PharmaSUG 2019 - Paper SS-117

Have You Met Define.xml 2.0?

Christine McNichol, Covance Inc.

ABSTRACT

Define.xml is critical for a reviewer getting to know a study's datasets. The define.xml facilitates the building of this relationship with the study data from casual introduction through the innermost workings of the datasets. To provide the best possible define.xml for reviewer's use, it is important to be comfortable with define.xml and how it communicates information about the datasets. Though define.xml might be intimidating at first, fear not. Following the flow through define.xml to get to know the data is like getting to know a new friend. Define.xml reveals levels of information about the study datasets from structure to details about where the values came from, even linking to documentation of issues encountered and decisions made. The anatomy of a define.xml, its purpose, what is special about each section and what can be learned about the study datasets from each will be discussed. Meet define.xml – friend, not foe.

INTRODUCTION

Define.xml is an important tool in communicating the structure and content of the study SDTM and ADaM datasets. In order to ensure that this information is communicated correctly and completely, it is critical to understand the parts of the define.xml and what each part is expressing. This paper will take a non-technical look at define.xml with the intent to provide understanding of what will be shown in the define.xml and how it relates to the data and to other sections. While this paper will not specifically discuss how to fill out spreadsheets or create define.xml, it will walk through the output that appears in the define.xml visual once created. After getting to know the define.xml output in this way, it will become clear how the datasets and derivations will appear to an end user such as a reviewer. This understanding will help to make that process of filling out input spreadsheets or coming up with the content to feed into a define.xml generation tool easier and ensure that it presents the most useful information for review.

DEFINE.XML BACKGROUND

Before the introductions to the study define.xml begin, it will be helpful to have some general background information about define.xml, what it is and its overall purpose.

WHAT IS A DEFINE.XML? WHAT IS ITS PURPOSE? WHO USES IT?

Define.xml contains information that describes the structure and content of the datasets used to describe a study. It can be created for either SDTM or ADaM datasets. The purpose of the define.xml is to describe for an end user the details about the datasets created for the study data, including structures, values and the source of those values whether that is from the CRF or derivations or another source. The define.xml is created with an agency reviewer in mind as the target audience in most cases, however the define.xml can also be useful for internal reviewers, new team members, and as documentation for longer running studies.

HOW DOES THE DEFINE.XML DIFFER FROM DATASET PROGRAMMING SPECIFICATIONS?

While dataset specifications and define.xml may have some similarities, they do have some important differences in content and purpose. Both will outline the study datasets, naming, attributes, derivations and content. But the purpose of dataset specifications is to describe how datasets and variables should be created. They might contain a bit of code or reference to some macros that must be used. But keeping an external reviewer in mind, the define.xml should describe derivations with text as needed for clarity and may not reference the specific macros or have heavy code bits. Programming specifications are the instructions. The define.xml on the other hand is not instruction how to program, but rather documentation of what was done and what has actually been created. It describes for the reviewer what has occurred and how the data arrived at the values and structures. The specs might contain a list of datasets and variables in the standard that may or may not be included depending on the presence of data, but the define.xml will only describe those that are submitted. For example, there may be a dataset programming specification for the Inclusion/Exclusion (IE) domain for a study. However, under current guidance, if there were no inclusion/exclusion violations present in the study, then the IE domain would be empty and would not be submitted. The define.xml itself would not contain any reference to that empty IE domain that was not submitted.

It is important to keep in mind that any text and description is being created on behalf of the study sponsor in the define.xml and supporting documentation to be submitted. If the define.xml or documentation is created by a CRO or someone other than the study sponsor, it would still be written in the voice of the study sponsor.

WHO NEEDS TO BE FAMILIAR WITH THE DETAILS OF WHAT IS DISPLAYED IN A DEFINE.XML?

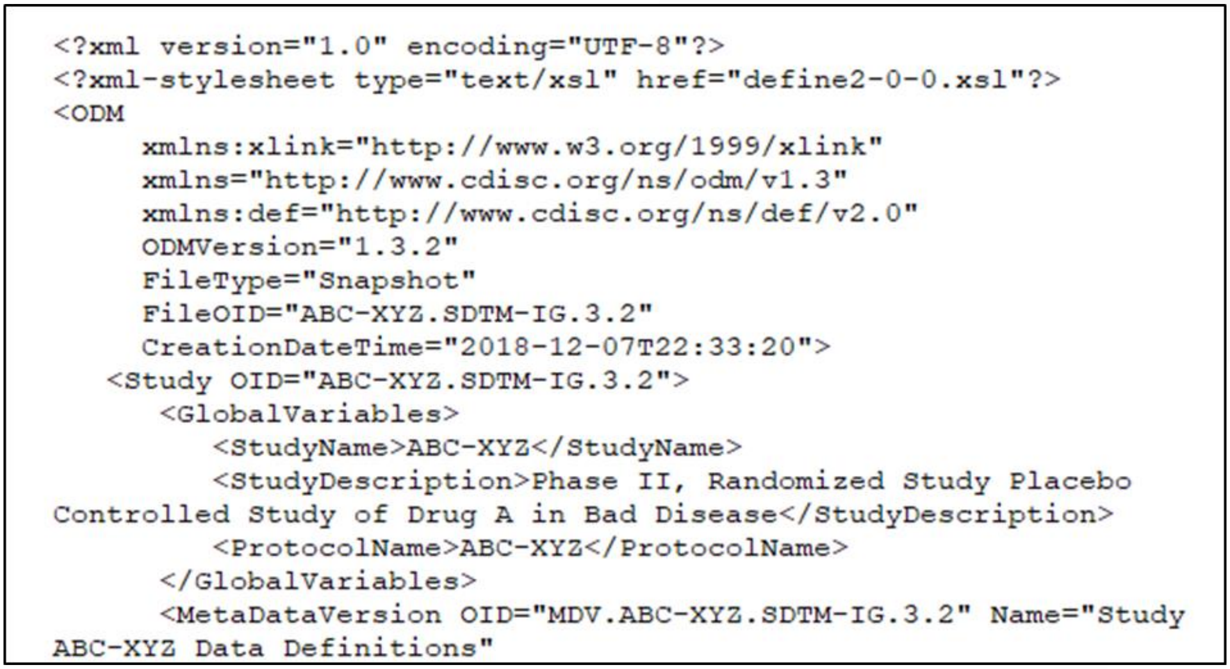
Companies approach define.xml creation in different ways. Sometimes this is created by the study team that creates the SDTM or ADaM datasets and other times a define.xml may be created by a core group of define.xml creation specialists or it may be outsourced. The reality is that those individuals ultimately responsible for the study may have little prior knowledge of the generation of the define.xml and its content. So, if you are not one of the individuals that creates the define.xml, do you really need to know the details about what it contains and how it describes the study? Of course!

Even if a team of specialists does the actual generation of the define.xml, it is important to understand define.xml in order to review the information that has been compiled and fully understand how the study is being presented to the reviewer through the information in the define.xml. The programmer and/or validator of a dataset may be asked to review sections of the define.xml for accuracy of derivations. A lead programmer, statistician, or oversight person for the study or compound may need to review for content, look for consistency, and check that the information for the study is presented to the reviewer in the most accurate and clearest way possible.

WHAT DOES DEFINE.XML LOOK LIKE?

The define.xml itself is a file composed of many lines of machine-readable xml code. The information contained in this code can be read, but not very easily.

Figure 1 is sample of define.xml code.

