

A Data-Driven Approach for Generating Define.xml v2.0 using ADaM Specifications and ADaM Datasets

Jaskaran Sing Gill, Hima Bhatia, ICON Plc.

ABSTRACT

Case Report Tabulation Data Definition Specifications (CRT-DDS) commonly known as Define.xml is an integral part of FDA's electronic submissions. Define.xml facilitates the review process by providing information about the structure and content of standard clinical data (SDTM and ADaM) in a machine readable format. There are two versions of define.xml v1.0 and v2.0 available. In March 2016, FDA announced to end support for Version 1.0 of define.xml and mandate the use of version 2.0 for all the electronic submissions. This paper presents a data-driven approach for generating define.xml v2.0 using ADaM ADS specifications and ADaM datasets. A Define ready ADaM ADS specification plus ADaM datasets are passed into a macro that automatically generates define.xml v2.0. This macro also provides an error report for any inconsistencies identified between ADaM datasets and ADaM specifications. This automation process considerably reduces the time and effort in generating a CDISC compliant define.xml for FDA submissions.

INTRODUCTION

Sponsor generate ADaM ADS specifications which serves as a programming guidance document for generating ADaM datasets. These ADaM ADS specifications are also submitted to FDA in the Case Report Tabulation Data Definition Specifications format or Define.xml prepared by CDISC define.xml team. Define.xml provide the structure and content of ADaM datasets in a machine readable format.

The process of generating a define.xml can be very time consuming if done manually. This paper depicts a process of generating define.xml v2.0 using ADaM ADS specifications and ADaM datasets with least manual input.

A Define.xml provides information about the following metadata's:

- Dataset Metadata:
 - o Name, Description, Class, Structure, Purpose, Key Variables, Location, Documentation
- Variable Metadata:
 - o Name, Label, Type, Length/Display Format, Controlled Terminology, Source/Derivation/Comment
- Parameter Value Level Metadata
 - o Describes method used to derive a particular variable for different PARAMCDs in an BDS ADaM dataset
- Controlled Terminology
- Analysis Derivation
- Comments
- External Documentation (ADRG, SAS codes)

PRE-REQUIREMENTS

Find enclosed the list of the items that are required to generate define.xml using this approach:

- Define Stylesheet(define2-0-0.xls)
- ADaM Datasets

- o Trimmed ADaM Datasets where character variable lengths are adjusted to the maximum length of the variable used across all the datasets in the study.
 - o SAS Transport file using Trimmed ADaM Datasets
- ADaM ADS specifications
- External Documentation
 - o ADRG
 - o SAS Codes
 - o Other external documentation that is being referred in ADaM ADS specifications

ADS SPECIFICATIONS

A standard ADaM ADS specification template is created to generate define.xml. A separate worksheet is created for each metadata that is being presented in define.xml.

Study Information

In this section, we populate Protocol Number, Study Description, Protocol Name, ADaM IG version and language used for writing ADaM specifications and generating ADaM datasets.

Parameter	Description
StudyName	111-P-222
StudyDescription	Define for Study XYZ
ProtocolName	Define for Study XYZ
StandardName	ADaM-IG
StandardVersion	1.0
Language	En

Figure 1. Study Information

Dataset Metadata

Dataset metadata serves as a TOC that lists out all the ADaM datasets. The following information is collected in dataset metadata:

- Dataset Name
- Dataset Description or Dataset label
- Dataset Structure
- Dataset Purpose
 - o Analysis for ADaM Datasets
- Dataset Class (ADSL, BDS, OCCD and ADaM OTHER)
- Key Variables
- Repeating
 - o Yes: For ADS that have more than one record for each USUBJID

- o No: For ADS that have one record for each USUBJID
- Documentation
- Document
 - o A unique short name given to each external document. Each short name has to start with a prefix "LF."

