CDISC Dataset-XML
A new Dataset Structure for Clinical Trial Data Transport
for Future Drug Submissions

Joerg Dillert
Consulting Solution Director
Oracle Health Sciences

PhUSE Annual Conference
October, 13th - 15th 2014
Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.
CDISC Interchange, Europe, Call for ABSTRACTS

The Programme Committee is now accepting abstracts for presentations and posters at the CDISC EU Interchange in Basel, Switzerland on the 5th and 6th May 2015.

The hot topics we see for this year are:

Measureable Benefits of CDISC Implementations
Implementing CDISC standards (e.g. SDTM, terminology, CDASH) into your R&D process and measure and metrics of the benefits.

Practical Implementation of Therapeutic Area Standards
Therapeutic area specific topics and their impact on existing CDISC standards; practical implementation experiences

Protocol Metadata for the CDISC end-to-end implementation
Structured protocol metadata as a starting point to automate setup of clinical databases and downstream CDISC standards deliverables such as CRF’s, datasets and reports. What is needed?
Credits & Acknowledgement
CDISC Dataset-XML team

- Sam Hume, CDISC
- Sally Cassells, Next Step Clinical Systems LLC
- Kevin Burges, Formedix
- Jozef Aerts, XML4Pharma
- Lex Jansen, SAS Institute
- Marcelina Hungria, DIconcore Group, LLC
- Mike Molter, D-Wise
- Peter Schaefer, Certara
- Paul Graham, Formedix
- Vanessa Nguyen, Baxter
- Veena Nataraj, Shire
- Vojtech Huser, NIH
- Priscilla Gathoni, Novartis
- Yoshiteru CHIBA, UMINCenter, Japan
CDISC Dataset-XML

1. Quick history of SAS®-XPT in Electronic Submissions
2. CDISC Dataset-XML
3. Data Structure
4. Demo
5. Status and Future
History of SAS®-XPT

... and the combined challenges
SAS®-XPT in Clinical Research and Electronic Submissions

• FDA, eCTD and Electronic Submissions

• Free available and documented formats: PDF and SAS®-XPT (V5)
Hmmm?

• 10 years later ...
  – IT technology improvement
  – PDF version issues
    • FDA PDF v1.4-1.7
    • Security issues
  – SAS®-XPT for non SAS systems

FDA PORTABLE DOCUMENT FORMAT (PDF) SPECIFICATIONS

PURPOSE

These specifications are for submitting documents in Portable Document Format (PDF). The purpose of this document is to provide specifications for submitting PDF files that align with the ICH M2 recommendations and that are in a format that the receiving Center currently supports. For purposes of this document, “supports” means the receiving Center has established processes and technology infrastructure to support the receipt, processing, review, and archive of files in the specified standard format. PDF is an open, published format created by Adobe Systems Incorporated (http://www.adobe.com). Software from a variety of sources can be used to create files in the PDF format.

VERSION

PDF versions 1.4 through 1.7 are acceptable. Submitted PDF files should be readable by Adobe Acrobat 8.0, should not require additional software or plug-ins to be read and navigated, and should be text searchable. If plug-ins are used during the creation of a PDF document, prior to submitting the document, ensure that a plug-in is not needed for review or archive.

SECURITY

Do not activate security settings or password protection. The integrity of the submitted files is maintained through Agency security and archival processes. A copy of the files, generated from the submitted files, will be provided to the reviewer. The reviewer should be able to print, select text and graphics, and make changes to text, notes and form fields using the provided copy.

FDA Forms downloaded from the FDA Forms website contain security settings that prevent changing the documents. These forms should be submitted as provided, with no additional security added and without removing the provided security settings.
Sure, documented, – but kind of binary

https://support.sas.com/techsup/technote/ts140.pdf
What’s wrong with XPT?