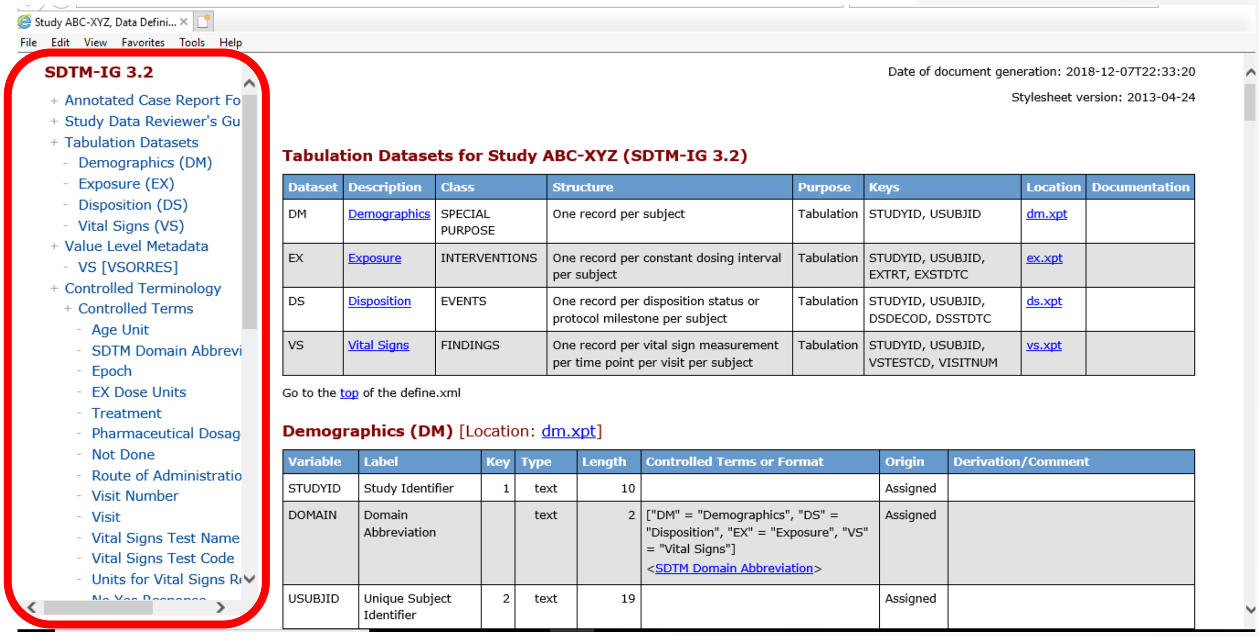


```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="define2-0-0.xsl"?>
<ODM
     xmlns:xlink="http://www.w3.org/1999/xlink"
     xmlns="http://www.cdisc.org/ns/odm/v1.3"
     xmlns:def="http://www.cdisc.org/ns/def/v2.0"
     ODMVersion="1.3.2"
     FileType="Snapshot"
     FileOID="ABC-XYZ.SDTM-IG.3.2"
     CreationDateTime="2018-12-07T22:33:20">
   <Study OID="ABC-XYZ.SDTM-IG.3.2">
      <GlobalVariables>
         <StudyName>ABC-XYZ</StudyName>
         <StudyDescription>Phase II, Randomized Study Placebo
Controlled Study of Drug A in Bad Disease</StudyDescription>
         <ProtocolName>ABC-XYZ</ProtocolName>
      </GlobalVariables>
      <MetaDataVersion OID="MDV.ABC-XYZ.SDTM-IG.3.2" Name="Study
ABC-XYZ Data Definitions"
```

Figure 1. Define.xml Code

To aid in the viewing of the define.xml, a stylesheet is used. The stylesheet is a file that has extension .xsl and will be stored in the same folder as the define.xml. It is the tool that allows the define.xml code to appear as something easily human-readable instead of just lines of xml code. When the define.xml is viewed using the stylesheet, the appearance is more user friendly. A stylesheet can be found with the define.xml 2.0 package. Recently, a new stylesheet from a PhUSE working group has been made available as well. Though the information shown by each is very similar, the appearance and some naming differs between the two. Where there is a notable difference, both will be displayed, otherwise the examples given are using the stylesheet that was included with the define.xml 2.0 package. Remember that the stylesheet is just changing the way the define.xml looks, but the underlying define.xml itself has not changed in the examples.

Figure 2 shows define.xml viewed with a stylesheet:


Study ABC-XYZ, Data Defini...
File Edit View Favorites Tools Help

SDTM-IG 3.2

+ Annotated Case Report Fo
+ Study Data Reviewer's Gu
+ Tabulation Datasets
- Demographics (DM)
- Exposure (EX)
- Disposition (DS)
- Vital Signs (VS)
+ Value Level Metadata
- VS [VSORRES]
+ Controlled Terminology
+ Controlled Terms
- Age Unit
- SDTM Domain Abbrevi
- Epoch
- EX Dose Units
- Treatment
- Pharmaceutical Dosag
- Not Done
- Route of Administratio
- Visit Number
- Visit
- Vital Signs Test Name
- Vital Signs Test Code
- Units for Vital Signs R
- No Yes Response

Date of document generation: 2018-12-07T22:33:20

Stylesheet version: 2013-04-24

Tabulation Datasets for Study ABC-XYZ (SDTM-IG 3.2)

Dataset	Description	Class	Structure	Purpose	Keys	Location	Documentation
DM	Demographics	SPECIAL PURPOSE	One record per subject	Tabulation	STUDYID, USUBJID	dm.xpt	
EX	Exposure	INTERVENTIONS	One record per constant dosing interval per subject	Tabulation	STUDYID, USUBJID, EXTRT, EXSTDTC	ex.xpt	
DS	Disposition	EVENTS	One record per disposition status or protocol milestone per subject	Tabulation	STUDYID, USUBJID, DSDECOD, DSSTDTC	ds.xpt	
VS	Vital Signs	FINDINGS	One record per vital sign measurement per time point per visit per subject	Tabulation	STUDYID, USUBJID, VSTESTCD, VISITNUM	vs.xpt	

Go to the top of the define.xml

Demographics (DM) [Location: dm.xpt]

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin	Derivation/Comment
STUDYID	Study Identifier	1	text	10		Assigned	
DOMAIN	Domain Abbreviation		text	2	["DM" = "Demographics", "DS" = "Disposition", "EX" = "Exposure", "VS" = "Vital Signs"] <SDTM Domain Abbreviation>	Assigned	
USUBJID	Unique Subject Identifier	2	text	19		Assigned	

Figure 2. Define.xml Viewed with a Stylesheet

On the left of the define.xml view, a list with links can be found to help navigate forward and backward to each section of the define.xml file and to some external files that will also describe the study and its datasets. The links to the sections of the define.xml itself will be present in both SDTM and ADaM define.xml. SDTM define.xml will also have links to external files of the annotated case report form and the Study Data Reviewer's Guide. ADaM define.xml will link to the Analysis Data Reviewer's Guide.

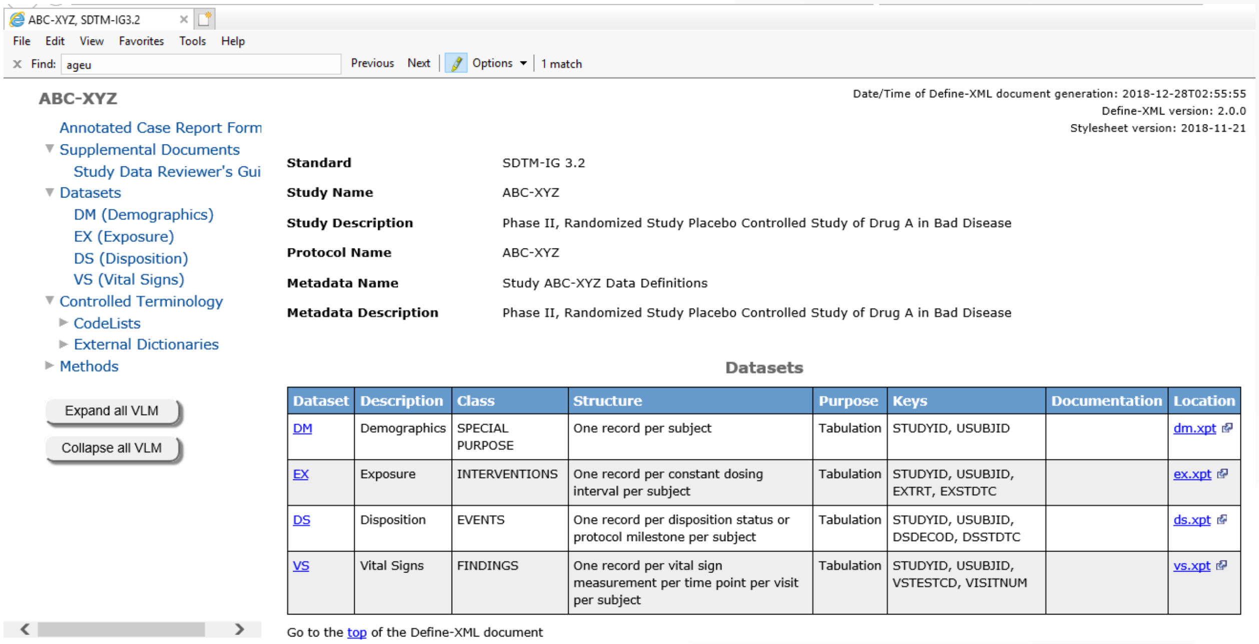
SDTM and ADaM define.xml files are overall very similar but have some differences to note in the section names though the content is very similar. Our focus will be SDTM define.xml but we will at times reference ADaM to highlight some differences. In the list below, the sections appear in both SDTM and ADaM define.xml unless noted specific to SDTM or ADaM.

Flowing from top to bottom, each of the define.xml sections describes the study data in increasing detail.

- Study information/header section – This is define.xml's high level introduction to us with study level information: study name, dataset standard type and version used for the datasets present.
- Tabulation Datasets (SDTM)/ Analysis Datasets ADaM) Lists – This is our quick introduction to which datasets define.xml will be introducing us to.
- Datasets section – This is essentially the variable level metadata section where define.xml will tell us a little more about its datasets and each variable present in them.
- Value Level Metadata (SDTM)/ Parameter Level Metadata (ADaM) – This section provides a deeper dive into some of the variables where derivations or source is not the same across all records.
- Controlled Terminology – This section is where we get to know the special language of this define.xml, the shorthand or nicknames that it uses in the data and structures.
- Computational Algorithms (SDTM)/Analysis Derivations (ADaM) and Comments sections - This section groups together and repeats the algorithms and comments that define.xml has shared in previous sections.
- External linked files – These external files are like meeting other friends of the define.xml that provide some more information about the history and creation details about the datasets.

