT Stylesheet
 dm.xml	XML Document
 ds.xml	XML Document
 eg.xml	XML Document
 ex.xml	XML Document
 ie.xml	XML Document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

What is Dataset-XML

Data Transport



Data Transport

Dataset-XML and Define-XML



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

```
proc contents data=sdm.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</TranslatedText>
  </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

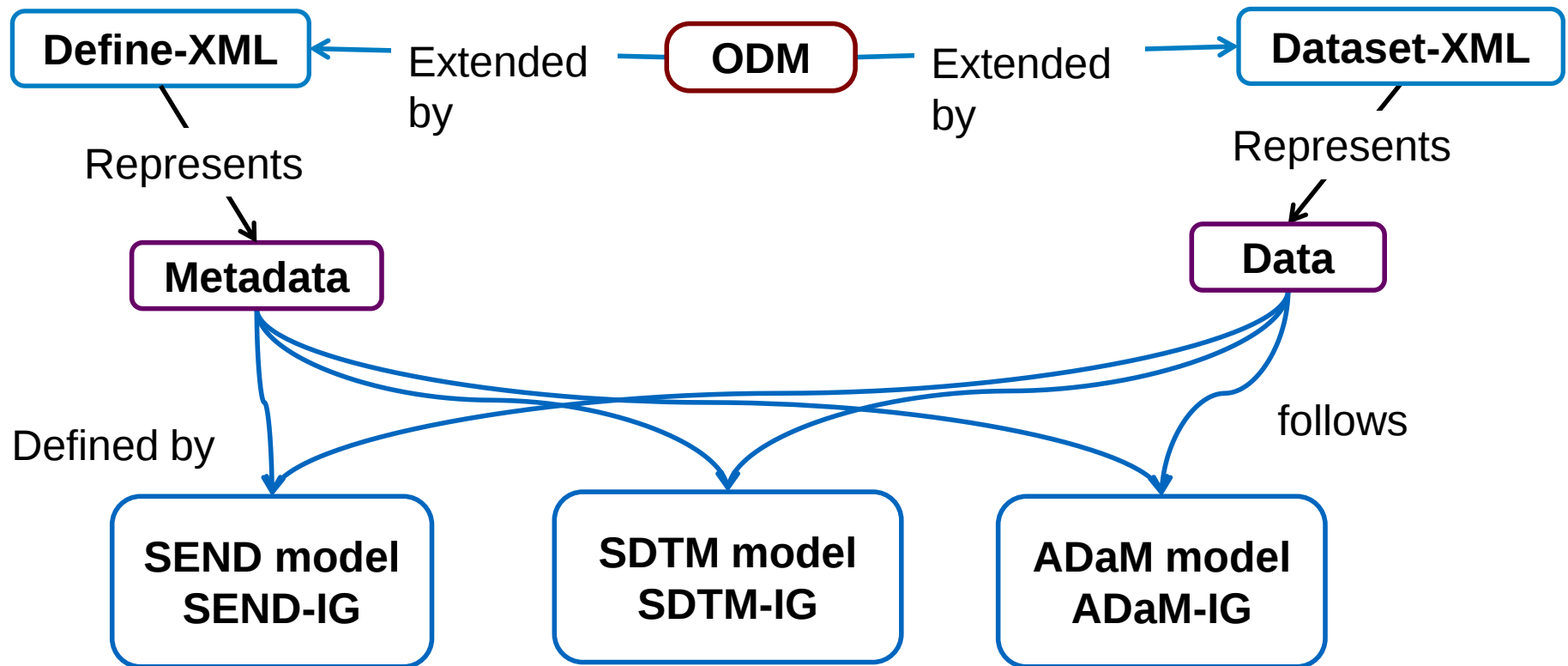
```
<ODM ...  
  <ClinicalData StudyOID="cdisc01" MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">  
    <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="1">  
      <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>  
      <ItemData ItemOID="IT.AE.DOMAIN" Value="AE"/>  
      <ItemData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

ae.xml

What is Dataset-XML

Data and Metadata

Relationship of Dataset-XML to other CDISC Standards



SAS Tools for Dataset-XML



SAS Tools for Dataset-XML

Available Now



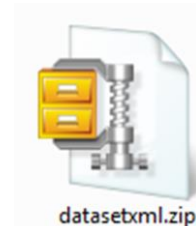
Sample 53447: SAS® Macros to support Dataset-XML v1.0.0

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:

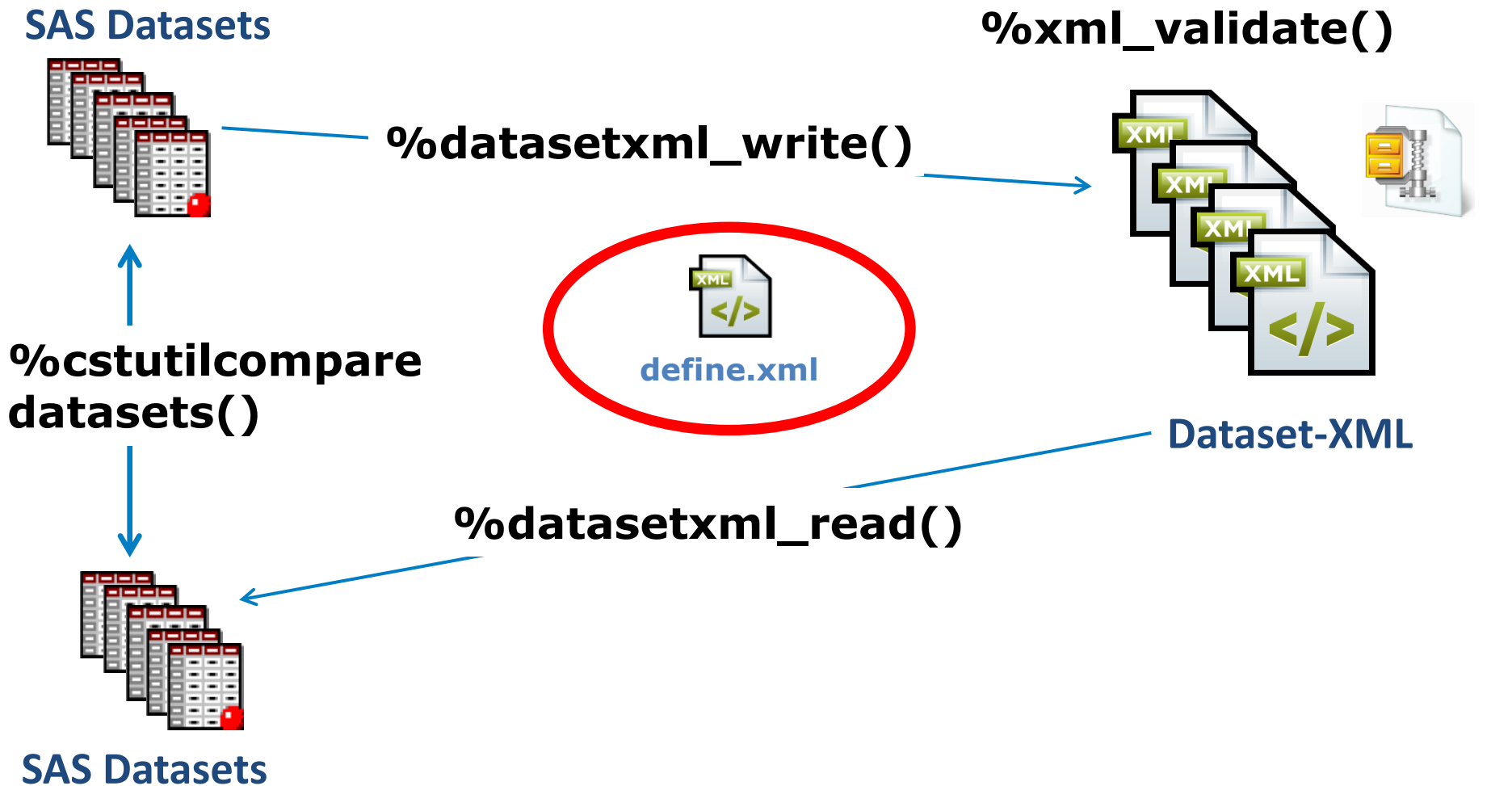
http://communities.sas.com/community/support-communities/sas_in_health_care_related_fields

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

CST 1.7

CDI 2.6

Dataset-XML SAS Tools - Macros



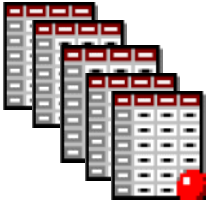
SAS® Tools for Working with Dataset-XML files:

<http://www.lexjansen.com/pharmasug/2015/SS/PharmaSUG-2015-SS09-SAS.pdf>

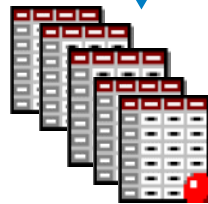
<http://www.lexjansen.com/pharmasug/2015/SS/PharmaSUG-2015-SS09-SAS-PPT.pdf>

