PhUSE Annual Conference2014

Paper CD15

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

Jörg Dillert, Oracle Health Sciences, Potsdam, Germany

ABSTRACT

CDISC has released the first version of the StudyDataSet-XML specification. StudyDataSet-XML is a CDISC XML format for CDISC SDTM, SEND, ADaM and legacy datasets. It has been developed to provide an alternative to SAS® V 5 Transport XPT for dataset transmissions.
This format is based on ODM 1.3.1 and extends the define.xml 2.0 standard in functionality. Its importance is thus at least as high as that of define.xml. The first advantages of this format are that no more 8-, 40-, or 200 character limitations exists and supplemental qualifiers can be stored in the parent dataset.

The initial release focuses on the basics, removing the XPT limitations and integrating with Define-XML metadata. This provides the functionality required to replace XPT files for FDA submission. To see if the FDA can adopt it, they announced a pilot project to determine if it is usable for submissions. The presentation will highlight the major features of the new standard.

INTRODUCTION

In the United States, the approval process for regulated human and animal health products requires the submission of data from clinical trials and other studies as expressed in the Code of Federal Regulations (CFR). The FDA established the regulatory basis for wholly electronic submission of data in 1997 with the publication of regulations on the use of electronic records in place of paper records (21 CFR Part 11). In 1999, the FDA standardized the submission of clinical and non-clinical data using the SAS XPORT Transport Format and the submission of metadata using Portable Document Format (PDF). In 2005, the Study Data Specifications published by the FDA included the recommendation that data definitions (metadata) be provided as a Define-XML file. On November 12, 2012, the FDA held a meeting entitled “Regulatory New Drug Review: Solutions for Study Data Exchange Standards”, the purpose of which was to solicit input regarding the advantages and disadvantages of current and emerging open, consensus-based standards for the exchange of regulated study data. Dataset-XML was presented as an ODM-based alternative for consideration.

Dataset-XML defines an ODM-based standard format for transporting tabular dataset data in XML between any two entities. That is, in addition to supporting the transport of datasets as part of a submission to the FDA, it may also be used to facilitate other data interchange use cases. For example, the Dataset-XML data format can be used by a CRO to transmit SDTM or ADaM datasets to a sponsor organization. Dataset-XML supports SDTM, ADaM, and SEND CDISC datasets, but can also be used to exchange any other type of tabular dataset.

RELATIONSHIP TO OTHER CDISC STANDARDS

There are relationships to other CDISC standards:

- CDISC ODM
- Define-xml
- SDTM
- ADaM
- SEND

CDISC ODM – OPERATIONAL DATA MODEL

The Dataset-XML standard is based on the CDISC ODM model. The ODM model includes the structure definition for clinical data and metadata and includes all information that needs to be shared among different software systems during study setup, operation, analysis and submission. ODM can be extended using a standardized mechanism for defining XML schema extensions.

DEFINE-XML

Define-XML is implemented as CDISC ODM extension and describes the model (metadata) that defines CDISC SDTM, ADaM and SEND datasets. Define-XML V 2.0 or later is recommended for use with Dataset-XML.

SDTM, SEND, ADAM

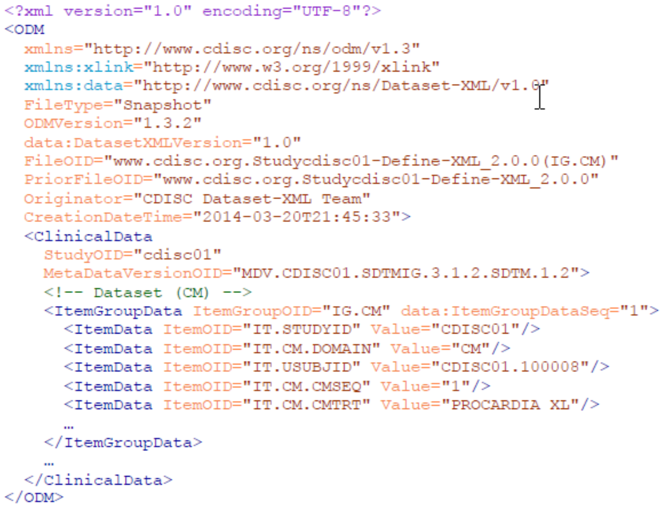
All the CDISC data standards - SDTM, SEND and ADaM can be used to be transmitted via Dataset-XML.