Using the newer stylesheet, the names of the sections are made consistent between SDTM and ADaM. Tabulation Datasets/Analysis Datasets are simply "Datasets". Computational Algorithms/Analysis Derivations/Comments were replaced with "Methods". The Value Level Metadata section no longer appears as a stand-alone section which will be discussed in more detail in the Value Level Metadata section.

Figure 3 shows define.xml viewed with the newer stylesheet:



Date/Time of Define-XML document generation: 2018-12-28T02:55:55
Define-XML version: 2.0.0
Stylesheet version: 2018-11-21

Standard	SDTM-IG 3.2
Study Name	ABC-XYZ
Study Description	Phase II, Randomized Study Placebo Controlled Study of Drug A in Bad Disease
Protocol Name	ABC-XYZ
Metadata Name	Study ABC-XYZ Data Definitions
Metadata Description	Phase II, Randomized Study Placebo Controlled Study of Drug A in Bad Disease

Datasets

Dataset	Description	Class	Structure	Purpose	Keys	Documentation	Location
DM	Demographics	SPECIAL PURPOSE	One record per subject	Tabulation	STUDYID, USUBJID		dm.xpt
EX	Exposure	INTERVENTIONS	One record per constant dosing interval per subject	Tabulation	STUDYID, USUBJID, EXTRT, EXSTDTC		ex.xpt
DS	Disposition	EVENTS	One record per disposition status or protocol milestone per subject	Tabulation	STUDYID, USUBJID, DSDECOD, DSSTDTC		ds.xpt
VS	Vital Signs	FINDINGS	One record per vital sign measurement per time point per visit per subject	Tabulation	STUDYID, USUBJID, VSTESTCD, VISITNUM		vs.xpt

Go to the top of the Define-XML document

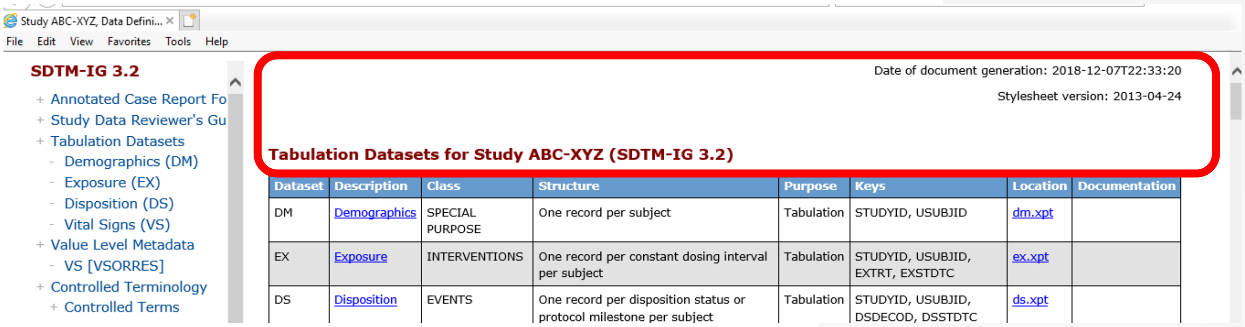
Figure 3. Define.xml Viewed with the Newer Stylesheet

STUDY INFORMATION/HEADER SECTION - HI, MY NAME IS STUDY ABC-XYZ.

First is a basic introduction. At the top of the define.xml output is the header section. In this section, basic introductory information about what will be described in this define.xml can be found. Either 'Tabulation Datasets' will be specified for an SDTM define.xml or 'Analysis Datasets' will be specified for an ADaM define.xml. Following that is the name of the study that the datasets have been created to describe and the version of SDTM Implementation Guide (IG) or ADaM IG that was used as a basis for creation of the datasets. Here, the study is 'ABC-XYZ' and SDTM IG version is 3.2. Since both SDTM and ADaM have multiple IG versions it is important to allow the reviewer to see at a glance which version of the dataset standard was used. When reviewing your study define.xml, do check to make sure that the version listed is accurate and that it is consistent with the version listed in the reviewer's guide.

Also listed in this section is the date and time of the define.xml creation as well as the version of the stylesheet used.

Figure 4 shows the header section of define.xml:



Study ABC-XYZ, Data Defini... ×
File Edit View Favorites Tools Help

SDTM-IG 3.2

- + Annotated Case Report Fo
- + Study Data Reviewer's Gu
- + Tabulation Datasets
  - - Demographics (DM)
  - - Exposure (EX)
  - - Disposition (DS)
  - - Vital Signs (VS)
- + Value Level Metadata
  - - VS [VSORRES]
- + Controlled Terminology
  - + Controlled Terms

Date of document generation: 2018-12-07T22:33:20

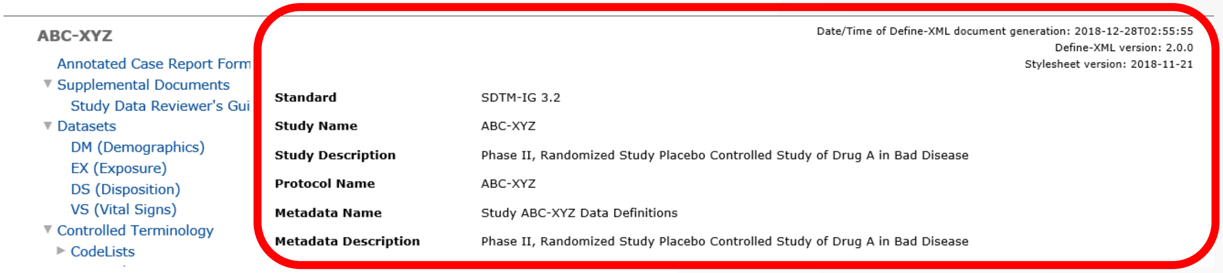
Stylesheet version: 2013-04-24

Tabulation Datasets for Study ABC-XYZ (SDTM-IG 3.2)

Dataset	Description	Class	Structure	Purpose	Keys	Location	Documentation
DM	Demographics	SPECIAL PURPOSE	One record per subject	Tabulation	STUDYID, USUBJID	dm.xpt	
EX	Exposure	INTERVENTIONS	One record per constant dosing interval per subject	Tabulation	STUDYID, USUBJID, EXTRT, EXSTDTC	ex.xpt	
DS	Disposition	EVENTS	One record per disposition status or protocol milestone per subject	Tabulation	STUDYID, USUBJID, DSDECOD, DSSTDTC	ds.xpt	

Figure 4. Header Section of Define.xml

Figure 5 shows the header section of define.xml using the newer stylesheet:



ABC-XYZ

- Annotated Case Report Form
- ▼ Supplemental Documents
  - Study Data Reviewer's Gui
- ▼ Datasets
  - DM (Demographics)
  - EX (Exposure)
  - DS (Disposition)
  - VS (Vital Signs)
- ▼ Controlled Terminology
  - ▶ CodeLists

Date/Time of Define-XML document generation: 2018-12-28T02:55:55
Define-XML version: 2.0.0
Stylesheet version: 2018-11-21

Standard	SDTM-IG 3.2
Study Name	ABC-XYZ
Study Description	Phase II, Randomized Study Placebo Controlled Study of Drug A in Bad Disease
Protocol Name	ABC-XYZ
Metadata Name	Study ABC-XYZ Data Definitions
Metadata Description	Phase II, Randomized Study Placebo Controlled Study of Drug A in Bad Disease

Figure 5. Header Section of Define.xml with Newer Stylesheet

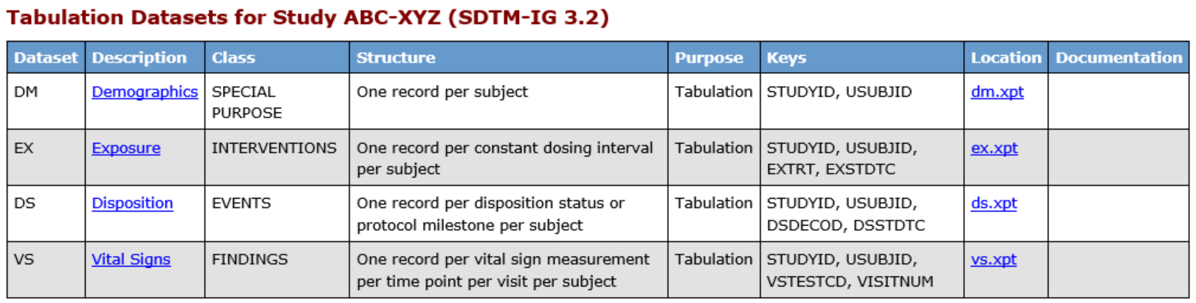
TABULATION DATASETS (ANALYSIS DATASETS) LIST THE BASICS – A QUICK INTRODUCTION TO THE DATASETS

Next comes the basics about which datasets have been included in the define.xml. In this section each submitted SDTM or ADaM dataset is listed in its own row. The dataset name is listed as well as a description of the dataset which will equate to the dataset label. Dataset class (SPECIAL PURPOSE, INTERVENTIONS, EVENTS, FINDINGS for SDTM; SUBJECT LEVEL ANALYSIS DATASET, BASIC DATA STRUCTURE, OCCURRENCE DATA STRUCTURE, ADAM OTHER for ADaM) and purpose ('Tabulation' for SDTM, 'Analysis' for ADaM) are specified.

