SAS® TOOLS FOR WORKING WITH DATASET-XML FILES

Lex Jansen
Principal Software Developer @ SAS
CDISC XML Technologies Team
Agenda

Dataset-XML

• Introduction
• What is Dataset-XML
• Dataset-XML and ODM
• Dataset-XML and Define-XML
• Dataset-XML – more detail
• SAS Tools for Dataset-XML
• FDA Pilot
“Regulatory New Drug Review: Solutions for Study Data Exchange Standards”

Nov 5, 2012 | FDA Study Data Exchange Standards Meeting

- Solicit input from industry, technology vendors and other members of the public
- What are the **advantages and disadvantages** of current and emerging **open, consensus-based** standards for the exchange of regulated study data
- Agenda based on federal register notice (FRN) with pre-meeting questions
Background

• “The current study data exchange format supported by FDA is the ASCII-based SAS Transport (XPORT) version 5 file format. Although XPORT has been an exchange format for many years, it is not an extensible modern technology. Moreover, it is not supported and maintained by an open, consensus-based standards development organization.”

• “FDA would like to discuss the current and emerging open study data exchange standards that will support interoperability.”
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
Five options were presented at the meeting

1. SAS Transport v5 extensions (SAS Version 8 Transport format, available in SAS 9.3), addresses the character size issues
2. **CDISC Operational Data Model (ODM)**
3. HL7 Version 3 – including Clinical Document Architecture (CDA)
4. Semantic Web Technologies:
   - Resource Description Framework (RDF)
   - Web Ontology Language (OWL)
5. Analytic Information Markup Language (AnIML)
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, ...)

| STIDYD | DOMAIN | USBJD | ASEQ | AESPD | AESTERM | AEMONDF | AEDECODE | AEBOOSYS | AESEV | AESER | AEACN | AEREL | AEIDOC | AEENDT | AEENDT | AEENDY | AEENDT | AEENDY | AEENDT | AEENDY | AEENDT | AEENDY | AEENDT | AEENDY | AEENDT | AEENDY | AEENDT | AEENDY | AEENDT | AEENDY |
|--------|--------|-------|------|-------|---------|---------|----------|----------|-------|-------|-------|-------|--------|--------|---------|--------|--------|--------|--------|--------|--------|---------|--------|---------|--------|--------|---------|--------|---------|---------|--------|---------|--------|---------|
| CDISC01 | AE     | CDISC01.10 | 1  | 1     | AGITATED | AGITATION | Agitation | Psychiatric d | MILD  | N     | DOSE NOT | POSSIBLY | 2003-06 | 3 | AFTER  |
| CDISC01 | AE     | CDISC01.10 | 2  | 2     | ANXIETY  | Anxiety    | Psychiatric d | MILD  | N     | DOSE NOT | POSSIBLY | 2003-05-33 | 15 | AFTER  |
| CDISC01 | AE     | CDISC01.10 | 3  | 3     | DECREASE | Diarrhea   | Diarrhea   | Gastroenteri | MILD  | N     | DOSE NOT | NOT RELAT | 2004-01-06 | 84 | AFTER  |
| CDISC01 | AE     | CDISC01.10 | 4  | 4     | VOMIT    | VOMITING   | Vomiting   | Gastroenteri | SEVERE | Y     | DRUG INT. | POSSIBLY | 2004-02-03 | 112 | 112     |
| CDISC01 | AE     | CDISC01.10 | 5  | 5     | VOMIT    | VOMITING   | Vomiting   | Gastroenteri | SEVERE | Y     | DRUG INT. | POSSIBLY | 2004-02-09 | 113 | 118     |
| CDISC01 | AE     | CDISC01.20 | 1  | 1     | ANXIETY  | Anxiety    | Psychiatric d | SEVERE | N     | DOSE NOT | POSSIBLY | 2003-10-16 | 17 | 21      |
| CDISC01 | AE     | CDISC01.20 | 2  | 2     | LEFT KNEE | Arthritis   | Musculoskel | SEVERE | N     | DRUG WTH. | NOT RELAT | 2004-02-02 | 126 | AFTER  |
| CDISC01 | AE     | CDISC01.20 | 3  | 3     | CONSTIPATION | Constipation | Gastroenteri | SEVERE | N     | DOSE NOT | POSSIBLY | 2003-12-25 | 87 | AFTER  |
| CDISC01 | AE     | CDISC01.20 | 4  | 4     | TARDINESS | Fatigue    | General diso | SEVERE | N     | DOSE NOT | POSSIBLY | 2003-10-16 | 17 | 21      |
| CDISC01 | AE     | CDISC01.20 | 5  | 5     | Nausea   | Nausea     | Gastroenteri | MILD  | N     | DOSE NOT | NOT RELAT | 2004-02-26 | 140 | 140     |
| CDISC01 | AE     | CDISC01.20 | 6  | 6     | MUSCLE S. | Muscle spas. | Musculoskel | MILD  | N     | DOSE NOT | NOT RELAT | 2004-01-05 | 88 | AFTER  |
| CDISC01 | AE     | CDISC01.20 | 7  | 7     | PALPITAT. | Palpitations | Cardiovascul | 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
What is Dataset-XML

