

SDTM FILE SIZE ISSUE (SAS LENGTH) UPDATE3

CJUG SDTM Team

LISaS Learning Industry Standard around SDTM

17th Jan 2013

Introduction

- CDISC standardized datasets are increasing file sizes of submissions using SAS transport v5.
 - File size limitations of available tools and machines hinder ability to conduct timely regulatory reviews
 - Over 650 datasets submitted / week to FDA CDER

Ref: 1) SDTM Column Resizing: Background and Industry Testing Results; Electronic Data (eData) Team, CDER FDA, October 13, 2011

- The “SDTM File Size Issue (SAS Length)” topic was discussed at the CJUG SDTM meeting on Sep. 14 2012, Apr. 12 2013 and May. 21 2013.
- This slide deck is focused on the following contents;
 1. Data Sizing Best Practices Recommendation
 2. StudyDataSet-XML (SDS-XML) Specification



DATA SIZING BEST PRACTICES RECOMMENDATION (DRAFT)

Based on The SDTM Validation Rules Project in the
FDA/PhUSE CSS Data Quality Working Group

Note : The Japanese translation was created and shared with CJUG-SDTM Team already.

Scope

- Two main factors that contribute to large data sets;
 1. The number of observations in a data set
 2. The space allocated to individual variables
- Scope
 - Process flow for managing the recommended solutions
 - How to manage the length of character values to avoid wasted space within datasets?
 - How to handle SAS xpt files when they exceed the maximum size allowed?
 - What to report in define documentation?

Process

1. Variable sizing process
2. Data set size examination
3. Split (based on the threshold)
 - Keeping in mind that these recommendations are strictly for the purpose of submission.
 - These practices as a final step in preparing submission data.

Start a final step in preparing submission data

Variable sizing

Flowchart Example

*1: In the CDER Common Data Standards Issues Document (v1.1), Sponsor splits the datasets into smaller datasets no larger than 1 GB in size.

*2: In this example, if a parent domain is smaller than 1.25GB but associated SUPP domain is larger than the criteria, then only SUPP domain is split even if a parent domain is not split.

*3: An associated SUPP/FA domain always be split, regardless of its size, according to the criteria by which the parent domain is split. The records in the split SUPP/FA domain correspond to the matching records in the split parent domain.

A dataset is larger than 1.25 GB? *1

Parent or SUPP/FA domain? *2

Parent

Split Parent

Split SUPP/FA

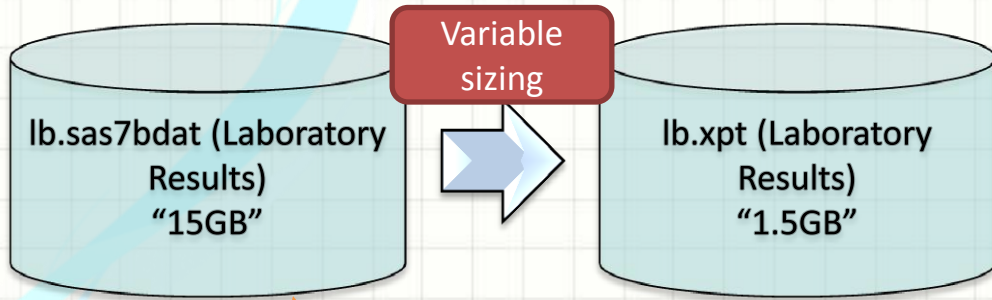
Exist SUPP/FA domain? *3

Split SUPP/FA

Create define.xml and Submit

- Normally XPT(.xpt) is smaller than SAS dataset (.sas7bdat).

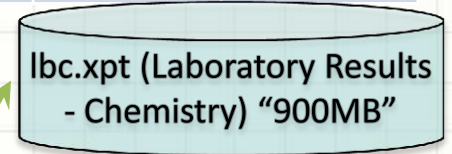
How to Split



LBCAT leng=200	LBTESTCD Leng=8	LBORRES leng=200
HEMA	HCT	20.0
CHEM	ALT	30

LBCAT leng=4	LBTESTCD leng=3	LBORRES leng=4
HEMA	HCT	20.0
CHEM	ALT	30

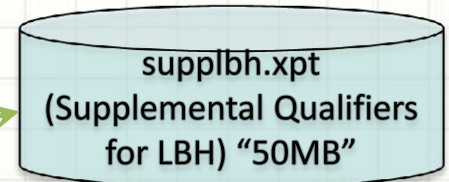
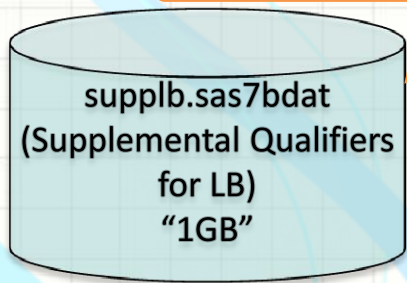
LBCAT leng=4	LBTESTCD leng=3	LBORRES leng=4
HEMA	HCT	20.0



