

Visualizing Data available in CDISC Dataset-XML Format

Monika Kawohl

Statistical Programming

Accovion GmbH



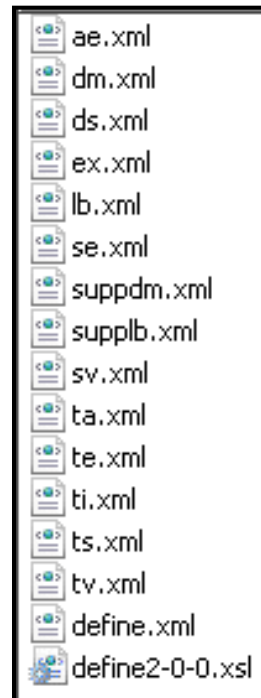
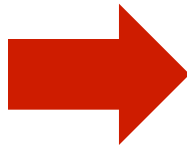
Presentation Overview

CDISC Dataset-XML

- ✓ What is it?
- ✓ Why is it useful?
- ✓ How does it work?
- ✓ In terms of visualization, are there any tools, yet?
- ✓ What are the interfaces to SAS?
("Once the data are available as SAS datasets, we can use the SAS visualization techniques, e.g., the G... procedures.")

What is it?

- ✓ Potential new data transport format for submissions



- ✓ FDA acceptance pending

- Pilot ongoing (about 6 companies were selected for participation)

Why is it useful?

- ✓ Applicable for
 - CDISC SDTM, ADaM, SEND
 - Legacy data
- ✓ SAS Version 5 transport format (XPT) restrictions are no longer an issue

DEMOGRAPHICS - Demographics and baseline characteristics in legacy data format		
Patient number (PATIENT_NUMBER)	Disease history / reason for participating in this study - free text (DISEASE_HTX)	...
1	Very, very long text greater than 200 characters	...



XPT



Dataset-XML



Impact on other CDISC standards?

- ✓ We will still have to adhere to the standards like SDTM, ADaM, SEND
 - Standard dataset and variable labels and names build based on XPT restrictions
- ✓ Some possible improvements for future SDTM, ADaM or SEND versions
 - More meaningful labels, e.g., instead of "Analysis Record Flag 01" we could add information about what is flagged
 - Simplify creating and programatically recognizing ADaM variable pairs, e.g., "numeric counterparts of the primary character variable: ALTBLGR1, ALTBLGR1N
 - When creating new variables, it might be easier to define a name and label
 - No need to split text values longer than 200 character into multiple variables, e.g., comment texts into COVAL, COVAL1, COVAL2, ...
- ✓ However, certain new restrictions may still be useful.

How does it work?

Data: dm.xml

```
...
<ItemGroupData ItemGroupOID="IG.DM"
  sds:ItemGroupDataSeq="1"> 1
  <ItemData ItemOID="IT.STUDYID"
    Value="CDISC01"/> 2
  <ItemData ItemOID="IT.DM.DOMAIN"
    Value="DM"/>
  <ItemData ItemOID="IT.USUBJID"
    Value="CDISC01.100008"/>
  ...
</ItemGroupData>
...
```