www.cdisc.org/dataset-xml

CDISC Dataset-XML 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

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Summary of Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-11-19</td>
<td>DRAFT 1.0</td>
<td>Version 1.0 for public comments, final version 1.0 incorporating all changes identified during public comment period, including removal of StudyDataset-XML (SDTM) namespace.</td>
</tr>
<tr>
<td>2014-04-22</td>
<td>FINAL 1.0</td>
<td>Version 1.0 for public comments, final version 1.0 incorporating all changes identified during public comment period, including removal of StudyDataset-XML (SDTM) namespace.</td>
</tr>
</tbody>
</table>

© 2014 Clinical Data Interchange Standards Consortium. All rights reserved.

Version 1.0 Final

Page 1 of 32
Dataset-XML  Tools

• Various tools under development to support
  • Validation
  • Data browsing (similar to SAS Viewer)
  • Conversion of SAS XPT files to Dataset-XML
  • Conversion of SAS data sets to Dataset-XML
  • Conversion of Dataset-XML to SAS data sets
  • Conversion of Dataset-XML to R
Dataset-XML Tools

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

CDISC Dataset-XML Resources

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Provided By</th>
<th>Links</th>
</tr>
</thead>
</table>

© Copyright 2014 SAS Institute Inc. All rights reserved.
## Dataset-XML Tool Summary

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Provided By</th>
<th>Links</th>
</tr>
</thead>
</table>
| 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 Cassels           | • 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                                                                                                            |
• Data and Metadata in Submissions Today

Data

SAS V5 XPT

Metadata

Define-XML
What is Dataset-XML

Data and Metadata

- Data and Metadata in Submissions Tomorrow

Data

Metadata

Dataset-XML

Define-XML

ODM-based Standards
What is Dataset-XML

Data and Metadata

Relationship of Dataset-XML to other CDISC Standards

- Define-XML
  - Represents
  - Defined by:
    - SEND model
      - SEND-IG
    - SDTM model
      - SDTM-IG
    - ADaM model
      - ADaM-IG

- Metadata
- ODM
  - Extended by

- Dataset-XML
  - Represents
  - follows
What is Dataset-XML

Data Transport

Convert SAS data sets to Dataset-XML

Send Dataset-XML

Receive Dataset-XML

Convert to SAS data sets or load into a data warehouse

Data Transport
Dataset-XML and ODM
Dataset-XML and ODM

- Vendor neutral XML Schema for exchange and archive of Clinical Trials metadata and data: snapshots, updates, archives
- In global production use since 2000 – currently at v1.3.2
- Supports Part 11 compliance and FDA Guidance on Computerized Systems
- Includes vendor extension capability
- Human and machine readable
Dataset-XML and ODM

- Hierarchical metadata structure: Study, protocol, events, forms, item groups, items
- Represents an entire clinical study:
  - Study metadata
  - Administrative metadata
  - Reference data
  - Subject data
  - Audit information
- Basis for Define-XML metadata description document used in submissions
- CDASH-ODM form metadata available
- SDM-XML represents BRIDG protocol/study design model (structure, workflow, timing)
- CT-XML delivers NCI-EVS controlled terminology
Dataset-XML and ODM - Extensions

- CRT-DDS v1
- Define-XML v2
- CT-XML
- Dataset-XML
- Analysis Results Metadata
- Study Design Model