• Variable name length limitation
• Variable label length limitation
• Variable data length limitation
• No support for international characters sets
• A kind of binary format ( structs )

-→ Hard to process on non-SAS® platforms

https://support.sas.com/techsup/technote/ts140.pdf
Requirements for a change

• A new, modern dataset transport format of regulated data

• Easy to process (with free tools) and faster software development

• Support tabular datasets including SDTM, ADaM, SEND

• Support of international character sets
CDISC Dataset-XML

... as the answer
Where does the name come from?

• Original name was

  Study Data Set XML (SDS-XML)

• To avoid confusion changed to

  Dataset-XML
Dataset-XML

• Based on CDISC ODM and Define-XML
  – Define.XML V 2.0 or later is recommended for Dataset-XML

• Related to SDTM, ADaM, SEND and any kind of tabular data in the future
What will change for submission / eCTD?
For the end user?

- Old

**SDTM-IG 3.1.2**
- Annotated Case Report Form
- Reviewers Guide
- Complex Algorithms
- Tabulation Datasets
- Value Level Metadata
- Controlled Terminology
- Computational Algorithms
- Comments

---

### Tabulation Datasets for Study CDISC01 (SDTM-IG 3.1.2)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Class</th>
<th>Structure</th>
<th>Purpose</th>
<th>Keys</th>
<th>Location</th>
<th>Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA</td>
<td>Trial Arms</td>
<td>TRIAL DESIGN</td>
<td>One record per planned Element per Arm</td>
<td>Tabulation</td>
<td>STUDYID, ARMCD, TAETORD</td>
<td>ta.xpt</td>
<td></td>
</tr>
<tr>
<td>TE</td>
<td>Trial Elements</td>
<td>TRIAL DESIGN</td>
<td>One record per planned Element</td>
<td>Tabulation</td>
<td>STUDYID, ETCD</td>
<td>te.xpt</td>
<td></td>
</tr>
<tr>
<td>TI</td>
<td>Trial Inclusion/Exclusion Criteria</td>
<td>TRIAL DESIGN</td>
<td>One record per I/E criterion</td>
<td>Tabulation</td>
<td>STUDYID, IETESTCD</td>
<td>ti.xpt</td>
<td></td>
</tr>
<tr>
<td>TS</td>
<td>Trial Summary</td>
<td>TRIAL DESIGN</td>
<td>One record per trial summary parameter value</td>
<td>Tabulation</td>
<td>STUDYID, TSPARMCD, TSSEQ</td>
<td>ts.xpt</td>
<td></td>
</tr>
<tr>
<td>TV</td>
<td>Trial Visits</td>
<td>TRIAL DESIGN</td>
<td>One record per planned Visit per Arm</td>
<td>Tabulation</td>
<td>STUDYID, VISITNUM, ARMCD</td>
<td>tv.xpt</td>
<td></td>
</tr>
<tr>
<td>DM</td>
<td>Demographics</td>
<td>SPECIAL PURPOSE</td>
<td>One record per subject</td>
<td>Tabulation</td>
<td>STUDYID, USUBID</td>
<td>dm.xpt</td>
<td>See Reviewer's Guide, Section 2.1 Demographics Reviewers Guide</td>
</tr>
</tbody>
</table>
What will change for submission / eCTD?