DATASET-XML DOCUMENT STRUCTURE

The Dataset-XML uses the <ClinicalData/> and <ReferenceData/> elements of the CDISC ODM definition. Pure clinical data domain like adverse event data (AE), concomitant medication (CM) must inserted on the <ClinicalData/> element. Reference data domains for the trial itself like Trial Arms (TA) must be included in the <ReferenceData/> element.

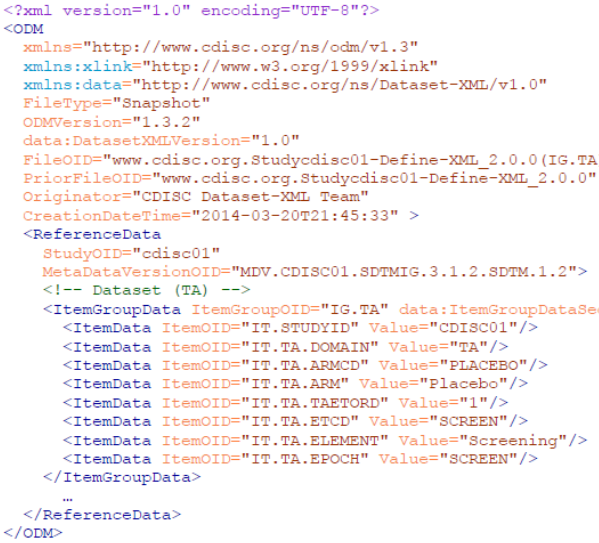
The “MetaDataVersionOID” attribute references the actual version of the data standard which is submitted (ex. SDTMMIG 3.1 and SDTM 1.2) and therefore the same in both elements, <ClinicalData/> as well as <ReferenceData/>. In the examples below the “MDV.CDISC01.SDTMMIG.3.1.2.SDTM.1.2” value characterizes the used SDTM version.

Dataset-XML document structure using the ClinicalData element:



```
<?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"
  FileOID="www.cdisc.org.Studycdisc01-Define-XML_2.0.0(IG.CM)"
  PriorFileOID="www.cdisc.org.Studycdisc01-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.CMTRT" Value="PROCARDIA XL"/>
      …
    </ItemGroupData>
    …
  </ClinicalData>
</ODM>
```

Dataset-XML document structure using the ReferenceData element:



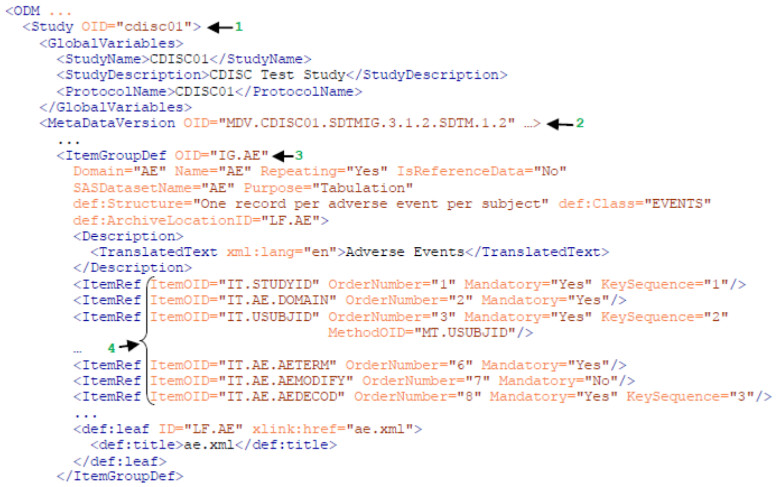
```
<?xml version="1.0" encoding="UTF-8"?>
<ODM
  xmlns="http://www.cdisc.org/ns/odm/v1.3"
  xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0"
  FileType="Snapshot"
  ODMVersion="1.3.2"
  data:DatasetXMLVersion="1.0"
  FileOID="www.cdisc.org.Studycdisc01-Define-XML_2.0.0(IG.TA)"
  PriorFileOID="www.cdisc.org.Studycdisc01-Define-XML_2.0.0"
  Originator="CDISC Dataset-XML Team"
  CreationDateTime="2014-03-20T21:45:33" >
  <ReferenceData
    StudyOID="cdisc01"
    MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
    <!-- Dataset (TA) -->
    <ItemGroupData ItemGroupOID="IG.TA" data:ItemGroupDataSeq="1">
      <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
      <ItemData ItemOID="IT.TA.DOMAIN" Value="TA"/>
      <ItemData ItemOID="IT.TA.ARMCD" Value="PLACEBO"/>
      <ItemData ItemOID="IT.TA.ARM" Value="Placebo"/>
      <ItemData ItemOID="IT.TA.TAETORD" Value="1"/>
      <ItemData ItemOID="IT.TA.ETCD" Value="SCREEN"/>
      <ItemData ItemOID="IT.TA.ELEMENT" Value="Screening"/>
      <ItemData ItemOID="IT.TA.EPOCH" Value="SCREEN"/>
    </ItemGroupData>
      …
  </ReferenceData>
</ODM>
```

RELATION OF OIDS BETWEEN DATASET-XML AND DEFINE-XML

Attributes whose names end with “OID” are used to uniquely identify specific metadata objects. For example, in the ItemData XML element the ItemOID attribute references a specific ItemDef in the define.xml file containing the variable metadata. Although the examples in this document use prefixes in the OIDs to indicate the object type, this is not required. The value of the OID attribute has no meaning by itself.

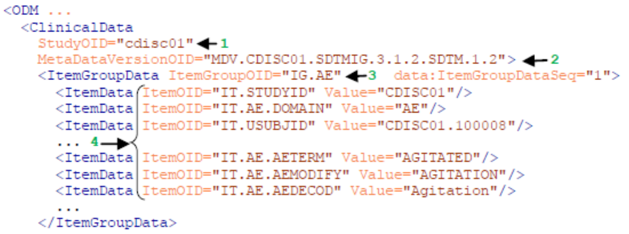
In the examples below the corresponding numbers are highlighted in green. The <Study> OID attribute (1) and the <Metadataversion> OID attribute (2) have the same value. The <ItemGroupDef> OID in define.xml names the <ItemGroupData> element in in dataset.xml. At the end, the ItemOIDs are the same in both files – in <ItemRef> (define.xml) and <ItemData> (dataset.xml).

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="Yes"/>
        <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>
```

ae.xml