ODM
Dataset-XML and ODM – Dataset-XML Extension
Dataset-XML and ODM – Dataset-XML Extension

```
<ODM>
  <ClinicalData>
    <SubjectData>
      <StudyEvenData>
        <FormData>
          <ItemGroupData>
            <ItemData>
            <ItemData>
          </ItemGroupData>
        </FormData>
      </StudyEvenData>
    </SubjectData>
  </ClinicalData>
</ODM>

/ODM/@data:DatasetXMLVersion
(The version of the Dataset-XML standard)

/ODM/ClinicalData/ItemGroupData/@data:ItemGroupDataSeq
(A unique sequence number for each ItemGroupData record in the dataset)
```
Dataset-XML and ODM
Dataset-XML and ODM – Unique Object Identifiers

• In ODM, there are many instances where one object needs to reference another -- both within the same file and across files within a series of ODM documents.
• To accomplish this, the target element is given a unique identifier (its OID).
• All elements that need to reference that target element just use its OID.
• The values used for OIDs can follow any convention, or even can be randomly generated.
• The only allowed use of OIDs is to define an unambiguous link between a definition of an object and references to it.
Dataset-XML and ODM – Unique Object Identifiers

```xml
<ClinicalData
    StudyOID="cdisc01"
    MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
    <!-- Dataset (AE) -->
    <ItemGroupData ItemGroupID="IG.AE" data:ItemGroupDataSeq="1">
        <ItemData ItemID="IT.STUDYID" Value="CDISC01"/>
        <ItemData ItemID="IT.AE.DOMAIN" Value="AE"/>
        <ItemData ItemID="IT.USUBJID" Value="CDISC01.100008"/>
        <ItemData ItemID="IT.AE.ASEQ" Value="1"/>
        <ItemData ItemID="IT.AE.AESEQID" Value="1"/>
        <ItemData ItemID="IT.AE.AETERM" Value="AGITATED"/>
    </ItemGroupData>
    <ItemGroupRef 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 ItemID="IT.STUDYID" OrderNumber="1" Mandatory="Yes" Key=""/>
        <ItemRef ItemID="IT.AE.DOMAIN" OrderNumber="2" Mandatory="Yes"/>
        <ItemRef ItemID="IT.USUBJID" OrderNumber="3" Mandatory="Yes" Key=""/>
        <ItemRef ItemID="IT.AE.ASEQ" OrderNumber="4" Mandatory="Yes"/>
        <ItemRef ItemID="IT.AE.AESEQID" OrderNumber="5" Mandatory="No"/>
        <ItemRef ItemID="IT.AE.AETERM" OrderNumber="6" Mandatory="Yes"/>
    </ItemGroupRef>

    <ItemDef OID="IT.AE.AETERM" Name="AETERM" DataTypes="text" Length="25" SASFieldName="AETERM">
        <Description>
            <TranslatedText xml:lang="en">Reported Term for the Adverse Event</TranslatedText>
        </Description>
    </ItemDef>
</ClinicalData>
```
Dataset-XML and Define-XML
Dataset-XML and Define-XML (data and metadata)

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

The SAS System

The CONTENTS Procedure

<table>
<thead>
<tr>
<th>Data Set Name</th>
<th>SDTM.AE</th>
<th>Observations</th>
<th>106</th>
</tr>
</thead>
<tbody>
<tr>
<td>Member Type</td>
<td>DATA</td>
<td>Variables</td>
<td>51</td>
</tr>
<tr>
<td>Engine</td>
<td>V9</td>
<td>Indexes</td>
<td>0</td>
</tr>
<tr>
<td>Created</td>
<td>09/26/2014 11:01:18</td>
<td>Observation Length</td>
<td>2200</td>
</tr>
<tr>
<td>Last Modified</td>
<td>09/26/2014 11:01:18</td>
<td>Deleted Observations</td>
<td>0</td>
</tr>
<tr>
<td>Protection</td>
<td>Compressed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Set Type</td>
<td>Sorted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td>Adverse Events</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Representation</td>
<td>WINDOWS_64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encoding</td>
<td>wlatin1 Western (Windows)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dataset-XML and Define-XML (data and metadata)