Nothing special

• New

Tabulation Datasets for Study CDISC01 (SDTM-IG 3.1.2)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Class</th>
<th>Structure</th>
<th>Purpose</th>
<th>Keys</th>
<th>Location</th>
<th>Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA</td>
<td>Trial Arms</td>
<td>TRIAL DESIGN</td>
<td>One record per planned Element per Arm</td>
<td>Tabulation</td>
<td>STUDYID, ARMCD, INETORODT</td>
<td>ts.xml</td>
<td></td>
</tr>
<tr>
<td>TE</td>
<td>Trial Elements</td>
<td>TRIAL DESIGN</td>
<td>One record per planned Element</td>
<td>Tabulation</td>
<td>STUDYID, ETCO</td>
<td>ts.xml</td>
<td></td>
</tr>
<tr>
<td>TI</td>
<td>Trial Inclusion/Exclusion Criteria</td>
<td>TRIAL DESIGN</td>
<td>One record per Inclusion/Exclusion Criteria</td>
<td>Tabulation</td>
<td>STUDYID, ITETESCO</td>
<td>ts.xml</td>
<td></td>
</tr>
<tr>
<td>TG</td>
<td>Trial Summary</td>
<td>TRIAL DESIGN</td>
<td>One record per trial Summary parameter value</td>
<td>Tabulation</td>
<td>STUDYID, TSPARMCO, TSSEQ</td>
<td>ts.xml</td>
<td></td>
</tr>
<tr>
<td>TV</td>
<td>Trial Visits</td>
<td>TRIAL DESIGN</td>
<td>One record per planned Visit per Arm</td>
<td>Tabulation</td>
<td>STUDYID, VISITNUM, ARMCD</td>
<td>ts.xml</td>
<td></td>
</tr>
<tr>
<td>DM</td>
<td>Demographics</td>
<td>SPECIAL PURPOSE</td>
<td>One record per subject</td>
<td>Tabulation</td>
<td>STUDYID, USSURID</td>
<td>dm.xml</td>
<td>See Reviewer’s Guide, Section 2.1 Demographics Reviewers Guide</td>
</tr>
<tr>
<td>SE</td>
<td>Subject Elements</td>
<td>SPECIAL</td>
<td>One record per actual</td>
<td>Tabulation</td>
<td>STUDYID, USSUSID, SE.ARXI</td>
<td>se.xml</td>
<td></td>
</tr>
</tbody>
</table>
Data Structure

Bring some light on it – no magic
ODM is my base for the structure (1) - Clinical Data

<ODM> - base ODM information as usual
  <ClinicalData> - study information, metadata version
    <ItemGroupData> - table/dataset level
      <ItemData> - data points, items

For all clinical domains, like AE, DM, CM etc.
First Clinical 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/v1.6"
    FileType="Snapshot"
    ODMVersion="1.3.2"
    data:DatasetXMLVersion="1.0"
    FileOID="www.cdisc.org.Studyodisc01-Define-XML_2.0.0(IG.CM)"
    PriorFileOID="www.cdisc.org.Studyodisc01-Define-XML_2.0.0"
    Originator="CDISC Dataset-XML Team"
    CreationDateTime="2014-03-20T21:45:33">
    <ClinicalData
        StudyOID="cdisc01"
        MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
        <!-- Dataset (CM) -->
        <ItemGroupData ItemGroupOID="IG.CM" data:ItemGroupDataSeq="1">
            <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
            <ItemData ItemOID="IT.CM.DOMAIN" Value="CM"/>
            <ItemData ItemOID="IT.USUBJID" Value="CDISC01.100008"/>
            <ItemData ItemOID="IT.CM.CMSEQ" Value="1"/>
            <ItemData ItemOID="IT.CM.CMRTT" Value="PROCARDIA XL"/>
            ...
        </ItemGroupData>
        ...
    </ClinicalData>
</ODM>
ODM is my base for the structure (2) – Reference Data

<ODM>

<ReferenceData>

<ItemGroupData>

<ItemData>

- base ODM information as usual
- study information, metadata version
- table/dataset level
- data points, items

For all reference / trial metadata domains, like TA, TV etc
Reference 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:="http://www.cdisc.org/ns/Dataset-XML/v1.0"
 FileVersion="Snapshot"
 ODMVersion="1.3.2"
 data:DatasetXMLVersion="1.0"
 FileID="www.cdisc.org.Studycdisc01-Define-XML_2.0.0(IG.TA)"
 PriorFileID="www.cdisc.org.Studycdisc01-Define-XML_2.0.0"
 Originator="CDISC Dataset-XML Team"
 CreationDateTime="2014-03-20T21:45:33" >
 <ReferenceData
    StudyID="cdisc01"
    MetaDataVersionID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
  <!-- Dataset (TA) -->
  <ItemGroupData ItemGroupID="IG.TA" data:ItemGroupDataSeq="1">
    <ItemData ItemID="IT.STUDYID" Value="CDISC01"/>
    <ItemData ItemID="IT.TA.DOMAIN" Value="TA"/>
    <ItemData ItemID="IT.TA.ARMCD" Value="PLACEBO"/>
    <ItemData ItemID="IT.TA.ARM" Value="Placebo"/>
    <ItemData ItemID="IT.TA.TARGETD" Value="1"/>
    <ItemData ItemID="IT.TA.TRCTD" Value="SCREEN"/>
    <ItemData ItemID="IT.TA.ELEMENT" Value="Screening"/>
    <ItemData ItemID="IT.TA.EPOCH" Value="SCREEN"/>
   </ItemGroupData>
  ...
 </ReferenceData>
</ODM>
Define.XML are my metadata description