```
<ODM ...
  <ClinicalData
    StudyOID="cdisc01" ← 1
    MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2"> ← 2
    <ItemGroupData ItemGroupOID="IG.AE" ← 3  data:ItemGroupDataSeq="1">
      <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
      <ItemData ItemOID="IT.AE.DOMAIN" Value="AE"/>
      <ItemData ItemOID="IT.USUBJID" Value="CDISC01.100008"/>
      ... 4 →
      <ItemData ItemOID="IT.AE.AETERM" Value="AGITATED"/>
      <ItemData ItemOID="IT.AE.AEMODIFY" Value="AGITATION"/>
      <ItemData ItemOID="IT.AE.AEDECOD" Value="Agitation"/>
      ...
    </ItemGroupData>
```

KEY DIFFERENCE BETWEEN THE DATASET-XML STRUCTURE AND ODM

As Dataset-XML represents just tabular data, the hierarchy has been simplified from

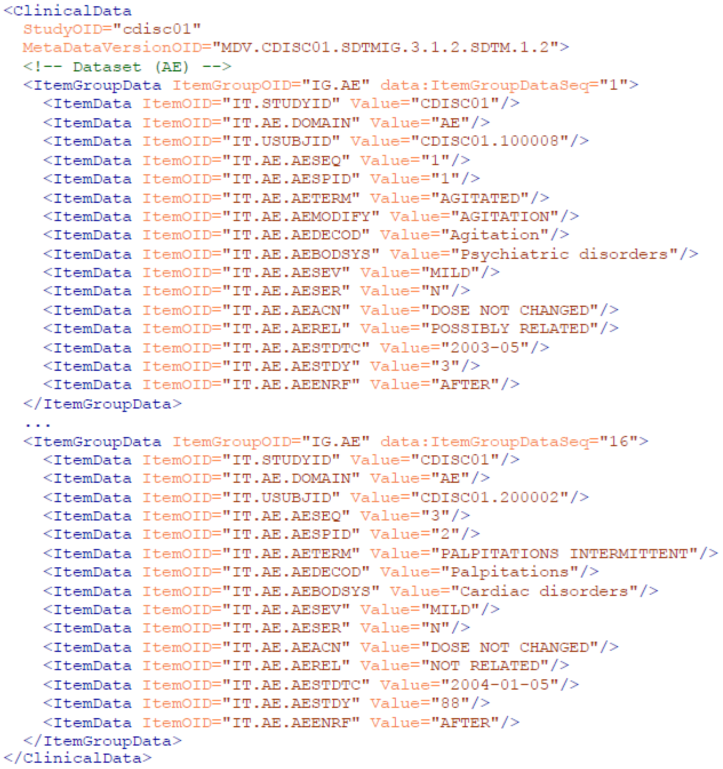
ODM/ClinicalData/SubjectData/StudyEventData/FormData/ItemGroupData/ItemData

to

ODM/ClinicalData/ItemGroupData/ItemData

EXAMPLE FOR SUBJECT DATA SET (CLINICAL DATA)

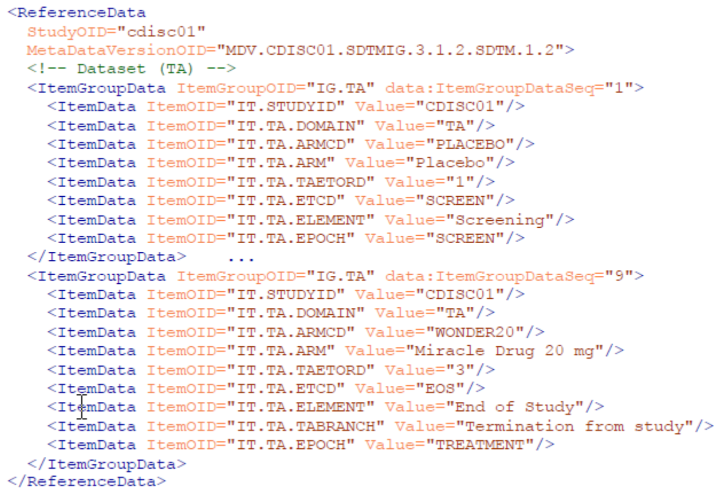
Every single domain must be included in a single xml file. For example, the AE domain should have a single ae.xml file. The example below shows two adverse event records. Missing and NULL values are not included as an ItemData element.



```
<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"/>
    <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>
  ...
  <ItemGroupData ItemGroupOID="IG.AE" data:ItemGroupDataSeq="16">
    <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
    <ItemData ItemOID="IT.AE.DOMAIN" Value="AE"/>
    <ItemData ItemOID="IT.USUBJID" Value="CDISC01.200002"/>
    <ItemData ItemOID="IT.AE.AESEQ" Value="3"/>
    <ItemData ItemOID="IT.AE.AESPID" Value="2"/>
    <ItemData ItemOID="IT.AE.AETERM" Value="PALPITATIONS INTERMITTENT"/>
    <ItemData ItemOID="IT.AE.AEDECOD" Value="Palpitations"/>
    <ItemData ItemOID="IT.AE.AEBODSYS" Value="Cardiac 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="NOT RELATED"/>
    <ItemData ItemOID="IT.AE.AESTDTC" Value="2004-01-05"/>
    <ItemData ItemOID="IT.AE.AESTDY" Value="88"/>
    <ItemData ItemOID="IT.AE.AEENRF" Value="AFTER"/>
  </ItemGroupData>
</ClinicalData>
```

EXAMPLE FOR REFERENCE DATA SET (REFERENCE DATA)

Every single domain must be included in a single xml file. For example, the TA domain should have a single ta.xml file. The example below shows two trial arm records. Missing and NULL values are not included as an ItemData element.