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

<table>
<thead>
<tr>
<th>#</th>
<th>Variable</th>
<th>Type</th>
<th>Len</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>STUDYID</td>
<td>Char</td>
<td>40</td>
<td>Study Identifier</td>
</tr>
<tr>
<td>2</td>
<td>DOMAIN</td>
<td>Char</td>
<td>8</td>
<td>Domain Abbreviation</td>
</tr>
<tr>
<td>3</td>
<td>USUBJID</td>
<td>Char</td>
<td>40</td>
<td>Unique Subject Identifier</td>
</tr>
<tr>
<td>4</td>
<td>AESEQ</td>
<td>Num</td>
<td>8</td>
<td>Sequence Number</td>
</tr>
<tr>
<td>5</td>
<td>AEGRPID</td>
<td>Char</td>
<td>40</td>
<td>Group ID</td>
</tr>
<tr>
<td>6</td>
<td>AEREFID</td>
<td>Char</td>
<td>40</td>
<td>Reference ID</td>
</tr>
<tr>
<td>7</td>
<td>AESPID</td>
<td>Char</td>
<td>40</td>
<td>Sponsor-Defined Identifier</td>
</tr>
<tr>
<td>8</td>
<td>AETERM</td>
<td>Char</td>
<td>200</td>
<td>Reported Term for the Adverse Event</td>
</tr>
</tbody>
</table>
Dataset-XML and Define-XML

Data set name? Variable names?

```xml
<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"/>
    </ItemGroupData>
</ClinicalData>
```
Dataset-XML and Define-XML

Data set name? Variable names?

```xml
<ClinicalData
  StudyOID="cdisc01"
  MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
  <ItemGroupData ItemGroupOID="IG1" data:ItemGroupDataSeq="1">
    <ItemData ItemOID="IT1" Value="CDISC01"/>
    <ItemData ItemOID="IT2" Value="AE"/>
    <ItemData ItemOID="IT3" Value="CDISC01.100008"/>
    <ItemData ItemOID="IT4" Value="1"/>
    <ItemData ItemOID="IT5" Value="1"/>
    <ItemData ItemOID="IT6" Value="AGITATED"/>
    <ItemData ItemOID="IT7" Value="AGITATION"/>
  </ItemGroupData>
</ClinicalData>
```
Dataset-XML and Define-XML

Data set name? Variable names?

```xml
<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"/>
  </ItemGroupData>
</ClinicalData>
```
<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" Meta
<ItemRef ItemOID="IT.AE.AESPID" OrderNumber="5" Mandatory="No" />
<ItemRef ItemOID="IT.AE.AETERM" OrderNumber="6" Mandatory="Yes" />
<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>
<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 ItemID="IT.STUDYID" OrderNumber="1" Mandatory="Yes" KeySequence="1"/>
      <ItemRef ItemID="IT.AE.DOMAIN" OrderNumber="2" Mandatory="Yes"/>
      <ItemRef ItemID="IT.USUBJID" OrderNumber="3" Mandatory="Yes" KeySequence="2" MethodOID="MT.USUBJID"/>
      ...
      <ItemRef ItemID="IT.AE.AETERM" OrderNumber="6" Mandatory="No"/>
      <ItemRef ItemID="IT.AE.AEMODIFY" OrderNumber="7" Mandatory="No"/>
      <ItemRef ItemID="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>

</ODM ...>

<ODM ...>

<ClinicalData StudyOID="cdisc01" MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
  <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupData_seq="1"/>
  <ItemData ItemID="IT.STUDYID" Value="CDISC01"/>
  <ItemData ItemID="IT.AE.DOMAIN" Value="AE"/>
  <ItemData ItemID="IT.USUBJID" Value="CDISC01.100006"/>
  ...
  <ItemData ItemID="IT.AE.AETERM" Value="AGITATION"/>
  <ItemData ItemID="IT.AE.AEDECOD" Value="Agitation"/>
  ...
</ItemGroupData>

</ODM ...>
**What is Dataset-XML Data Transport**

<table>
<thead>
<tr>
<th>ae.xpt</th>
<th>SAS Xport Transport File</th>
</tr>
</thead>
<tbody>
<tr>
<td>blankcrf.pdf</td>
<td>Adobe Acrobat Document</td>
</tr>
<tr>
<td>cm.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
<tr>
<td>complexalgorithms.pdf</td>
<td>Adobe Acrobat Document</td>
</tr>
<tr>
<td>da.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
<tr>
<td>define.xml</td>
<td>XML Document</td>
</tr>
<tr>
<td>define2-0-0.xsl</td>
<td>XSLT Stylesheet</td>
</tr>
<tr>
<td>dm.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
<tr>
<td>ds.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
<tr>
<td>eg.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
<tr>
<td>ex.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
<tr>
<td>ie.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
<tr>
<td>lb.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
<tr>
<td>mh.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
<tr>
<td>pe.xpt</td>
<td>SAS Xport Transport File</td>
</tr>
</tbody>
</table>