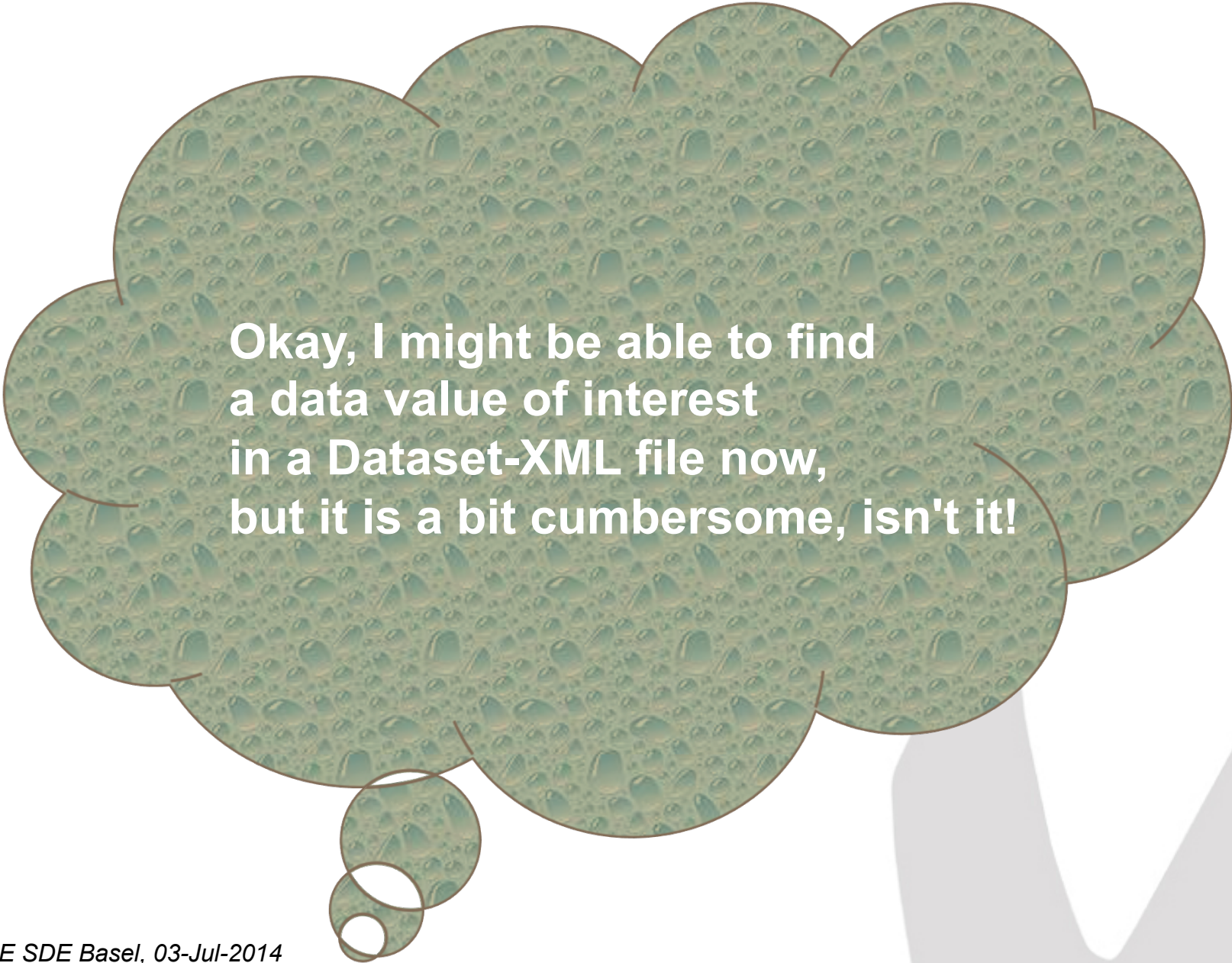
Data and metadata linked via unique OIDs, here: **IG.DM**, **IT.STUDYID**

Metadata: define.xml (Define-XML 2.0)

```
...
<ItemGroupDef OID="IG.DM" Name="DM"...>
  <Description>
    <TranslatedText
      xml:lang="en">Demographics</TranslatedText>
    </Description>
    <ItemRef ItemOID="IT.STUDYID" .../>
    <ItemRef ItemOID="IT.DM.DOMAIN" .../>
    <ItemRef ItemOID="IT.USUBJID" .../>
  ...
</ItemGroupDef>
...
<ItemDef OID="IT.STUDYID" Name="STUDYID" 3
  DataType="text" Length="7"...>
  <Description>
    <TranslatedText xml:lang="en">Study Identifier 4
    </TranslatedText>
  </Description>
  ...
</ItemDef>
...
```

DM (Demographics) dataset in a tabular view (e.g. , SAS)

Obs.	Study Identifier (STUDYID) 3	Domain (DOMAIN)	Unique Subject Identifier (USUBJID)	...
1 1	CDISC01 2	DM	CDISC01.100008	...