Structure is a text description of the structure of the dataset in terms of the planned uniqueness of a record. For example, the dataset may be one record per subject or one record per constant dosing interval or one record per subject per parameter per visit. Keys are the variable names that are important to describing structure. Structure and keys should be consistent and describe fields that are contained in the dataset. If timepoint is important to the structure and uniqueness, then timepoint should be referenced in both structure and keys. Only variables present in the final dataset should be listed in keys.

A link to the location of the dataset in the submission is listed as well in the Location column. Check the case of the dataset file in the folder as well as in the link checking that the link works. The Documentation column may be used for dataset level information or reference to the reviewers guide as needed.

Figure 6 shows the tabulation datasets list:



Tabulation Datasets for Study ABC-XYZ (SDTM-IG 3.2)

Dataset	Description	Class	Structure	Purpose	Keys	Location	Documentation
DM	Demographics	SPECIAL PURPOSE	One record per subject	Tabulation	STUDYID, USUBJID	dm.xpt	
EX	Exposure	INTERVENTIONS	One record per constant dosing interval per subject	Tabulation	STUDYID, USUBJID, EXTRT, EXSTDTC	ex.xpt	
DS	Disposition	EVENTS	One record per disposition status or protocol milestone per subject	Tabulation	STUDYID, USUBJID, DSDECOD, DSSTDTC	ds.xpt	
VS	Vital Signs	FINDINGS	One record per vital sign measurement per time point per visit per subject	Tabulation	STUDYID, USUBJID, VSTESTCD, VISITNUM	vs.xpt	

Figure 6. Tabulation Datasets List

DATASETS SECTIONS (VARIABLE LEVEL METADATA) – TELL ME MORE… AN INITIAL DESCRIPTION OF EACH DATASET

The next layer is to find out a little bit more about the datasets that were defined at a high level in the Tabulation Datasets sections. Following the datasets list, a section is present for each of the datasets. This section contains a description of each variable contained in the dataset. There are links to each dataset section from both the left section under Tabulation Datasets as well as from the Description column for each dataset from the Tabulation Datasets cells. The columns present in this section that describe each variable in an SDTM are: Variable, Label, Key, Type, Length, Controlled Terms or Format, Origin, and Derivation/Comment. For an ADaM dataset, the columns to describe each variable are: Variable, Label, Type, Length/Display Format, Controlled Terms or Format, and Source/Derivation/Comment.

Variable, Label – The variable and label columns list the variable name and label exactly as they appear in the submitted dataset.

Key (SDTM only) – This column lists the order of the key variables. This should align with the variables and order defined in the Keys column on the datasets table at the top of the define.xml.

Type – This identifies variable type of integer, float, text or datetime. Numeric fields are identified as either integer or float. The type of datetime is used for type for character dates in the ISO8601 format.

Length (SDTM) or Length/Display Format (ADaM) – SDTM define.xml will list simply the length of the field in the dataset that will be submitted. If post processing was done on a dataset before submission to shorten the length to the shortest necessary length, this is that shortened length. It may not be equal to the initial planned length that appears in the programming specification. In addition to length, ADaM define.xml may list a display format in this column. An example of display format in ADaM would be date9 format applied to a numeric SAS date for display. SDTM variables do not have display formats, so this part of the field is only applicable to ADaM.

Controlled Terms or Format – This column lists a reference or link to the values that can be present in the variable described by the row when applicable. It would be used where there is a finite list of allowed values. When the list is short, this column lists both the name of the code list as well as the code and decode values.

Example: ["N" = "No", "Y" = "Yes"] <No Yes Response>

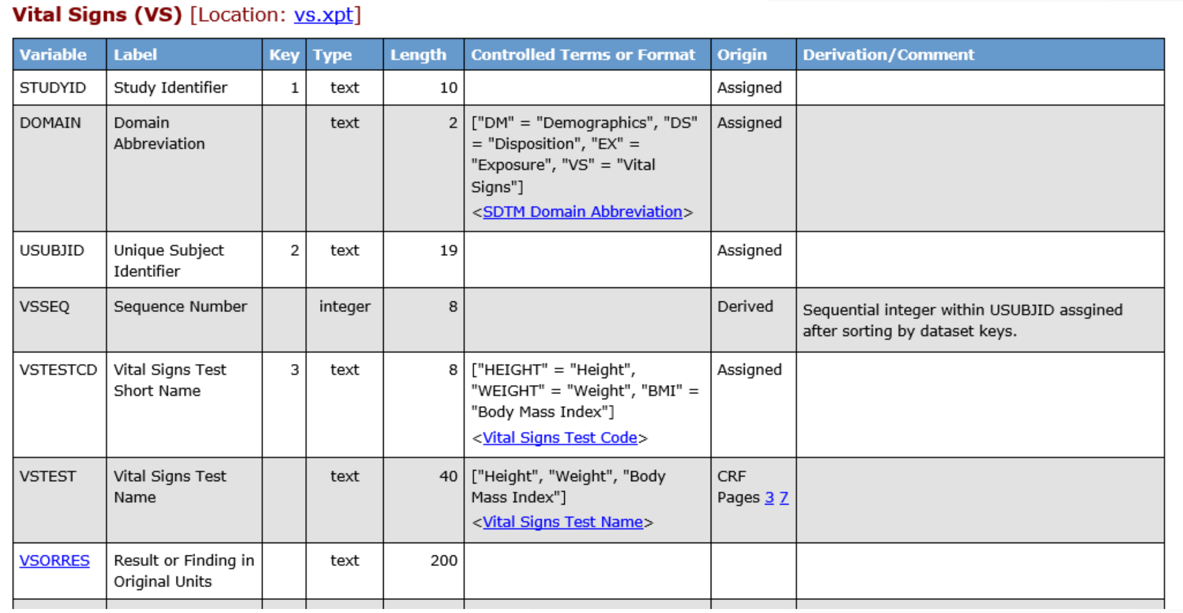
However, if the list is longer, just the name and link to the code list may be present. Each code list that is present should be defined in the Controlled Terminology section which will be described below.

Origin (SDTM only) – Defines the source of the value. CRF Page(s) x is used for values that come directly from raw variables captured from the CRF. In this case, the raw dataset/variable name is not referenced here, but rather the page number in the linked acrf.pdf on which the variable can be found. Derived is used when there is a calculation or derivation needed to be performed to obtain the variable value.

Assigned can be used when a constant value is assigned to a variable or a decode value is assigned based on a code value for example. eDT would be for electronically transferred data.

Derivation/Comment (SDTM) or Source/Derivation/Comment (ADaM) – This column provides the detail that the Origin column has noted is coming. If Origin is Derived, then there would be a derivation present in this column to describe the details. There may be a comment present for origins of Assigned. This cell may also contain an external reference and link to the Reviewer's Guide for example for derivations or explanations requiring more space or format than possible in the define.xml cell. In ADaM define.xml, this field also contains a Source component which is similar in function to Origin in describing the source from which the value has come. If the value is sourced directly from SDTM, it can use the Source of 'Predecessor' and might look like: Predecessor: DM.STUDYID. For those ADaM fields that are assigned or derived, the value of 'Assigned:' or 'Derived:' would appear in the cell with the derivation/comment text. This Derivation/Comment section is one that is important to review and pay close attention. While many of the other fields so far either come from the data or have a set list of values that can be present, this section is not as limited. The derivation/comment is free text and can be a manual effort, so it deserves a bit of extra attention in both creation and review. It is a description of what was done to get to the value. Keep various reviewers in mind and make sure that the derivation is clear and accurate to allow a reviewer to understand what was used as source and what methods or derivations were applied to that source to arrive at the value in the field. It is possible that someone looking at the define.xml may not have a programming background, so a textual description may be much more effective to get the concept across to a wider audience than code or pseudo code.