XML Document

Adobe Acrobat Document

XML Document

Adobe Acrobat Document

XML Document

XML Document

XML Document

XML Document

XML Document

XML Document

XML Document

XML Document

XML Document

XML Document

XML Document

XML Document

XML Document

XML Document
```xml
<?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"
FileID="www.cdisc.org.Studydisc01-Define-XML_2.0.0(IG.AE)"
PriorFileID="www.cdisc.org.Studydisc01-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>
  </ClinicalData>
</ODM>
```

<table>
<thead>
<tr>
<th>STUDYID</th>
<th>DOMAIN</th>
<th>USUBJID</th>
<th>AESEQ</th>
<th>AESPID</th>
<th>AETERM</th>
<th>AEMODIFY</th>
<th>AEDECOD</th>
<th>AEBODSYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01.100008</td>
<td>1</td>
<td>1</td>
<td>AGITATED</td>
<td>AGITATION</td>
<td>Agitation</td>
<td>Psychiatric disorders</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AESEV</th>
<th>AESER</th>
<th>AEACN</th>
<th>AEREL</th>
<th>AESTDTC</th>
<th>AEENDTC</th>
<th>AESTDY</th>
<th>AEENDY</th>
<th>AEENRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>MILD</td>
<td>N</td>
<td>DOSE NOT CHANGED</td>
<td>POSSIBLY RELATED</td>
<td>2003-05</td>
<td></td>
<td>3</td>
<td></td>
<td>AFTER</td>
</tr>
</tbody>
</table>
• Fields that are not populated do not have any `<ItemData>` elements
• The following examples are incorrect in Dataset-XML

```
<ItemData ItemOID="IT.AE.AEENTC" Value=""/>
<ItemData ItemOID="IT.AE.AEENTC" />
<ItemData ItemOID="IT.AE.AEENTC" IsNull="Yes"/>
```
<?xml version="1.0" encoding="UTF-8"?>
<OM>
  <xmlns:cdisc="http://www.cdisc.org/ns/odm/v1.3"
   FileType="Snapshot" ODMVersion="1.3.2" data:DatasetXMLVersion="1.0.0"
   FileOID="www.cdisc.org.Studydisc01-Define/XML_2.0.0(IG.TS)"
   PriorFileOID="www.cdisc.org.Studydisc01-Define/XML_2.0.0"
   Originator="CDISC Dataset-XML Team" CreationDateTime="2014-04-01T09:31:03">
    <ReferenceData>
      <StudyOID>cdisc01</StudyOID>
      <MetaDataVersionOID>MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2</MetaDataVersionOID>
      <ItemGroupData ItemGroupOID="IG.TS" data:ItemGroupDataSeq="1">
        <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
        <ItemData ItemOID="IT.TS.DOMAIN" Value="TS"/>
        <ItemData ItemOID="IT.TS.TSEQ" Value="1"/>
        <ItemData ItemOID="IT.TS.TSPARMCD" Value="ADDON"/>
        <ItemData ItemOID="IT.TS.TSPARM" Value="Added on to Existing Treatments"/>
        <ItemData ItemOID="IT.TS.TSVAL" Value="N"/>
      </ItemGroupData>
      <ItemGroupData ItemGroupOID="IG.TS" data:ItemGroupDataSeq="2">
        <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
        <ItemData ItemOID="IT.TS.DOMAIN" Value="TS"/>
        <ItemData ItemOID="IT.TS.TSEQ" Value="1"/>
        <ItemData ItemOID="IT.TS.TSPARMCD" Value="AGEMAX"/>
        <ItemData ItemOID="IT.TS.TSPARM" Value="Planned Maximum Age of Subjects"/>
        <ItemData ItemOID="IT.TS.TSVAL" Value="85"/>
      </ItemGroupData>
    </ReferenceData>
  </ODM>
<xml version="1.0" encoding="UTF-8"?>
<ODM
xmlns="http://www.cdisc.org/ns/odm/v1.3"
FileType="Snapshot" ODMVersion="1.3.2" data:DatasetXMLVersion="1.0.0"
FileOID="www.cdisc.org.StudyCDISC01-Define-XML_2.0.0(IG.SUPPDM)"
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.SUPPDM" data:ItemGroupDataSeq="1">
      <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
      <ItemData ItemOID="IT.RDOMAIN" Value="DM"/>
      <ItemData ItemOID="IT.USUBJID" Value="CDISC01.100008"/>
      <ItemData ItemOID="IT.SUPPDM.QNAM" Value="RACEOTH"/>
      <ItemData ItemOID="IT.SUPPDM.QLABEL" Value="Race, Other"/>
      <ItemData ItemOID="IT.SUPPDM.QVAL" Value="EASTERN INDIAN"/>
      <ItemData ItemOID="IT.SUPPDM.QORIG" Value="CRF"/>
    </ItemGroupData>
    <ItemGroupData ItemGroupOID="IG.SUPPDM" data:ItemGroupDataSeq="2">
      <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
      <ItemData ItemOID="IT.RDOMAIN" Value="DM"/>
      <ItemData ItemOID="IT.USUBJID" Value="CDISC01.100008"/>
      <ItemData ItemOID="IT.SUPPDM.QNAM" Value="RAND"/>
      <ItemData ItemOID="IT.SUPPDM.QLABEL" Value="Randomized Population Flag"/>
      <ItemData ItemOID="IT.SUPPDM.QVAL" Value="Y"/>
      <ItemData ItemOID="IT.SUPPDM.QORIG" Value="CRF"/>
    </ItemGroupData>
  </ClinicalData>
</ODM>
SAS Tools for Dataset-XML