Okay, I might be able to find
a data value of interest
in a Dataset-XML file now,
but it is a bit cumbersome, isn't it!

Any tools for visualization support available, yet?

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

Dataset-XML Tool Summary

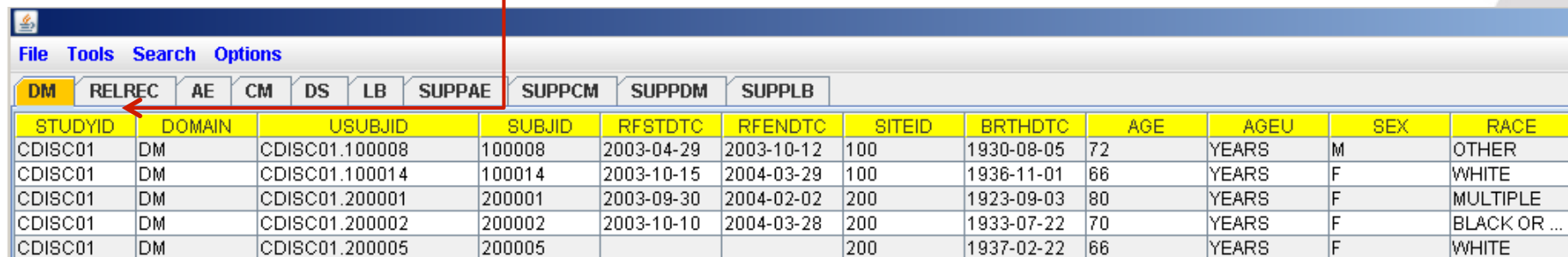
Name	Description	Provided By	Links
XPT2DatasetXML	- Transforms XPT datasets into Dataset-XML datasets - Freely available	XML4Pharma	Available under the Smart SDS-XML View project on source forge
Smart Dataset-XML Viewer	- Similar to the SAS Viewer, but with additional functionality - Supports working 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
EZ Convert	- Converts Dataset-XML files into SAS datasets - Supports Define-XML Version 1 or Version 2 - Open Source	@Sally Cassells	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.	SAS Institute Inc.	SAS Clinical Standards Toolkit
OpenCDISC v1.5	OpenCDISC v1.5 works with Dataset-XML files and Define-XML v2.0	OpenCDISC	OpenCDISC.org

Smart Dataset-XML Viewer

✓ Open Source tool for viewing Dataset-XML data in a tabular format

- Viewing one or more data files
- Sorting by one or more variables
- Changing order of variables via drag and drop
- Filtering/subsetting
- Display metadata as tool tips
- Highlighting relationships
- Export as text file
- Basic validation

Label: Study Identifier
Mandatory = Yes
Datatype: text
Length: 7



STUDYID	DOMAIN	USUBJID	SUBJID	RFSTDTC	RFENDTC	SITEID	BRTHDTC	AGE	AGEU	SEX	RACE
CDISC01	DM	CDISC01.100008	100008	2003-04-29	2003-10-12	100	1930-08-05	72	YEARS	M	OTHER
CDISC01	DM	CDISC01.100014	100014	2003-10-15	2004-03-29	100	1936-11-01	66	YEARS	F	WHITE
CDISC01	DM	CDISC01.200001	200001	2003-09-30	2004-02-02	200	1923-09-03	80	YEARS	F	MULTIPLE
CDISC01	DM	CDISC01.200002	200002	2003-10-10	2004-03-28	200	1933-07-22	70	YEARS	F	BLACK OR ...
CDISC01	DM	CDISC01.200005	200005			200	1937-02-22	66	YEARS	F	WHITE