Figure 7 shows a sample of the VS domain section variable level metadata:



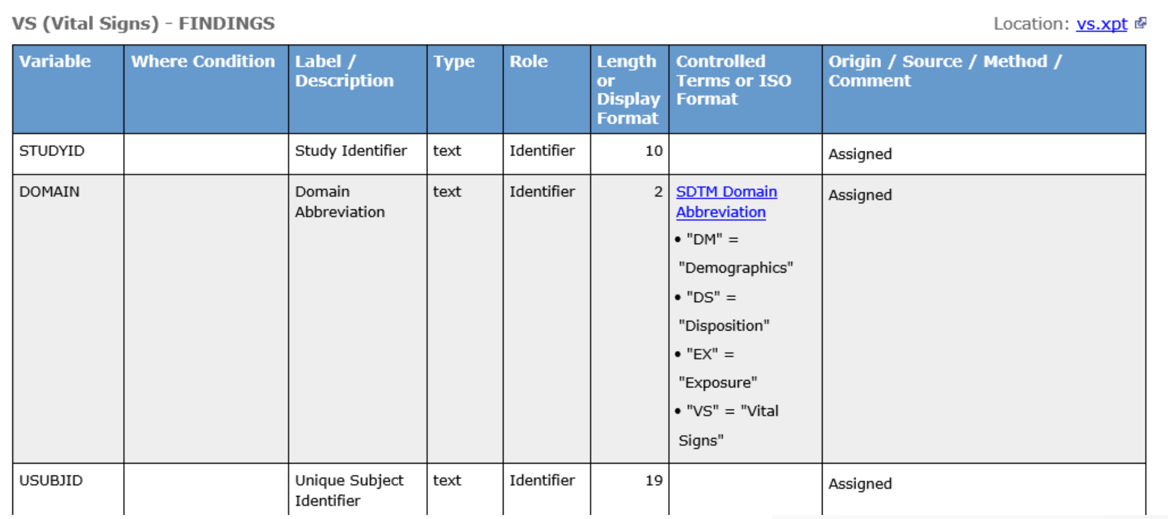
Vital Signs (VS) [Location: vs.xpt]

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin	Derivation/Comment
STUDYID	Study Identifier	1	text	10		Assigned	
DOMAIN	Domain Abbreviation		text	2	["DM" = "Demographics", "DS" = "Disposition", "EX" = "Exposure", "VS" = "Vital Signs"] <SDTM Domain Abbreviation>	Assigned	
USUBJID	Unique Subject Identifier	2	text	19		Assigned	
VSSEQ	Sequence Number		integer	8		Derived	Sequential integer within USUBJID assgined after sorting by dataset keys.
VSTESTCD	Vital Signs Test Short Name	3	text	8	["HEIGHT" = "Height", "WEIGHT" = "Weight", "BMI" = "Body Mass Index"] <Vital Signs Test Code>	Assigned	
VSTEST	Vital Signs Test Name		text	40	["Height", "Weight", "Body Mass Index"] <Vital Signs Test Name>	CRF Pages 3 7	
VSORRES	Result or Finding in Original Units		text	200			

Figure 7. Dataset Section for a Sample of Variable Level Metadata for the VS Domain

Using the newer stylesheet, the appearance of the dataset section differs with the removal of the Key column and the addition/combination of a few other columns. The Where Condition column has been added which will now be used for the Value Level Metadata instead of appearing in its own section. Also added in this view is a column for Role which displays the variable role as defined by the standards. Additionally, Origin and Derivation/Comment columns have been combined into one column for Origin/Source/Method/Comment.

Figure 8 shows a sample of the VS domain section variable level metadata viewed with the newer stylesheet:



VS (Vital Signs) - FINDINGS

Location: vs.xpt

Variable	Where Condition	Label / Description	Type	Role	Length or Display Format	Controlled Terms or ISO Format	Origin / Source / Method / Comment
STUDYID		Study Identifier	text	Identifier	10		Assigned
DOMAIN		Domain Abbreviation	text	Identifier	2	SDTM Domain Abbreviation • "DM" = "Demographics" • "DS" = "Disposition" • "EX" = "Exposure" • "VS" = "Vital Signs"	Assigned
USUBJID		Unique Subject Identifier	text	Identifier	19		Assigned

Figure 8. Dataset Section for a Sample of the VS Domain Viewed with Newer Stylesheet

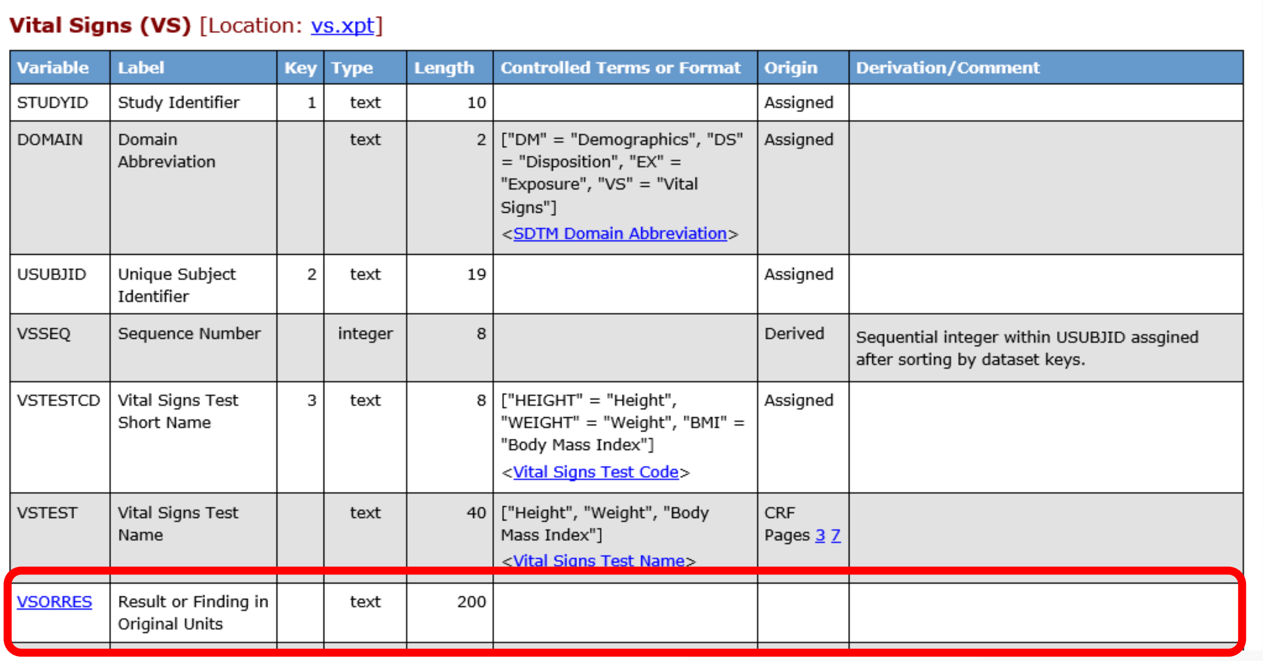
Within the Datasets Section with variable level descriptions, take care to make comments understandable to the reviewer or someone not already very familiar with the study. Remember this is meant to help the reviewer get to know the dataset. Also make sure to use references from within the submission such as the CRF page, not the raw data names. Use language explaining what has occurred in plain terms and not complex code.

Also, check that the same variables are referenced the same way across datasets (i.e. USUBJID across all datasets has the same attributes). Similarly, in ADaM check that core variables from ADSL that are present in other datasets are defined consistently across those datasets.

VALUE LEVEL METADATA – GETTING TO KNOW THE DETAILS

In the variable level description rows in the dataset section, certain variable names may be hyperlinked. In SDTM, this is commonly xxORRES for example, and in ADaM may be AVAL and possibly other variables. These highlighted/hyperlinked variable names will link to the Value Level Metadata section where the deeper dive into the data description and derivation continues. This section is used when derivations or source differs on records depending on another value, for example: xxTESTCD or PARAMCD value.

Figure 9 shows variable VSORRES with a link to value level metadata:



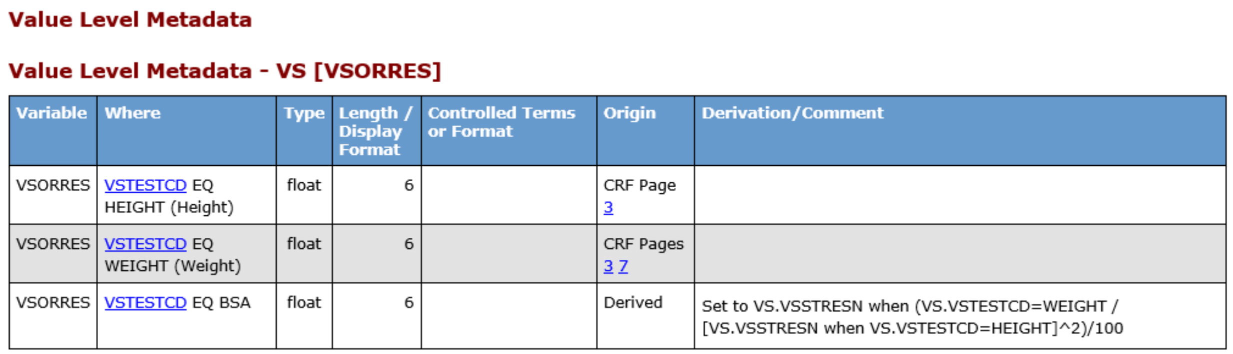
Vital Signs (VS) [Location: vs.xpt]

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin	Derivation/Comment
STUDYID	Study Identifier	1	text	10		Assigned	
DOMAIN	Domain Abbreviation		text	2	["DM" = "Demographics", "DS" = "Disposition", "EX" = "Exposure", "VS" = "Vital Signs"] <SDTM Domain Abbreviation>	Assigned	
USUBJID	Unique Subject Identifier	2	text	19		Assigned	
VSSEQ	Sequence Number		integer	8		Derived	Sequential integer within USUBJID assgined after sorting by dataset keys.
VSTESTCD	Vital Signs Test Short Name	3	text	8	["HEIGHT" = "Height", "WEIGHT" = "Weight", "BMI" = "Body Mass Index"] <Vital Signs Test Code>	Assigned	
VSTEST	Vital Signs Test Name		text	40	["Height", "Weight", "Body Mass Index"] <Vital Signs Test Name>	CRF Pages 3 7	
VSORRES	Result or Finding in Original Units		text	200			

Figure 9. VSORRES Variable with Link to Value Level Metadata

The columns present in the Value Level Metadata are very similar in label and content to those described in the Dataset variable level section. One difference in the value level section is that each table refers to just one variable. For that variable, for example VSORRES, there is then a row for each of the different origins or derivations as needed. The Where column defines the cases where the derivation/origin will differ. For example, in the VS domain, the original value VSORRES has 3 different sources defined with 3 different where clauses based on VSTESTCD value. Note that in the Dataset variable level section, Origin and Derivation/Comment are null since they differ at the variable level and will be defined separately at the value level. In the Value Level Metadata section, they are described on the table in 3 rows. Where VSTESTCD is HEIGHT (Height), the height value comes from CRF Page 3. Where VSTESTCD is WEIGHT (Weight), the weight value comes from both CRF Pages 3 and 7. Where VSTESTCD is BSA (Body Surface Area), the BSA is derived using the calculation provided and the standardized weight and height values.