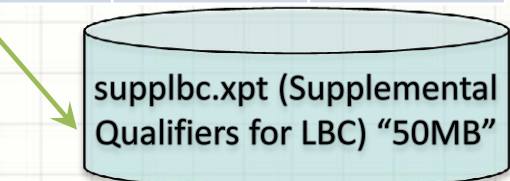
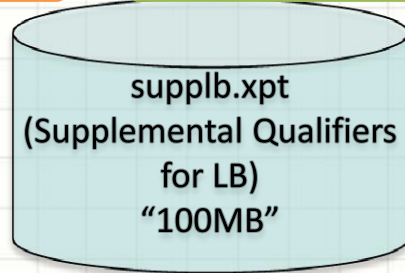
LBCAT leng=4	LBTESTCD leng=3	LBORRES leng=4
CHEM	ALT	30

Operational SDTM

Submission SDTM



IDVAR leng=5	IDVARVAL leng=1	QVAL leng=8
LBSEQ	1	Abnormal



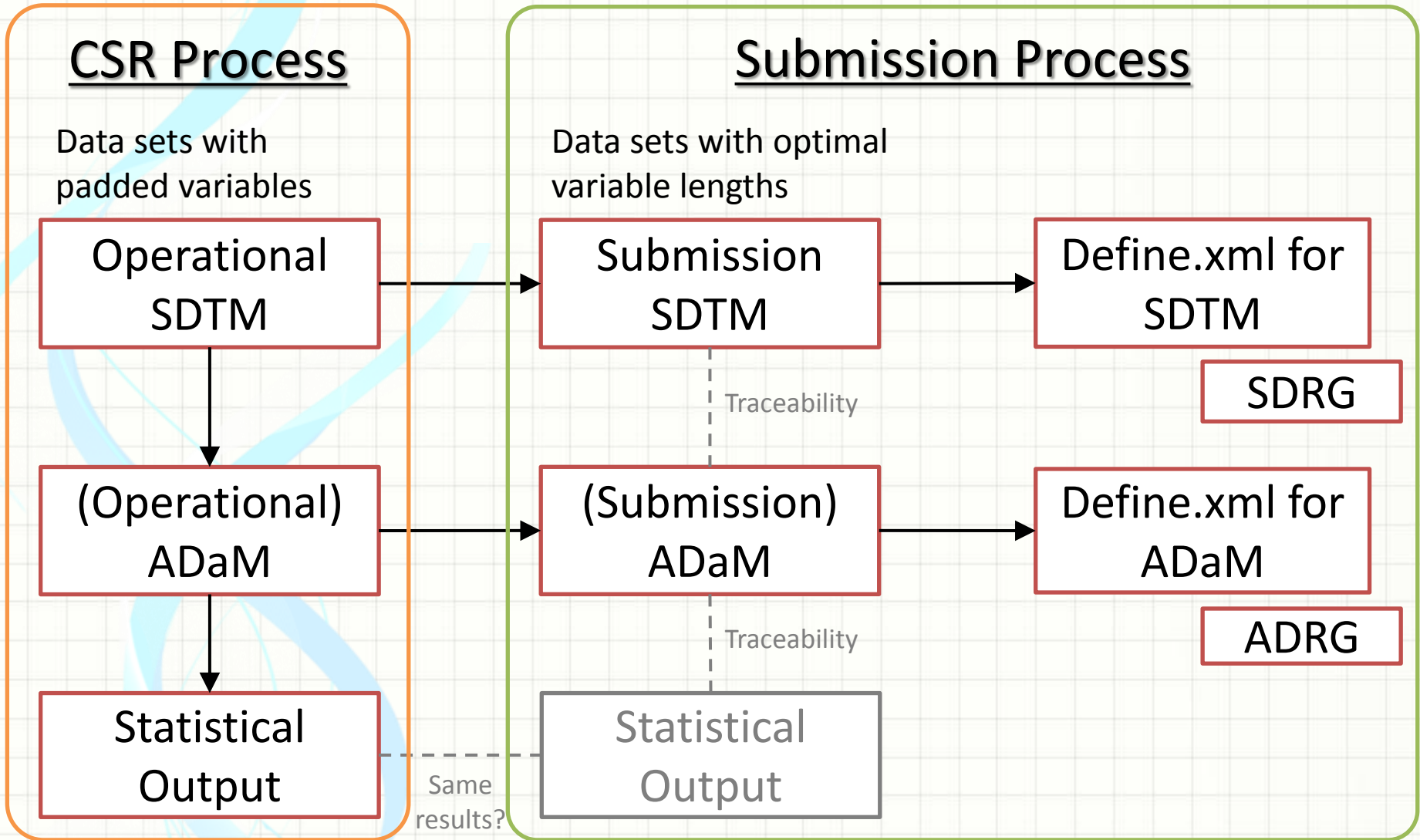
IDVAR leng=5	IDVARVAL leng=1	QVAL leng=8
LBSEQ	2	Abnormal