Available Now

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

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, cstutilcomparemetadatasas define, cstutilsqcolumndefinition, cstutilsqgeneratetable, cstutilfindfixextdasciichars, cstutilregistercsubstrtype, and cstutilxmlvalidate.
SAS Tools for Dataset-XML

Available Now

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

---

**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](https://support.sas.com/rnd/base/cdisc/cst/index.html) 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 NOI controlled terminology.
- New framework macros have been added, including cstutilmanagecolumns, cstutilcomparemetadatadataset, cstutilsqlobjectdefinition, cstutilsqlobjectentable, cstutilfindfixextdasciichars, cstutilregistercstsubtype, and cstutilxmlvalidate.
Sample 53447: SAS® Macros to support Dataset-XML v1.0.0

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 Data

%datasetxml_write()

define.xml

%xml_validate()

Dataset-XML

%datasetxml_read()

cstutilcompare 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.
* 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;
* 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;
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

**Dataset-XML**  
**FDA Pilot**

- **Objectives:**
  - test the transport functionality of DS-XML, which included ensuring that data integrity was maintained and that DS-XML format would support longer variable names, labels, and text fields

---

**Pilot Report:**
http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm380756.htm
Dataset-XML  
FDA Pilot – Testing

Pilot Report:
http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm380756.htm
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”).
An XML document starts with an optional XML declaration:

```xml
<?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)

<table>
<thead>
<tr>
<th>ISO 8859-1</th>
<th>Windows Latin-1</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Character Table" /></td>
<td><img src="image" alt="Character Table" /></td>
</tr>
</tbody>
</table>

**ISO 8859-1**

- Characters in the range 0x00 to 0x7F are the same in both Windows Latin-1 and ISO 8859-1.
- Characters in the range 0x80 to 0xFF are different between the two encodings.

**Windows Latin-1**

- Characters in the range 0x00 to 0x7F are the same in both Windows Latin-1 and ISO 8859-1.
- Characters in the range 0x80 to 0xFF are different between the two encodings.

---

*Copyright © 2014, SAS Institute Inc. All rights reserved.*
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.
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.
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 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**

<table>
<thead>
<tr>
<th></th>
<th>SAS (compress)</th>
<th>XPT</th>
<th>XML</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB</td>
<td>301.51 MB</td>
<td>636.07 MB</td>
<td>1.75 GB</td>
<td>52.91 MB</td>
</tr>
<tr>
<td>QS</td>
<td>432.08 MB</td>
<td>776.73 MB</td>
<td>2.04 GB</td>
<td>53.68 MB</td>
</tr>
<tr>
<td>SUPPLB</td>
<td>338.98 MB</td>
<td>717.81 MB</td>
<td>1.79 GB</td>
<td>29.25 MB</td>
</tr>
<tr>
<td>SUPPQS</td>
<td>39.23 MB</td>
<td>37.28 MB</td>
<td>214.05 MB</td>
<td>3.73 MB</td>
</tr>
</tbody>
</table>
FDA envisions conducting several pilots to evaluate new transport formats before a decision is made to support a new format.
THANK YOU!

QUESTIONS?