Dataset-XML SAS Tools

SAS Datasets



**%cstutilcompare
datasets()**



SAS Datasets

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

FDA Pilot



Dataset-XML 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

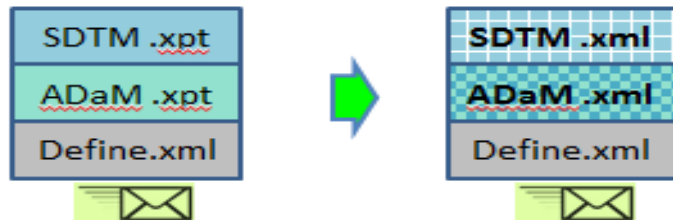
- Test Report for Pilot published on April 8, 2015
- Additional testing will be needed to evaluate cost versus effectiveness as an alternate transport format
- FDA envisions conducting several pilots to evaluate new transport formats before a decision is made to support a new format

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

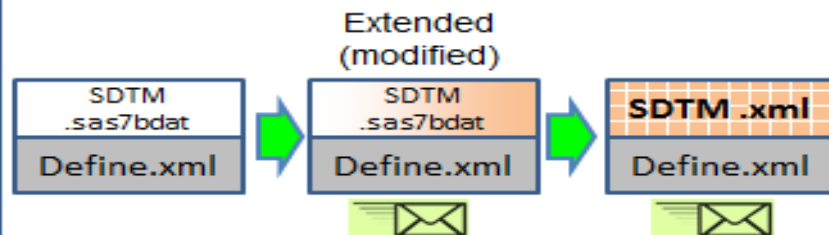
Dataset-XML FDA Pilot – Testing

Sponsor

- 1 BATCH 1: XPT Datasets**
Convert original Phase 3 SDTM and ADaM (if avail) datasets into DS-XML with Define.xml; send both to FDA

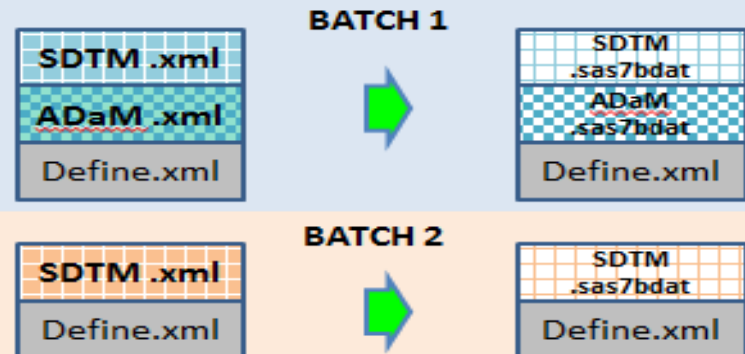


- 2 BATCH 2: Extended (modified) Dataset**
Modify original Phase 3 SDTM dataset to contain longer variable names, labels, and text fields; then convert the Extended (modified) dataset into DS-XML; then send both to FDA

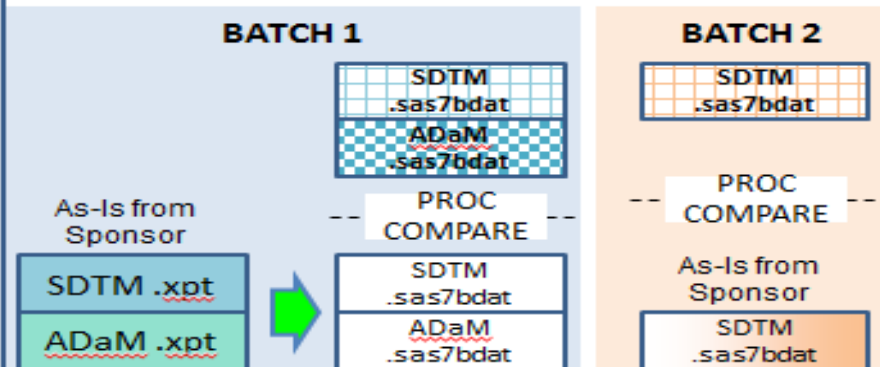


FDA

- 3 Data Processing Test:**
Convert Batches 1 and 2 into .sas7bdat



- 4 Data Matching Test:**
Compare converted .sas7bdat files with original .sas7bdat files using PROC COMPARE



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

Dataset-XML FDA Pilot – some challenges

- Memory issue in the SAS xmlv2 libname engine when processing huge XML files
(Hotfix available:
<http://support.sas.com/kb/54/333.html>)
- Data conversion issues with two sponsors' datasets due to the following SAS error: "Some code points did not transcode." This was caused due to an incorrect XML [encoding](#) of the XML files, which contained non ASCII characters (*"MS Curly quotes"*)

Dataset-XML 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 **strict 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.