Figure 10 shows value level metadata for VSORRES:



Value Level Metadata

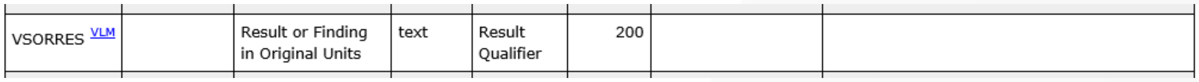
Value Level Metadata - VS [VSORRES]

Variable	Where	Type	Length / Display Format	Controlled Terms or Format	Origin	Derivation/Comment
VSORRES	VSTESTCD EQ HEIGHT (Height)	float	6		CRF Page 3	
VSORRES	VSTESTCD EQ WEIGHT (Weight)	float	6		CRF Pages 3 7	
VSORRES	VSTESTCD EQ BSA	float	6		Derived	Set to VS.VSSTRESN when (VS.VSTESTCD=WEIGHT / [VS.VSSTRESN when VS.VSTESTCD=HEIGHT]^2)/100

Figure 10. Value Level Metadata for VSORRES

The newer stylesheet displays the value level metadata within the dataset section itself. Instead of the VSORRES field linking to the VSORRES Value Level Metadata table, there is a hyperlink labeled VLM that will allow the dataset table itself to expand to display the value level metadata under the field using the VLM.

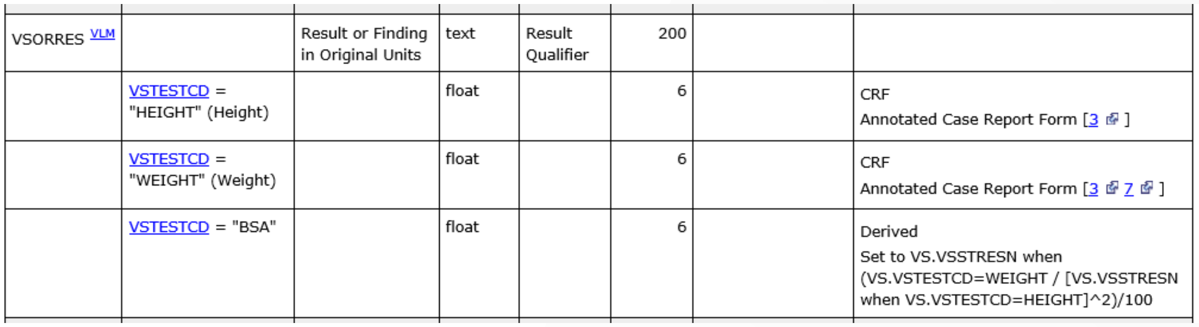
Figure 11 shows define.xml using newer stylesheet with VLM collapsed:



VSORRES VLM		Result or Finding in Original Units	text	Result Qualifier	200		

Figure 11. Collapsed Value Level Metadata for VSORRES with Newer Stylesheet

Figure 12 shows define.xml using newer stylesheet with VLM expanded:



VSORRES VLM		Result or Finding in Original Units	text	Result Qualifier	200		
	VSTESTCD = "HEIGHT" (Height)		float		6		CRF Annotated Case Report Form [3]
	VSTESTCD = "WEIGHT" (Weight)		float		6		CRF Annotated Case Report Form [3 7]
	VSTESTCD = "BSA"		float		6		Derived Set to VS.VSSTRESN when (VS.VSTESTCD=WEIGHT / [VS.VSSTRESN when VS.VSTESTCD=HEIGHT]^2)/100

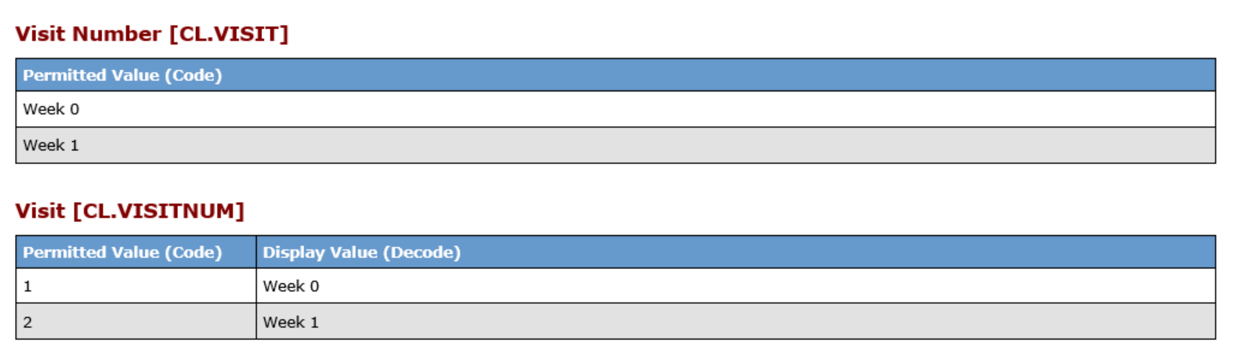
Figure 12. Expanded Value Level Metadata for VSORRES with Newer Stylesheet

CONTROLLED TERMINOLOGY – THE SHORTHAND OR NICKNAMES OF THIS DEFINE.XML AND DATASETS

The Controlled Terminology section is like a dictionary for the particular define.xml and set of datasets. It describes the shorthand or 'nicknames' used for this specific study and datasets. There are two sections in the Controlled Terminology section: Controlled Terms and External Dictionaries. The Controlled Terms section contains the lists of finite values for the controlled terms lists that were referenced in the Datasets and Value Level Metadata sections.

There are a couple different types of lists. For those code lists that are for values which are text only and do not have a decode necessary to make sense of the value and do not have a paired longer text field, only the value itself is listed in a single column. An example of this is VISIT. Values in VISIT make sense on their own and there is not a decode field paired with VISIT, so the text values of VISIT are listed under the Visit Number code list. Where decode text is necessary to clarify the meaning of the value or where there is a paired decode value, both the value and its decode are listed in separate columns. An example of this is VISITNUM. The value of 1 for VISITNUM does not clearly define what VISITNUM refers to. This could be VISIT 1, WEEK 1, or not have any correlation to the number 1. VISIT and VISITNUM are paired variables where VISIT contains the decode of the value in VISITNUM. For the VISITNUM code list, VISITNUM values are listed as the code value and the corresponding VISIT text value is listed as the decode to show that connection so that some context is given to the value VISITNUM=1.

Figure 13 shows examples of terminology lists for VISIT and VISITNUM:



Visit Number [CL.VISIT]

Permitted Value (Code)
Week 0
Week 1

Visit [CL.VISITNUM]