Dataset	Description	Structure	Purpose	Class
ADSL	Subject Level Analysis	One record per screened subject	Analysis	SUBJECT LEVEL ANALYSIS DATASET
ADVS	Vital Signs	One record per subject per parameter per visit	Analysis	BASIC DATA STRUCTURE
Key Variables		Repeating	Documentation	Document
USUBJID		No	SAP, Protocol	LF.ADRG; LF.ADSL.PGM;
USUBJID,PARAMCD,AVISITN		Yes	SAP, Protocol	LF.ADVS.PGM;

Figure 2. Dataset Metadata

Variable Metadata

All the variables that are included in each ADaM dataset are described in variable metadata.

- Dataset Name
- Variable Name
- Variable Label
- Variable Type as per data types give in CDISC ODM
 - o text, integer, float, date, time, datetime, partialDate, partialTime, partialDatetime, incompleteDatetime
- Variable length
- Sigdigit: Significant digits. Only populated where variable type is float
- Key: Integer value assigned to all the key variables
- Format: date and time formats associated with date/time numeric variables
- Codelist: A unique name assigned to standard or sponsor defined valid values and decodes
- Origin:
 - o Derived, Assigned and Predecessor
- Source/Derivation: Describes the derivation of a variable. If the derivation is more than 1000 characters, the derivation will be saved in an external document. A short name for this document will be listed in Document column
- Valuelist
 - o Populated only for BDS type of ADaM datasets

- o Populated with a value of “Y” for those variables whose derivation changes for each parameter
- o Value Level Metadata will be populated for the variable for whom valuelist is marked as “Y”
- Mandatory
 - o Populated with values of Yes/No depending upon whether a variable is required or not.
- Document
 - o A unique short name given to each external document. Each short name has to start with a prefix “LF.”

Dataset	Variable	Label	Type	Length	Sigdigit	Key	Format
ADSL	STUDYID	Study Identifier	text	15			
ADSL	SITEID	Study Site Identifier	text	25			
ADSL	INVID	Investigator Identifier	text	25			
ADSL	INVNAM	Investigator Name	text	200			
ADSL	USUBJID	Unique Subject Identifier	text	25		1	
ADSL	SUBJID	Subject Identifier for the Study	text	25			
CodeList	Origin	Source	ValueList		Mandatory		Document
	Predecessor	DM.STUDYID			Yes		
	Predecessor	DM.SITEID			Yes		
	Predecessor	DM.INVID			Yes		
	Predecessor	DM.INVNAM			No		
	Predecessor	DM.USUBJID			Yes		
	Predecessor	DM.SUBJID			No		

Figure 3. Variable Metadata

Value Level Metadata

Value level metadata will be populated for all the variables for whom ValueList column is marked as “Y” in the variable metadata.

- Dataset Name
- Variable attributes for which value level metadata is generated
 - o Variable Name
 - o Variable Label
 - o Variable Type
 - o Variable Length
 - o Sigdigit: Significant digits. Only populated when variable type is float
 - o Origin: Derived/Assigned
- PARAMCD: List of PARAMCD that are being generated in a particular ADaM dataset
- Source: Describes the derivation of a variable for each PARAMCD. If the derivation is more than 1000 characters, the derivation will be saved in an external document. A short name for this document will be listed in Document column

- Document
 - o A unique short name given to each external document. Each short name has to start with a prefix "LF."