IDVAR leng=8	IDVARVAL leng=200	QVAL leng=200
LBSEQ	1	Abnormal
LBSEQ	2	Abnormal

IDVAR leng=5	IDVARVAL leng=1	QVAL leng=8
LBSEQ	1	Abnormal
LBSEQ	2	Abnormal

Parallel Method Example

Ref: 8) Traceability between SDTM and ADaM, Mamiko Hayashi, CJUG SDTM KANSAI, 28 November 2013



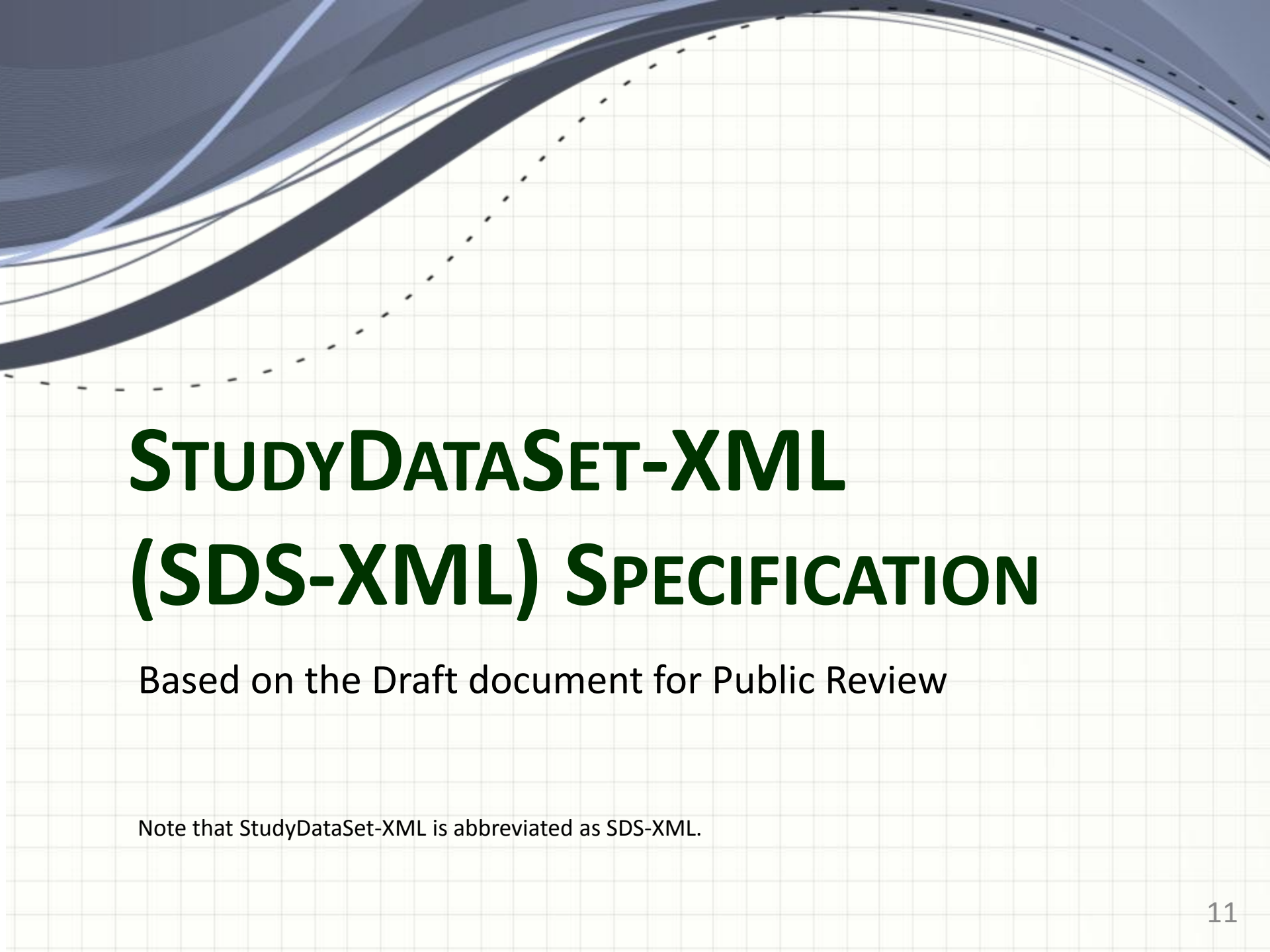
Note that this recommendation is focused on the SDTM only, No mention of ADaM.