Smart Dataset-XML Viewer - GUI

Standard: SDTM

Define.xml: ataset-XML_1-0\Example\Untyped_Data_Example-sdtm\define2-0-0-example-sdtm(2013-11-09).xml Browse

Define.xml version: ☒ 2.0 ☐ 1.0

Dataset-XML data files:

- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\ae.xml
- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\cm.xml
- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\dm.xml
- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\ds.xml
- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\lb.xml
- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\relrec.xml
- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\suppae.xml
- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\suppcm.xml
- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\suppdm.xml
- ISC-Dataset-XML 1.0 (Final)\Dataset-XML_1-0\Example\Untyped_Data_Example-sdtm\supplb.xml

Add Remove Clear

☐ Use TYPED ItemData (ItemDataString, ItemDataDate, ...)

☐ Bring SUPPQUAL data back to original dataset

Progress: 4% 10/10 files read 100% % validation done

Start Interrupt

Select Standard
Select define file

Select 1 or more
Dataset-XML
files to be
viewed

Load the data

Well, wait, we
may want to set
some validation
options first.

Smart Dataset-XML Viewer - Options

Options

?

☒ Check Study OID against define.xml
☐ Check uniqueness of USUBJID in DM/ADSL dataset
☐ Check USUBJID versus DM/ADSL dataset
☐ Check dataType correctness
☐ Check age from birthdate and reference start date
☐ Check values against reference range
☒ Check values against codelist
☐ Check uniqueness of USUBJID/SEQ combinations - SLOW

☐ Show date of first study treatment exposure in DM
☐ Show date of last study treatment exposure in DM
☐ Show --DY value on --DTC
☐ Get and show VISIT name on VISITNUM

☐ Sort tables automatically by USUBJID and --SEQ

☐ Use data caching for very large tables

OK

Cancel

Cells violating the selected validation criteria will be highlighted in red (ERRORS) or orange (WARNINGS) in the data tables

AEREL	AEOU	AESCAN	AESCONG	AESDISAB	AESDTH
PROBABLE	NOT RECO...	N	N	N	N
PROBABLE	NOT RECO...	N	N	N	N
REMOTE	RECOVER...	N	N	N	N
POSSIBLE	NOT RECO...	N	N	N	N
POSSIBLE	NOT RECO...	N	N	N	N
IMPOSSIBLE	NOT RECO...	N	N	N	N
POSSIBLE	RECOVER...	N	N	N	N
POSSIBLE	ERROR: Value IMPOSSIBLE is not in codelist [AECAUS] (AEREL)				
PROBABLE	NOT RECO...	N	N	N	N

Smart Dataset-XML Viewer - Subsetting

STUDYID	DOMAIN	USUBJID	SUBJID	AGE
CDISC01	DM	CDISC01.1...	100014	66
CDISC01	DM	CDISC01.2...	200005	66
CDISC01	DM	CDISC01.2...	200002	70
CDISC01	DM	CDISC01.1...	100008	72
CDISC01	DM	CDISC01.2...	200001	80

Select USUBJID

☐ All Subjects

☐ Subjects from current view

☒ All currently selected Subjects (3 Subjects)

CDISC01.100008
CDISC01.100014
CDISC01.200001
CDISC01.200002
CDISC01.200005

☒ Apply to all datasets

Filter title:
Subjects with age >=70

OK Cancel

- ✓ Sort DM by age
- ✓ Select subjects of interest (e.g., age >=70)
- ✓ Select "Tools - Filtering - Filter on USUBJID"
- ✓ Choose "All currently selected Subjects"
- ✓ Filter can be named and applied to all datasets
- ✓ Display of the filtered data subset
- ✓ Filter can be expanded by additional conditions, e.g., "Subjects with age >=70 and severe AEs"
 - go to the AE worksheet
 - sort by severity
 - proceed as shown for the age based selection