Dataset	Variable	Label	Type	Length	Sigdigit
ADVS	PARAM	Parameter	text	200	
ADVS	PARAM	Parameter	text	200	
ADVS	PARAM	Parameter	text	200	
ADVS	PARAM	Parameter	text	200	
ADVS	PARAM	Parameter	text	200	
ADVS	PARAM	Parameter	text	200	
ADVS	PARAM	Parameter	text	200	
Origin	PARAMCD	Source			Document
Derived	VSALL	PARAM= trim(left(VS.VSTEST))			
Derived	BMI	PARAM= trim(left(VS.VSTEST))			
Derived	DIABP	PARAM= trim(left(VS.VSTEST)) " (" trim(left(VS.VSSTRESU)) ")"			
Derived	HEIGHT	PARAM= trim(left(VS.VSTEST)) " (" trim(left(VS.VSSTRESU)) ")"			
Derived	PULSE	PARAM= trim(left(VS.VSTEST)) " (" trim(left(VS.VSSTRESU)) ")"			
Derived	SYSBP	PARAM= trim(left(VS.VSTEST)) " (" trim(left(VS.VSSTRESU)) ")"			
Derived	WEIGHT	PARAM= trim(left(VS.VSTEST)) " (" trim(left(VS.VSSTRESU)) ")"			

Figure 4. Value Level Metadata

Controlled Terminology

Controlled terminology collects standard or sponsor defined valid values and decodes.

- Codelist Name: A unique short name assigned to a codelist
- Codelist Label: A short description of codelist
- Codelist Type: text/integer/float
- Code: Standard or sponsor defined valid values
- Decode: Coded value for each standard or sponsor defined value
- Rank: Populated only for codelists that have a numeric significance. Rank should not be used to define a display order
- OrderNumber: OrderNumber is used to define a display order for the items in a particular codelist
- CodeListNum: C-codes provided in National Cancer Institute's Enterprise Vocabulary System if unique short name of a codelist matches with standard codelist names provided in National Cancer Institute's Enterprise Vocabulary System.
- CodeNum: C-codes provided in National Cancer Institute's Enterprise Vocabulary System if any of standard or sponsor defined valid values matches with standard names provided in National Cancer Institute's Enterprise Vocabulary System.

Codelist Name	CodeList Label	CodeList Type	Code	Decode	Rank	OrderNumber	CodeListNum	Codenum
AGEU	Age Unit	text	YEARS			1	C66781	C29848
SEX	Sex	text	M			1	C66731	C20197
SEX	Sex	text	F			2	C66731	C16576
AVISITN	AVISITN	integer	1	Screening	1	1		
AVISITN	AVISITN	integer	2	Day -2 to -1	2	2		
AVISITN	AVISITN	integer	3	Day 1 to 3	3	3		
AVISITN	AVISITN	integer	4	Day 8	4	4		
RACE	Race	text	WHITE			5	C74457	C41261

Figure 5. Controlled Terminology

Documents

This section collects the following information about each external documentation that is being referred in ADaM ADS specifications:

- Document ID: A unique short name with a prefix of "LF.", given to each external document.
- Title: Text that will be displayed in define.xml
- HRef: Physical location of the file with the actual file name and file extension. If define.xml and external documentation is located in the same folder, populate actual file name and file extension only.

Document ID	Title	HRef
LF.ADRG	Analysis Data Reviewer's Guide	analysis-data-reviewers-guide.pdf
LF.ADSL.PGM	adsl.sas	d0_adsl.sas
LF.ADVS.PGM	adv.sas	d1_adv.sas