Define Document

- “Length” is required if data type is;
 - text
 - integer
 - float
- Sponsors must reflect in the define documentation the actual variable and value level lengths in the submitted files, or in other words, the reduced lengths.
- For split domains, metadata should be provided for each data set separately.

Action Items

- (OpenCDISC) An Information note is issued for data sets between 1 and 1.25 GB in future version.
- (CDISC) Change SDTM IG 4.1.2.9 and any other suggestions of mandated variable lengths.
- (OpenCDISC) Some updates regarding Split domains.



STUDYDATASET-XML (SDS-XML) SPECIFICATION

Based on the Draft document for Public Review

Note that StudyDataSet-XML is abbreviated as SDS-XML.

Why SDS-XML?

- In 1999, the FDA standardized the submission of clinical and non-clinical data using the SAS XPORT Transport Format v5.
- However there are many limitations of the current exchange format : SAS XPT v5.
- The purpose of StudyDataSet-XML is to support the interchange of tabular clinical research data using ODM-based XML technologies.
 - This was presented as an ODM-based alternative solution in the FDA meeting on November 5, 2012.

Limitations of SAS XPT v5

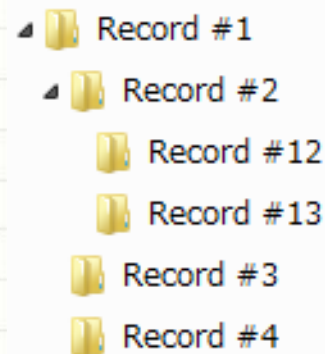
Ref: 9) FDA public meeting entitled "Regulatory New Drug Review: Solutions for Study Data Exchange Standards" on November 5, 2012

– Technical perspective;

- Limitations for variable names (8 char), variable name characters (Traditional alphanumeric only), labels (40 chars), character fields (200 bytes)
- Large file sizes due to "empty space"

– Structural perspective;

- Two-dimensional "flat" data structure for hierarchical/multi-relational "round" data; lack of a robust information model with the standard. Important meaning is lost when exchanging 2-dimensional flat files, making some interpretations and analyses difficult or impossible, i.e. decreased semantic interoperability.
- The solution will also necessitate a shift in how data are collected.
- More challenging to solve.



Data Submission Process

- FDA will convert (de-serialize) to SAS datasets for their review.

SDS-XML for Data Transport

Convert SAS
Datasets to SDS-
XML

Send SDS-XML

Receive SDS-
XML

Convert to SAS
Datasets or load
into a data
warehouse

Data Transport

Other topics

- SDS-XML can be used to transmit SDTM, ADaM and SEND datasets.
- This specification provides two ways to represent data - “Typed” and “Untyped”. Either of these methods may be used with SDS-XML.
 - Typed data: using elements named ItemData[Type], e.g., integer variable -> ItemDataInteger.
 - You can ensure each data value has the correct datatype in the define.xml
 - Untyped data: using elements named ItemData are used for all data regardless of the datatype
- It could easily use Japanese characters rather than XPTs.
 - However draft SDS-XML Smart Viewer and OpenCDISC cannot read.



Q&A

CJUG SDTM Team

LISaS Learning Industry Standard around SDTM

17th Jan 2013

References

- 1) SDTM Column Resizing: Background and Industry Testing Results; Electronic Data (eData) Team, CDER FDA, October 13, 2011
http://www.cdisc.org/stuff/contentmgr/files/0/4f05d8426369051905a247002c87e38e/files/dhananjay_chhatre_session_9.pdf
- 2) SDTM File size issue (SAS Length), CJUG SDTM Team, September 14, 2012
- 3) SDTM File size issue (SAS Length) UPDATE, CJUG SDTM Team, April 12, 2013
- 4) SDTM File size issue (SAS Length) UPDATE2, CJUG SDTM Team, May 21, 2013
- 5) Data Sizing Best Practices Recommendation / PhUSE
http://www.cdisc.org/stuff/contentmgr/files/0/4f05d8426369051905a247002c87e38e/files/dhananjay_chhatre_session_9.pdf
- 6) CDISC Standards Webinar - Latest Updates and Additions, November 21, 2013
- 7) StudyDataSet-XML Specification Version 1.0 Draft for Public Review
- 8) Traceability between SDTM and ADaM, Mamiko Hayashi, CJUG SDTM KANSAI, 28 November 2013
- 9) FDA public meeting entitled “Regulatory New Drug Review: Solutions for Study Data Exchange Standards” on November 5, 2012
<http://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm332003.htm>