Filter: Subjects with age >=70

File Tools Search Options

DM	AE	CM	DS	EX	SUPPAE	SUPPCM	SUPPDM	SUPI
STUDYID	DOMAIN	USUBJID	SUBJID	AGE				
CDISC01	DM	CDISC01.100008	100008	72				
CDISC01	DM	CDISC01.200001	200001	80				
CDISC01	DM	CDISC01.200002	200002	70				

Smart Dataset-XML Viewer - Showing Relations

File Tools Search Options

DM	AE	CM	DS	EX	SUPPAE	SUPPCM	SUPPDM	SUPPVS	VS	RELREC
STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	RELTYPE	RELID				
CDISC01	AE	CDISC01.200001	AESPID	5		1				
CDISC01	DS	CDISC01.200001	DSSEQ	3		1				

Message

There are 2 related records with RELID '1':
 Domain = AE - USUBJID = CDISC01.200001 - AESPID = 5
 Domain = DS - USUBJID = CDISC01.200001 - DSSEQ = 3

OK

- ✓ In Worksheet RELREC, click on a record of interest
- ✓ Select "Tools - Show related records"
- ✓ A message about the related records is displayed

DM	AE	CM	DS	EX	SUPPAE	SUPPCM	SUPPDM	SUPPVS	VS	RELREC
STUDYID	DOMAIN	USUBJID	AESID	AESPID	AETERM					
CDISC01	AE	CDISC01.100008	1	1	AGITATED					
CDISC01	AE	CDISC01.100008	2	2	ANXIETY					
CDISC01	AE	CDISC01.100008	3	3	DECREASED APPETITE					
CDISC01	AE	CDISC01.100014	1	1	DIARRHEA					
CDISC01	AE	CDISC01.100014	2	2	HEMORRHOIDS					
CDISC01	AE	CDISC01.100014	3	3	HEADACHE					
CDISC01	AE	CDISC01.100014	4	4	VOMIT					
CDISC01	AE	CDISC01.100014	5	5	VOMIT					
CDISC01	AE	CDISC01.200001	1	1	ANXIETY					
CDISC01	AE	CDISC01.200001	2	5	LEFT KNEE PAIN WORSENING					
CDISC01	AE	CDISC01.200001	3	3	CONSTIPATION					

- ✓ The respective records in the parent datasets are highlighted in green

DM	AE	CM	DS	EX	SUPPAE	SUPPCM	SUPPDM	SUPPVS	VS	RELREC
STUDYID	DOMAIN	USUBJID	DSSEQ	DSTERM						
CDISC01	DS	CDISC01.100008	1	INFORMED CONSENT OBTAINED						
CDISC01	DS	CDISC01.100008	2	RANDOMIZED						
CDISC01	DS	CDISC01.100008	3	COMPLETED						
CDISC01	DS	CDISC01.100014	1	INFORMED CONSENT OBTAINED						
CDISC01	DS	CDISC01.100014	2	RANDOMIZED						
CDISC01	DS	CDISC01.100014	3	COMPLETED						
CDISC01	DS	CDISC01.200001	1	INFORMED CONSENT OBTAINED						
CDISC01	DS	CDISC01.200001	2	RANDOMIZED						
CDISC01	DS	CDISC01.200001	3	DISCONTINUED DUE TO ADVERSE EVENT						
CDISC01	DS	CDISC01.200002	1	INFORMED CONSENT OBTAINED						

- ✓ Similarly, parent records for selected data in supplemental qualifier datasets can be highlighted

What are the interfaces to SAS?

✓ Refer to list of Dataset-XML tools on CDISC Wiki

Future Version of SAS Clinical Standards Toolkit (CST)	• Converts SAS datasets into Dataset-XML files and vice versa • Validates Dataset-XML files
OpenSource Tool: EZ Convert	• Converts Dataset-XML files into SAS datasets or SAS programs to create the respective datasets
OpenSource Tool: XPT2DatasetXML	• (Converts XPT files into Dataset-XML files)

✓ DIY - Do It Yourself!