Relation of OIDS

<table>
<thead>
<tr>
<th>Define.XML</th>
<th>Dataset.XML</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;Study OID = ...&gt;</code></td>
<td><code>&lt;ClinicalData StudyOID = ...&gt;</code></td>
</tr>
<tr>
<td><code>&lt;MetaDataVersion OID = ...&gt;</code></td>
<td><code>&lt;ClinicalData MetaDataVersionOID = ...&gt;</code></td>
</tr>
<tr>
<td><code>&lt;ItemGroupDef OID = ...&gt;</code></td>
<td><code>&lt;ItemGroupData ItemGroupOID = ...&gt;</code></td>
</tr>
<tr>
<td><code>&lt;ItemRef ItemOID = ...&gt;</code></td>
<td><code>&lt;ItemData ItemOID = ...&gt;</code></td>
</tr>
</tbody>
</table>

Remark: For `<ReferenceData>` same like `<ClinicalData>`
Define.XML and Dataset.XML OID mapping
One XML file per (SDTM) domain

• One XML file per (SDTM) domain

• NULL and MISSING values not included as an ItemData element

• Same for <ReferenceData>
XML schemas and extensions

For Technical Reference

```xml
<OM version="1.0" encoding="UTF-8">
  <OM
    xmlns=http://www.cdisc.org/ns/omd/cdm/v1.3
    xmlns:xlink=http://www.w3.org/1999/xlink
    xlink:namespace=http://www.cdisc.org/ns/omd/cdm/v1.3
    xlink:namespaceVersion="1.3.2"
    xlink:fileOID="http://www.cdisc.org/studyCDISC01-Define-XML_2.0.0(IG.TA)"
    xlink:primaryFileOID="http://www.cdisc.org/studyCDISC01-Define-XML_2.0.0"
    xlink:originator="CDISC Dataset-XML Team"
    xlink:CreateDate="2014-03-01T01:45:33">
    <ReferenceData
      xlink:fileOID="http://www.cdisc.org/studyCDISC01-Define-XML_2.0.0(IG.TA)"
      xlink:primaryFileOID="http://www.cdisc.org/studyCDISC01-Define-XML_2.0.0"
      xlink:originator="CDISC Dataset-XML Team"
      xlink:CreateDate="2014-03-01T01:45:33">
      <ItemGroupData xlink:ItemGroupOID="IG.TA"
        xlink:ItemGroupDataSeq="1">
        <ItemData xlink:ItemOID="ID.TA.STUDYID" Value="CDISC01"/>
        <ItemData xlink:ItemOID="ID.TA.TA.DOMAIN" Value="TA"/>
        <ItemData xlink:ItemOID="ID.TA.ARMCD" Value="PLACEBO"/>
        <ItemData xlink:ItemOID="ID.TA.TA.TAETOPD" Value="1"/>
      </ItemGroupData>
    </ReferenceData>
  </OM>
</OM>
```
Data Types

• All data types follow the definition in the define.xml specification.

• For converting floats, the define.xml should contain entries in the ‘SignificantDigits’ and ‘Length’ attributes for the item definition in <ItemDef> element.
Tools & Ressourcess available
http://wiki.cdisc.org/display/PUB/CDISC+Dataset-XML+Resources
Status and Future
Question: XML, size and process in IT systems?