Figure 6. Documents

PROCESS FLOW OF GENREATING DEFINE.XML

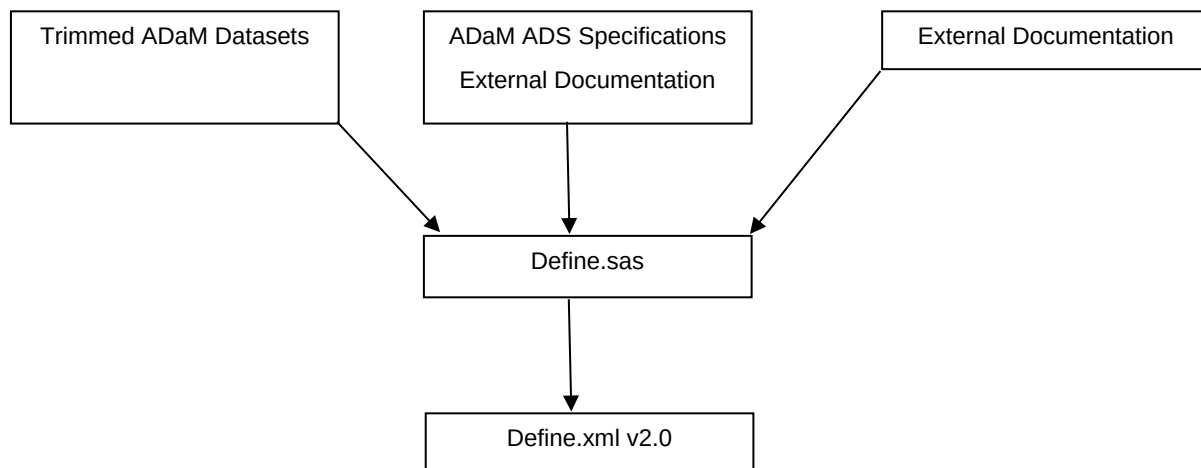


Figure 7. Process Flow of Generating Define.xml

The location of trimmed ADaM datasets and final ADaM ADS specifications are passed into define.sas macro which generates define.xml v2.0. Define.sas macro refers to trimmed ADaM datasets for variable attributes (Name, label, type, length), order of the variables in a particular ADaM dataset and controlled terminology for derived variables.

ADaM-IG 1.0

Analysis Data Reviewers Guide

- Analysis Datasets
- Parameter Value Level Metadata
- Controlled Terminology
- Analysis Derivations
- Comments

Analysis Datasets for Study 111-P-222 (ADaM-IG 1.0)

Dataset	Description	Class	Structure	Purpose	Keys	Location	Documentation
ADSL	Subject Level Analysis	SUBJECT LEVEL ANALYSIS DATASET	One record per screened subject	Analysis	USUBJID	adsl.xpt	SAP, Protocol d0_adsl.sas
ADVS	Vital Signs	BASIC DATA STRUCTURE	One record per subject per parameter per visit	Analysis	USUBJID, PARAMCD, AVISITN	adv.spt	SAP, Protocol d1_adv.sas

Go to the [top](#) of the define.xml

Subject Level Analysis (ADSL) [Location: [adsl.xpt](#)]

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
STUDYID	Study Identifier	text	8		Predecessor: DM.STUDYID
SITEID	Study Site Identifier	text	2		Predecessor: DM.SITEID
INVID	Investigator Identifier	text	2		Predecessor: DM.INVID
INVNAM	Investigator Name	text	23		Predecessor: DM.INVNAM
USUBJID	Unique Subject Identifier	text	15		Predecessor: DM.USUBJID
SUBJID	Subject Identifier for the Study	text	6		Predecessor: DM.SUBJID

Figure 8. Define.xml v2.0 viewed via Microsoft Internet Explorer

CONCLUSION

This data driven approach of generating define.xml using ADaM datasets and ADaM ADS specifications is a time-saving, straightforward and easy way of generating define.xml v2.0. Programmers doesn't need to have XML knowledge for generating define.xml via this method.

REFERENCES

CDISC Define-XML Specification, Version 2.0, March 5, 2013 <http://www.cdisc.org/define-xml>

Jansen, Lex. 2017. "Creating Define-XML version 2 including Analysis Results Metadata with the SAS® Clinical Standards Toolkit. Pharma SUG 2017.

<https://www.pharmasug.org/proceedings/2017/SS/PharmaSUG-2017-SS08.pdf>

RECOMMENDED READING

- CDISC Define-XML Specification, Version 2.0
- Study Data Technical Conformance Guide, March 2016

CONTACT INFORMATION

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

Jaskaran Singh Gill
ICON Plc
Jaskaran.Gill@iconplc.com
www.iconplc.com

Hima Bhaita
ICON Plc
Hima.Bhatia@iconplc.com
www.iconplc.com