- Write macros
 - Writing: sas2datasetxml
 - Reading: datasetxml2sas

Custom SAS code to write Dataset-XML

- ✓ Dastep programming, i.e., write XML files with PUT statements (one of other options)
 - Nest the following elements:
 1. Write **XML header**
 2. Specify the root **ODM** element (e.g., incl. Study information)
 3. Specify **ClinicalData** or **ReferenceData** element depending on dataset contents
 - ClinicalData for subject data (e.g., DM, EX, VS, AE)
 - ReferenceData for non-subject data (e.g., trial design domains: TA, TS, etc.)
 4. Write **ItemGroupData** element for each record
 - Naming convention for ItemGroupOID: IG.<dataset name>
 5. Write **ItemData** element for each non-missing data value within a record
 - Naming convention for ItemOID: IT.<dataset name>.<variable name>

Note: define.xml not needed as input if we follow the same OID naming conventions

Custom SAS code to read Dataset-XML

- ✓ The more interesting and challenging part...
 - Needed in order to use the visualization procedures we are familiar with
- ✓ Here is what you could do:
 - Use define.xml to create dataset templates
 - Use the Dataset-XML file to populate the dataset with the data values

Obs	Study Identifier (STUDYID)	Domain (DOMAIN)	Unique Subject Identifier (USUBJID)	...
1	CDISC01	DM	C01-1001	
2	CDISC01	DM	C01-1002	

← define.xml

← dm.xml

```
<ItemGroupData ItemGroupOID="IG.DM" data:ItemGroupDataSeq="1">
  <ItemData ItemOID="IT.STUDYID" Value="CDISC01"/>
  <ItemData ItemOID="IT.CM.DOMAIN" Value="DM"/>
  <ItemData ItemOID="IT.USUBJID" Value="C01-1001"/>
  ...
</ItemGroupData>
```

Extracting data/metadata from XML files

✓ PROC XSL (extract information from XML and transform acc. to XLS into OUT)

```
*) Generate SAS program which writes the required metadata from define into dataset METADATA;
*) Structure of dataset METADATA: ITEMGROUPOID (dataset identifier), MEMNAME, MEMLABEL,;
*) VARNUM, ITEMDEFOID (variable identifier), NAME, LABEL, TYPE, LENGTH ;

PROC XSL IN="define.xml" XSL="read-metadata.xsl" OUT="read-metadata.sas"; RUN;
```

✓ Output: Program read-metadata.sas

```
data metadata;
length itemgroupoid $200 memname $8 memlabel $40
      varnum 8 itemdefoid $200 name $8 label $40 type $8 length 8;

itemgroupoid="IG.DM"; memname="DM"; memlabel="Demographics";

varnum=1; itemdefoid="IT.STUDYID"; name="STUDYID";
      label="Study Identifier"; type="text"; length=7; output;
varnum=2; itemdefoid="IT.DM.DOMAIN"; name="DOMAIN";
      label="Domain Abbreviation"; type="text"; length=2; output;
varnum=3; itemdefoid="IT.USUBJID"; name="USUBJID";
      label="Unique Subject Identifier"; type="text"; length=14; output;
...
```

datasetxml2sas

- ✓ Use previously described approach to generate metadata and data datasets

Metadata Dataset

ITEMGROUPOID	MEMNAME	MEMLABEL	VARNUM	ITEMDEFOID	NAME	LABEL	TYPE	LENGTH
IG.DM	DM	Demographics	1	IT.STUDYID	STUDYID	Study Identifier	text	7
IG.DM	DM	Demographics	2	IT.DM.DOMAIN	DOMAIN	Domain Abbreviation	text	2