```
<ReferenceData
  StudyOID="cdisc01"
  MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2">
  <!-- Dataset (TA) -->
  <ItemGroupData ItemGroupOID="IG.TA" data:ItemGroupDataSeq="1">
    <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
    <ItemData ItemOID="IT.TA.DOMAIN" Value="TA"/>
    <ItemData ItemOID="IT.TA.ARMCD" Value="PLACEBO"/>
    <ItemData ItemOID="IT.TA.ARM" Value="Placebo"/>
    <ItemData ItemOID="IT.TA.TAETORD" Value="1"/>
    <ItemData ItemOID="IT.TA.ETCD" Value="SCREEN"/>
    <ItemData ItemOID="IT.TA.ELEMENT" Value="Screening"/>
    <ItemData ItemOID="IT.TA.EPOCH" Value="SCREEN"/>
  </ItemGroupData>    ...
  <ItemGroupData ItemGroupOID="IG.TA" data:ItemGroupDataSeq="9">
    <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
    <ItemData ItemOID="IT.TA.DOMAIN" Value="TA"/>
    <ItemData ItemOID="IT.TA.ARMCD" Value="WONDER20"/>
    <ItemData ItemOID="IT.TA.ARM" Value="Miracle Drug 20 mg"/>
    <ItemData ItemOID="IT.TA.TAETORD" Value="3"/>
    <ItemData ItemOID="IT.TA.ETCD" Value="EOS"/>
    <ItemData ItemOID="IT.TA.ELEMENT" Value="End of Study"/>
    <ItemData ItemOID="IT.TA.TABRANCH" Value="Termination from study"/>
    <ItemData ItemOID="IT.TA.EPOCH" Value="TREATMENT"/>
  </ItemGroupData>
</ReferenceData>
```

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.

AVAILABLE TOOLS AND RESOURCES

As part of the implementation of the Dataset-XML standard the first toolsets have been developed. An overview of these tools can be found at the resource website. There are tools available to convert datasets from XML to SAS, from SAS to XML. The package R4CDISC supports functions for reading Dataset-XML and Define.xml into R. SAS announced the support in the next SAS Clinical Standards Toolkit. A Smart Dataset-XML Viewer is also available.

CONCLUSION / SUMMARY

CDISC Dataset-XML is the new data structure for the submission of Clinical Data. It works in conjunction with define-xml, which holds the metadata information. CDISC Dataset-XML is simplified CDISC ODM and able to hold all data standard structures like SDTM, ADaM and SEND as well as any tabular structured data.

First tools are available to support the review and transformation from or to SAS datasets.

REFERENCES

CDISC – Clinical Data Interchange Standards Consortium: www.cdisc.org
CDISC Dataset-XML: http://www.cdisc.org/dataset-xml
CDISC Dataset-XML specification: http://www.cdisc.org/system/files/all/article/application/zip/dataset_xml_1_0_1_.zip
CDISC Dataset-XML Resources: http://wiki.cdisc.org/display/PUB/CDISC+Dataset-XML+Resources
CDISC define.xml specification: http://www.cdisc.org/define-xml

ACKNOWLEDGEMENT

Acknowledgement goes to the 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, DIcore 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

CONTACT INFORMATION

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

Jörg Dillert
Oracle Deutschland B.V: & Co. KG
Schiffbauergasse 14
14467 Potsdam
Work Phone: +49-331 2007388
Email: joerg.dillert@oracle.com
Web: www.oracle.com

Brand and product names are trademarks of their respective companies.
Date of this article: 5th August 2014