Dataset-XML FDA Pilot – some challenges – File size

	SAS (compress)	XPT	XML	ZIP
LB	301.51 MB	636.07 MB	1.75 GB	52.91 MB
QS	432.08 MB	776.73 MB	2.04 GB	53.68 MB
SUPPLB	338.98 MB	717.81 MB	1.79 GB	29.25 MB
SUPPQS	39.23 MB	37.28 MB	214.05 MB	3.73 MB

Thank you!

Questions?



Backup Slides



Dataset-XML FDA Pilot – some challenges - Encoding

- An XML document starts with an optional XML declaration:
`<?xml version="1.0" encoding="UTF-8" ?>`
- The XML declaration is the very first statement of the XML document
- Leaving out the encoding means: UTF-8
- UTF-8 is a superset of ASCII; the first 128 characters of UTF-8 are identical to (7-bit) ASCII
- The first 256 codes of UTF-8 are identical to ISO 8859-1

Dataset-XML FDA Pilot – some challenges - Encoding



Windows Latin-1
(Code page 1252)
is not the same as
ISO 8859-1
(ISO Latin 1)

ISO
8859-1

	-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-		!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/	
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	¢	£	¤	¥	¦	§	¨	©	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	

Dataset-XML FDA Pilot – some challenges - Encoding



Windows Latin-1

This causes issues when
this gets encoded as UTF-8
Be careful when copying
from Excel or Word ...

You may want to review the
"AutoCorrect" options in
Word, Excel and
PowerPoint.

	-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	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
	0030	0031	0032	0033	0034	0035	0036	0037	0038	0039	003A	003B	003C	003D	003E	003F
4-	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	0040	0041	0042	0043	0044	0045	0046	0047	0048	0049	004A	004B	004C	004D	004E	004F
5-	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
	0050	0051	0052	0053	0054	0055	0056	0057	0058	0059	005A	005B	005C	005D	005E	005F
6-	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
	0060	0061	0062	0063	0064	0065	0066	0067	0068	0069	006A	006B	006C	006D	006E	006F
7-	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
	0070	0071	0072	0073	0074	0075	0076	0077	0078	0079	007A	007B	007C	007D	007E	007F
80	€		/	f	"	...	†	‡	~	%	š	<	Œ		ž	
	20AC		201A	0192	201E	2026	2020	2021	02C6	2030	0160	2039	0152		017D	
90		\	/	"	"	•	—	—		™	š	>	œ		ž	ÿ
		2018	2019	201C	201D	2022	2013	2014	02DC	2122	0161	203A	0153		017E	0178
A-		ı	ç	£	¤	¥	¦	§	¨	©	ª	«	¬	-	®	¯
	00A0	00A1	00A2	00A3	00A4	00A5	00A6	00A7	00A8	00A9	00AA	00AB	00AC	00AD	00AE	00AF
B-	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
	00B0	00B1	00B2	00B3	00B4	00B5	00B6	00B7	00B8	00B9	00BA	00BB	00BC	00BD	00BE	00BF
C-	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
	00C0	00C1	00C2	00C3	00C4	00C5	00C6	00C7	00C8	00C9	00CA	00CB	00CC	00CD	00CE	00CF
D-	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
	00D0	00D1	00D2	00D3	00D4	00D5	00D6	00D7	00D8	00D9	00DA	00DB	00DC	00DD	00DE	00DF
E-	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
	00E0	00E1	00E2	00E3	00E4	00E5	00E6	00E7	00E8	00E9	00EA	00EB	00EC	00ED	00EE	00EF
F-	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ
	00F0	00F1	00F2	00F3	00F4	00F5	00F6	00F7	00F8	00F9	00FA	00FB	00FC	00FD	00FE	00FF

Dataset-XML FDA Pilot – some challenges - Encoding

```
ERROR: [CSTLOGMESSAGE] Invalid byte 1 of 1-byte UTF-8 sequence.  
com.sun.org.apache.xerces.internal.impl.io.MalformedByteSequenceException: Invalid byte 1 of  
1-byte UTF-8 sequence.  
    at com.sun.org.apache.xerces.internal.impl.io.UTF8Reader.invalidByte(UTF8Reader.java:687)  
    at com.sun.org.apache.xerces.internal.impl.io.UTF8Reader.read(UTF8Reader.java:557)  
  
ERROR: Some code points did not transcode.  
    occurred at or near line 14, column 62  
ERROR: XML parsing error. Please verify that the XML content is well-formed.  
ERROR: File WORK.ClinicalItemData.DATA has not been saved because copy could not be completed.
```

[Back](#)