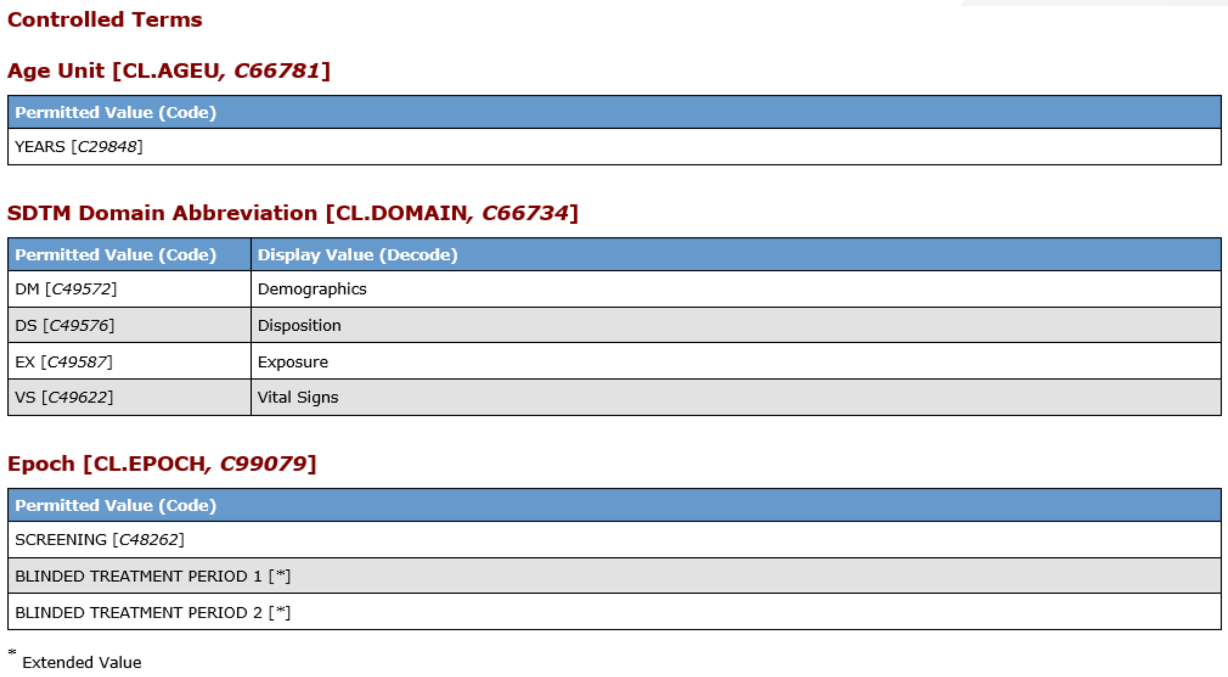
Permitted Value (Code)	Display Value (Decode)
1	Week 0
2	Week 1

Figure 13. VISIT and VISITNUM Terminology Lists

Controlled terminology lists could be from CDISC terminology or be user defined. CDISC sourced terminology lists will have the code number of the referenced terminology list following the code list name. User defined code lists will not have this number. Within extensible CDISC terminology lists, the rows present could reference values from the CDISC terminology list or extended values. The CDISC values will have the code number of the value itself appearing next to the code value. Where there is a value added to an extensible CDISC code list, this value will appear with a * in place of the code number and "* Extended value" will appear under the code list.

Note that the code lists will contain values applicable to the data in the study. That is, if the CDISC unit code list is referenced, not all values of unit will appear. Just selected values will be displayed. Also, there may be multiple code lists defined in the define.xml that reference the same CDISC code number. The unit code list is also an example of this. The unit code list values may be used in both LB and EX. However, the unit code list for EX would only reference those units that are appropriate to describe study drug dosing, not all units that were also used in LB.

Figure 14 shows terminology with CDISC code list codes, value codes, and extensible value notations:



Controlled Terms

Age Unit [CL.AGEU, *C66781*]

Permitted Value (Code)
YEARS [*C29848*]

SDTM Domain Abbreviation [CL.DOMAIN, *C66734*]

Permitted Value (Code)	Display Value (Decode)
DM [*C49572*]	Demographics
DS [*C49576*]	Disposition
EX [*C49587*]	Exposure
VS [*C49622*]	Vital Signs

Epoch [CL.EPOCH, *C99079*]

Permitted Value (Code)
SCREENING [*C48262*]
BLINDED TREATMENT PERIOD 1 [*]
BLINDED TREATMENT PERIOD 2 [*]

* Extended Value

Figure 14. Terminology with CDISC Code Lists, CDISC Code Values and Extensible Values

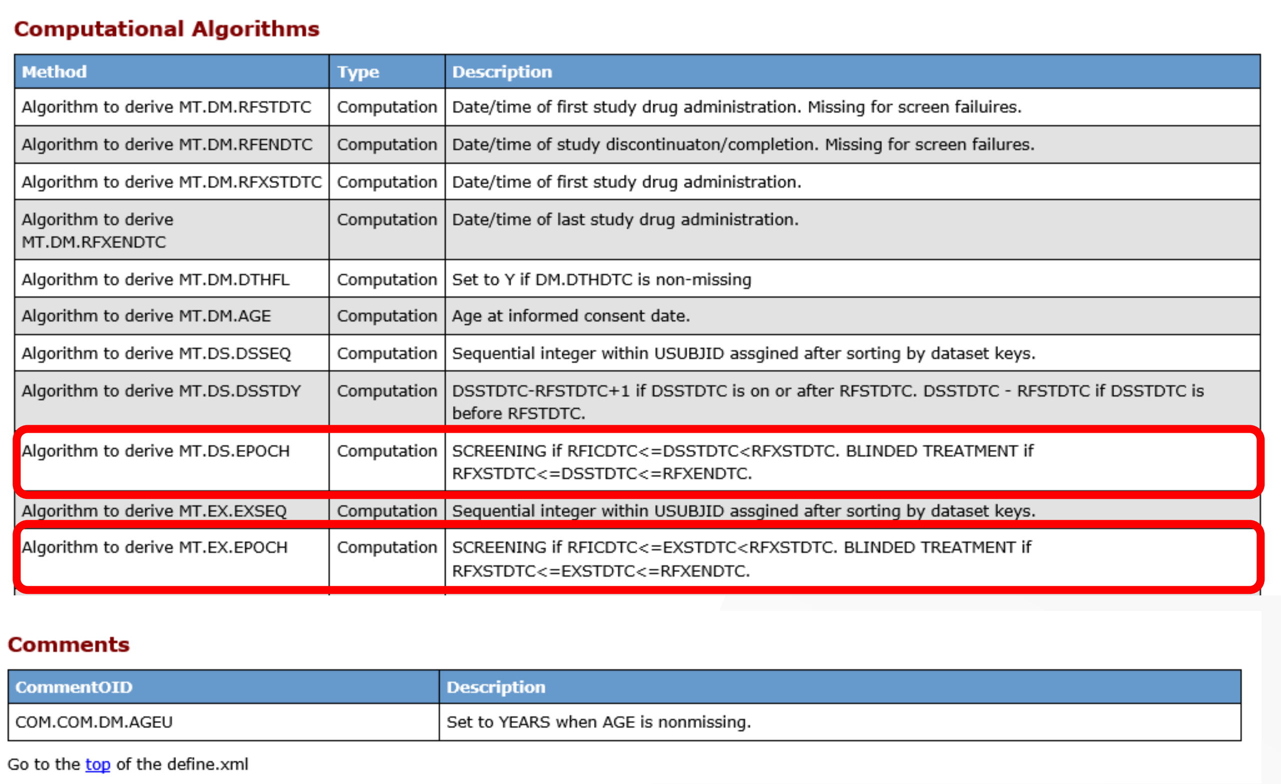
The External Dictionaries portion of the Controlled Terminology section contains references to the external dictionaries used but does not show a list of values. The name used for reference, dictionary name and dictionary version are listed. Examples of external dictionaries that may be referenced are ISO8601, ISO3166, MedDRA or WHODrug.

COMPUTATIONAL ALGORITHMS/COMMENTS SECTIONS

Now that we know the study's define.xml pretty well, like talking with an old friend, sometimes the stories repeat. The lore or epic stories that are worth repeating are rehashed.

The Computational Algorithms and Comments sections group together and repeat each of the algorithms and comments that have been stated in previous sections. While these sections may seem a bit repetitive, they can actually be great tools to aid in review. Here, seeing all of the derivations and comments consolidated, it is helpful to review for consistencies. EPOCH, xxDY, xxSEQ or other variables that would have similar definitions across datasets can be cross checked for wording and definition. For example, instead of looking across each domain to see how EPOCH is worded, it is quite easy to page through the Computational Algorithms section to see the various EPOCH descriptions without quite as many variables or page space in between them. In addition to checking that the derivations were worded similarly, it is also helpful to look in between to make sure that there are differences where they should be. For example, that EPOCH for DS references DSSTDTC and that EPOCH for EX has been changed to reference EXSTDTC.

Figure 15 shows Computational Algorithms and Comments sections:



Computational Algorithms

Method	Type	Description
Algorithm to derive MT.DM.RFSTDTC	Computation	Date/time of first study drug administration. Missing for screen failuires.
Algorithm to derive MT.DM.RFENDTC	Computation	Date/time of study discontinuaton/completion. Missing for screen failures.
Algorithm to derive MT.DM.RFXSTDTC	Computation	Date/time of first study drug administration.
Algorithm to derive MT.DM.RFXENDTC	Computation	Date/time of last study drug administration.
Algorithm to derive MT.DM.DTHFL	Computation	Set to Y if DM.DTHDTC is non-missing
Algorithm to derive MT.DM.AGE	Computation	Age at informed consent date.
Algorithm to derive MT.DS.DSSEQ	Computation	Sequential integer within USUBJID assgined after sorting by dataset keys.
Algorithm to derive MT.DS.DSSTDY	Computation	DSSTDTC-RFSTDTC+1 if DSSTDTC is on or after RFSTDTC. DSSTDTC - RFSTDTC if DSSTDTC is before RFSTDTC.
Algorithm to derive MT.DS.EPOCH	Computation	SCREENING if RFICDTC<=DSSTDTC<RFXSTDTC. BLINDED TREATMENT if RFXSTDTC<=DSSTDTC<=RFXENDTC.
Algorithm to derive MT.EX.EXSEQ	Computation	Sequential integer within USUBJID assgined after sorting by dataset keys.
Algorithm to derive MT.EX.EPOCH	Computation	SCREENING if RFICDTC<=EXSTDTC<RFXSTDTC. BLINDED TREATMENT if RFXSTDTC<=EXSTDTC<=RFXENDTC.