• Answer: You should not open an XML file in ‘notepad’, instead use ‘vi(m)’
• Answer: For transport you may compress the file (ASCII compress perfect)
• Answer: Serialize / and or process serial
• Answer: In-Memory usage

• XML file size should not be a limit
Where we are (all) today?

• Scope: Functionality required to replace XPT files for FDA submission

• XPT limitations removed for a modern, vendor neutral format

• One file per dataset

• Supports SDTM concepts incl. SuppQuals, RELREC and tabular structures in general
Where we are (all) today? (cont.)

• Support international character sets

• FDA have announced a pilot project
  http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm380756.htm

• Several tools developed
Future

• Incremental transfers
• Audit records
• Different ways of grouping data
  – All in one file, one file per subject, one file per domain per subject
• Typed data
• Inline Supplemental Qualifiers (once in SDTM)
• Enables SDTM to remove 200 char limit on fields
Demo
Study Dataset XML-Viewer

<table>
<thead>
<tr>
<th>STUDBID</th>
<th>RELECID</th>
<th>AE</th>
<th>DS</th>
<th>LB</th>
<th>SUPPAF</th>
<th>STUDYID</th>
<th>RDOMAIN</th>
<th>USUBJID</th>
<th>CDISC01</th>
<th>CDISC01:200001</th>
</tr>
</thead>
<tbody>
<tr>
<td>01-701-1145</td>
<td>1145</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>STUDBID</td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>01-701-1015</td>
<td>1015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01-701-1015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01-701-1145</td>
<td>1146</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01-701-1015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01-701-1148</td>
<td>1148</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01-701-1015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01-701-1153</td>
<td>1153</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01-701-1015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013-03-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2013-02-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013-03-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2013-03-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014-07-09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2014-07-09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**XML Viewer Message**
There are 2 related records with RELID '1':
- Domain = AE - USUBJID = CDISC01:1200001 - AESPId = 1
- Domain = CD - USUBJID = CDISC01:200001 - DSSEQ = 3

<table>
<thead>
<tr>
<th>DM</th>
<th>RELREC</th>
<th>AE</th>
<th>DS</th>
<th>LB</th>
<th>SUPPAF</th>
<th>STUDYID</th>
<th>RDOMAIN</th>
<th>USUBJID</th>
<th>CDISC01</th>
<th>CDISC01:200001</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:1:100008</td>
<td>1</td>
<td>1</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:1:100008</td>
<td>2</td>
<td>2</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:1:100008</td>
<td>3</td>
<td>3</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:1:100014</td>
<td>1</td>
<td>1</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:1:100014</td>
<td>2</td>
<td>2</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:1:100014</td>
<td>3</td>
<td>3</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:1:100014</td>
<td>4</td>
<td>4</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:1:100014</td>
<td>5</td>
<td>5</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:1:100014</td>
<td>1</td>
<td>1</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC01</td>
<td>AE</td>
<td>CDISC01:200001</td>
<td>2</td>
<td>2</td>
<td></td>
<td>CDISC01</td>
<td>CDISC01:200001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**LBDTC Table**
- 2014-08-25T13:40: 0.3
- 2014-10-21T09:52: 0.4
- 2014-11-18T12:23: 0.4
- 2014-12-2014-10-21T06:52: LBDTO: LBDY = 113

**COMPLT24**
- Label: Completers of Week 24 Population Flag
- Mandatory: No

Copyright © 2014, Oracle and/or its affiliates. All rights reserved.
Study Dataset XML-Viewer WebService version

- Calls external WebServices for more information
- Idea is to call CDISC SHARE in the future
Study Dataset XML-Viewer WebService version

• Problem to see anything?

• VPN? PROXY?
Study Dataset XML-Viewer – 64bit JAVA recommended
XPT2DataSetXML Converter
EZ Convert – OpenCDISC Validator
No Magic!
Hardware and Software
Engineered to Work Together