Data Dataset

ITEMGROUPOID	RECORDNO	ITEMDEFOID	VALUE
IG.DM	1	IT.STUDYID	CDISC01
IG.DM	1	IT.DM.DOMAIN	DM

- ✓ Merge and create comma-delimited text files with data, e.g.,

dm.dat

```
CDISC01,DM,CDISC01.100008,100008,2003-04-29,2003-10-12,100,1930-08-05,72,YEARS,M,OTHER, ...
CDISC01,DM,CDISC01.100014,100014,2003-10-15,2004-03-29,100,1936-11-01,66,YEARS,F,WHITE, ...
...
```

datasetxml2sas (continued)

- ✓ Use previously described approach to generate metadata and data datasets

Metadata Dataset

ITEMGROUPOID	MEMNAME	MEMLABEL	VARNUM	ITEMDEFOID	NAME	LABEL	TYPE	LENGTH
IG.DM	DM	Demographics	1	IT.STUDYID	STUDYID	Study Identifier	text	7
IG.DM	DM	Demographics	2	IT.DM.DOMAIN	DOMAIN	Domain Abbreviation	text	2

Data Dataset

ITEMGROUPOID	RECORDNO	ITEMDEFOID	VALUE
IG.DM	1	IT.STUDYID	CDISC01
IG.DM	1	IT.DM.DOMAIN	DM

- ✓ Merge and create comma-delimited text files with data, e.g.,

dm.dat

```
CDISC01,DM,CDISC01.100008,100008,2003-04-29,2003-10-12,100,1930-08-05,72,YEARS,M,OTHER, ...
CDISC01,DM,CDISC01.100014,100014,2003-10-15,2004-03-29,100,1936-11-01,66,YEARS,F,WHITE, ...
...
```

datasetxml2sas (continued)

- ✓ Create Program to read the data files (.dat) and execute

Program read-dm.sas

```
data sdtm.DM (label="Demographics");

infile "dm.dat" dsd truncover lrecl=2000;

input
  STUDYID : $7.      DOMAIN : $2.      USUBJID : $14.      SUBJID : $6.      RFSTDTC : $10.
  RFENDTC : $10.     SITEID : $3.      BRTHDTC : $10.      AGE      AGEU : $5.      SEX : $1.
  RACE : $40.        ETHNIC : $22.     ARMCD : $8.        ARM : $20.        COUNTRY : $3.;

label STUDYID = "Study Identifier";
label DOMAIN  = "Domain Abbreviation";
label USUBJID = "Unique Subject Identifier";
label SUBJID  = "Subject Identifier for the Study";
label RFSTDTC = "Subject Reference Start Date/Time";
label RFENDTC = "Subject Reference End Date/Time";
label SITEID  = "Study Site Identifier";
label BRTHDTC = "Date/Time of Birth";
label AGE     = "Age";
label AGEU    = "Age Units";
label SEX     = "Sex";
label RACE    = "Race";
label ETHNIC  = "Ethnicity";
label ARMCD   = "Planned Arm Code";
label ARM     = "Description of Planned Arm";
label COUNTRY = "Country";

run;
```

Dataset-XML in Summary

✓ Structure

- Data values in <dataset name>.xml
- Metadata (name, label, etc.) in define.xml (Define-XML 1.0 or 2.0)
- Linked via OIDs

✓ Benefit: No XPT-format restrictions

✓ Acceptance by regulatory authorities pending

✓ OpenSource viewer tool available

- Smart Dataset-XML Viewer

✓ SAS Interfaces

- Future Version of SAS Clinical Standards Toolkit (CST)
- If necessary, you could do it on your own

Thank You!

Questions?

ACCOVION

Monika Kawohl

Principal Statistical
Programmer

www.accovion.com

Accovion GmbH
Softwarecenter 3
D-35037 Marburg, Germany

Tel. +49 6421 948 49-20

monika.kawohl@accovion.com