Comments

CommentOID	Description
COM.COM.DM.AGEU	Set to YEARS when AGE is nonmissing.

Go to the top of the define.xml

Figure 15. Computational Algorithms and Comments Sections

With the newer stylesheet, this section is labeled as simply Methods and does not contain the comments from the fields with origin of Assigned. However, the columns present and content remain the same.

EXTERNAL LINKED FILES – MEETING SOME OTHER FRIENDS

In addition to the define.xml itself, there are other files external to the define.xml that help in this description and familiarization of the study data. These external files are like meeting other friends of the define.xml that provide more information about the history and creation details about the datasets. Each of these files is a topic unto itself, so for purposes here, the input to the define.xml story will be kept to a high level. Some of the other files that will be briefly described are:

acrf.pdf – Annotated Case Report Form (referenced by SDTM define.xml)

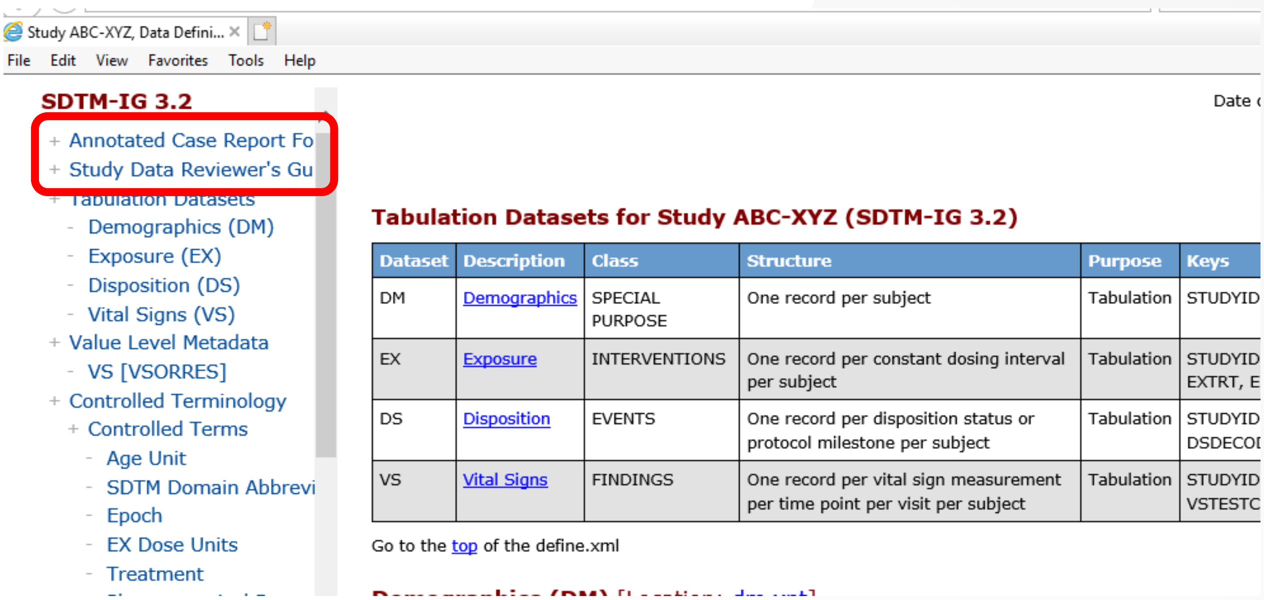
csdrg.pdf – Clinical Study Data Reviewer's Guide (referenced by SDTM define.xml)

adrg.pdf – Analysis Data Reviewer's Guide (referenced by ADaM define.xml)

The aCRF gives a picture of some of the data collection sources so that we can see a visual of where the data point has been collected. In addition to the link on the left of the define.xml to the aCRF, there are also direct page links to the CRF. As described in the Origin column, when the source of the data is the CRF, text will appear that describes the origin of "CRF Page xx". Each of these page numbers will also link to the acrf.pdf file to show where that particular data point has come from. The acrf.pdf file will have annotations of the SDTM domain and variable names that will correspond to those found in the define.xml file.

The csdrg.pdf and adrg.pdf files have further descriptions about the study and data. Where the descriptions of derivations in the define.xml are somewhat limited in space, the SDRG and ADRG files can use additional space for complex algorithms and/or pictures, flow charts or descriptions in tabular form where needed. Both of these files tell some of the background story that is important for a reviewer to know about the dataset structure and content but define.xml may not have been able to or is not designed to share directly. They contain descriptions of the protocol, historical information, decisions leading up to the dataset algorithms, and even problems with data. They will also list any data that was planned but not submitted. Earlier, the IE domain was given as an example of a domain that may not be submitted if there were not any subjects with inclusion/exclusion violations. But the planned, but not submitted IE domain could be described in the csdrg.pdf. Additionally, the reviewer's guides contain a list of all datasets giving us the ability to describe custom domains as well as any dataset dependencies. The Reviewer's Guides will also show how the datasets fare in terms of compliance and describe reasoning for any messages that remain after final compliance checking. These issues may not be apparent looking at dataset/metadata from what the define.xml is able to show but will give this information so that a reviewer will have the background information.

Figure 16 shows links to the Study Data Reviewers Guide and Annotated CRF:



SDTM-IG 3.2

+ Annotated Case Report Fo
+ Study Data Reviewer's Gu
+ Tabulation Datasets
- Demographics (DM)
- Exposure (EX)
- Disposition (DS)
- Vital Signs (VS)
+ Value Level Metadata
- VS [VSORRES]
+ Controlled Terminology
+ Controlled Terms
- Age Unit
- SDTM Domain Abbrevi
- Epoch
- EX Dose Units
- Treatment

Tabulation Datasets for Study ABC-XYZ (SDTM-IG 3.2)

Dataset	Description	Class	Structure	Purpose	Keys
DM	Demographics	SPECIAL PURPOSE	One record per subject	Tabulation	STUDYID
EX	Exposure	INTERVENTIONS	One record per constant dosing interval per subject	Tabulation	STUDYID EXTRT, E
DS	Disposition	EVENTS	One record per disposition status or protocol milestone per subject	Tabulation	STUDYID DSDECOI
VS	Vital Signs	FINDINGS	One record per vital sign measurement per time point per visit per subject	Tabulation	STUDYID VSTESTC

Go to the top of the define.xml

Figure 16. Links to Study Data Reviewer's Guide and Annotated CRF

WHAT'S NEXT? DEFINE.XML 2.1 - SOME OF THE PROPOSED CHANGES

Two draft versions of the next define.xml standard have been released for public review, but define.xml 2.1 has not yet been released as final. Some of the proposed changes that may be coming in define.xml version 2.1 are:

- The ability for the controlled terminology package to be specified as one of the standards in the define.xml metadata. Multiple content standards and terminology standards can be specified, but one of each must be designated as the default.
- The ability to define and include non-standard domains or variables in define.xml. Though these non-standard pieces may be possible in the define.xml, this does not mean that they are an acceptable part of the data standards, so do check compliance to the standard version being used.
- Origin may be updated in two ways. First enhanced origin can be used to describe both the type (collected, derived, assigned, protocol, predecessor) and corresponding source (example, subject,

investigator, DATASET.VARIABLE). Additionally, 2.1 may allow for multiple origin elements to be defined, but they need to be described and should be used only when no other option. The existing method is to use value level metadata when a clearly defined where statement will apply.

- Description elements added to def:ValueListDef and CodeList and comments can be added to MetaDataVersion and CodeList.
- The ability for links to external documents to be made more specific to a particular section rather than referencing only the document itself.
- Addition of SubClass to be utilized for ADaM dataset classes.
- Addition of HasNoData which can be used in cases where an item has been defined for the study but is not present in the data.

CONCLUSION

While define.xml may look intimidating at first, after getting to know our new friend section by section it is a very effective tool to help to understand the study data. Those that generate define.xml, those that review it and the end user reviewers or internal team members benefit from a define.xml that communicates the derivations and descriptions clearly and is well put together. By reviewing each section of the define.xml and now understanding what each piece is describing, the define.xml becomes a bit less of an unknown. After becoming familiar with define.xml we can now see its friendly and helpful way of communicating information about SDTM and ADAM. Creating and reviewing content in order to provide the best possible define.xml to your reviewer should now be a less intimidating and very achievable task.

REFERENCES

CDISC Define-XML Specification Version 2.0 www.cdisc.org

Define-XML 2.0 Release Package www.cdisc.org

CDISC Define-XML Specification Version 2.1 (DRAFT and DRAFT2) www.cdisc.org

Changes in Define-XMLV2-1.ppt www.cdisc.org

CONTACT INFORMATION

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

Christine McNichol
Covance, Inc
Christine.McNichol@Covance.com

Any brand and product names are trademarks of their respective companies.