

Generating Define.xml v2.0 and Analysis Result Metadata using ADaM Specifications, ADaM Datasets and TFL Annotation

Mark Crangle, ICON Clinical Research, Marlow, UK
Hima Bhatia, ICON Clinical Research, North Wales, U.S.A

ABSTRACT

Case Report Tabulation Data Definition Specifications (CRT-DDS) commonly known as Define-xml is an integral part of FDA's electronic submissions. The most recent addition to define.xml package is Analysis Result Metadata. 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 whereas Analysis Result Metadata provides information about the purpose of the analysis, ADaM datasets and analysis methods used for analysis. This paper presents a data-driven approach for generating define.xml v2.0 where a Define ready ADaM specification plus ADaM datasets are passed into a macro that automatically generates define.xml v2.0. For Analysis Result metadata, a standard TFL annotation spreadsheet is also passed into a macro that automatically generates analysis result metadata, providing information about the relationship between analysis results, analysis datasets and documentation of the analysis.

INTRODUCTION

Sponsor generate ADaM specifications which serves as a programming guidance document for generating ADaM datasets. These ADaM 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.

Analysis Results Metadata(ARM.XML) provides traceability for a given analysis result to the specific ADaM data used as input. It also provides standard format for identification of key analyses, information about analysis method used and reason for analysis which can assist regulatory reviewers by linking results, documentation and datasets.

The process of generating a define.xml and ARM.xml can be very time consuming if done manually. This paper depicts a process of generating the documents using ADaM specifications, ADaM datasets and ARM.xml 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 macro:

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

PhUSE US Connect 2018

- 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

ADaM Specifications

A standard ADaM 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 analysis 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 analysis 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 "LR."

PhUSE US Connect 2018

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 analysis 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 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
- 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
- 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	CodeList
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				

PhUSE US Connect 2018

ADSL	USUBJID	Unique Subject Identifier	text	25		1		
ADSL	SUBJID	Subject Identifier for the Study	text	25				
Origin	Source	ValueList	Mandatory	FormalExpression	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 will be created
 - 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 analysis dataset
- Source: Describes the derivation of a variable for each PARAMCD. If the derivation is more than 1000 characters, this 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)) ")"			

PhUSE US Connect 2018

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: Populate with 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: Populate with 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.

PhUSE US Connect 2018

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	advs.sas	d1_advs.sas

Figure 6. Documents

ARM.XML

For generating ARM.xml, a standard template will be populated TLF mocks annotation (similar to ADaM specs template) with below tabs.

- Analysis
- WhereClause
- Documents

Analysis

- Table Number: Table number defined in mock
- Table Title: Table title defined in mock
- CSR Ref: Any referene
- Analysis: Type of analysis need to be performed
- Merge: Any merge variable
- Reason: Reference documents e.g SAP/Protocol
- Purpose: Primary/Secondary Outcome
- Dataset: Dataset name
- Analysis Variable: Analysis variables used in the analysis
- PARAMCD: Parameter Code
- Description: Description to support the analysis
- SAP: Reference to SAP section
- Code: Statistical procedure

	A	B	C	D	E	F	G	H	I	J	K	L	M
	Table Number	Table Title	CSR Ref	Analysis	Merge	Reason	Purpose	Dataset	Analysis Variable	PARAMCD	Description	SAP	Code
1	Table 14.2.1.1.2.1	Statistical Analysis (Repeated Measures Model) of Change from Baseline in Mean Daily Average Index Knee Pain While Walking (24h recall) by Week up to Week 4 - Primary Analysis	LF Table 14.2.1.1.2.1.4	Difference between treatment by week		SPECIFIED IN SAP	PRIMARY OUTCOME MEASURE	ADSL	CHG	Y	Adjusted least square mean	LF.SAP-SEC-10.4.1	PROC MIXED DATA=USL; RUN; CLASS WEEK TRT USUBJID COUNTRY; MODEL CHG_BASE - TRT WEEK BASE_VAL TRT WEEK COUNTRY / DIFF=KP; REPEATED WEEK TYPE=UN SUBJECT=USUBJID; ESTIMATE 'ISS1 100mg vs Placebo' TRT=110 / CL ALPHA=0.05; ESTIMATE 'ISS2 200mg vs Placebo' TRT=110 / CL ALPHA=0.05; ESTIMATE 'ISS3 100mg vs ISS1 100mg' TRT=0-11 / CL ALPHA=0.05; ODS OUTPUT LMEANS=lm.DIFFS=diff.ESTIMATES=est; RUN;
2	Table 14.2.1.1.2.1	Statistical Analysis (Repeated Measures Model) of Change from Baseline in Mean Daily Average Index Knee Pain While Walking (24h recall) by Week up to Week 4 - Primary Analysis	LF Table 14.2.1.1.2.1.4	Difference between treatment by week		SPECIFIED IN SAP	PRIMARY OUTCOME MEASURE	ADSL	CHG	Y	Adjusted least square mean	LF.SAP-SEC-10.4.1	PROC MIXED DATA=USL; RUN; CLASS WEEK TRT USUBJID COUNTRY; MODEL CHG_BASE - TRT WEEK BASE_VAL TRT WEEK COUNTRY / DIFF=KP; REPEATED WEEK TYPE=UN SUBJECT=USUBJID; ESTIMATE 'ISS1 100mg vs Placebo' TRT=110 / CL ALPHA=0.05; ESTIMATE 'ISS2 200mg vs Placebo' TRT=110 / CL ALPHA=0.05; ESTIMATE 'ISS3 100mg vs ISS1 100mg' TRT=0-11 / CL ALPHA=0.05; ODS OUTPUT LMEANS=lm.DIFFS=diff.ESTIMATES=est; RUN;
3	Table 14-5.02	Incidence of Treatment Emergent Serious Adverse Events by Treatment Group		Incidence of Treatment Emergent Serious Adverse Events by Treatment Group	Get denominator for percentage from ADSL and counts and numerator from ADAE Join ADAE with ADSL based on the unique subject identifier (USUBJID) keeping only records in ADAE for the numerator.	SPECIFIED IN SAP	PRIMARY OUTCOME MEASURE	ADSL			Unique count of subjects that experienced an Adverse Event by Preferred Term, System Organ Class, and Treatment Group and percent each based on the number of subjects in the safety population within each treatment group. The total number of times an event occurred is recorded by Preferred Term, System Organ Class, and Treatment Group. Fisher's exact test was used for treatment comparison of event rates.	LF Table 14-5.02.zw	
4	Table 14-5.02	Incidence of Treatment Emergent Serious Adverse Events by Treatment Group		Incidence of Treatment Emergent Serious Adverse Events by Treatment Group	Get denominator for percentage from ADSL and counts and numerator from ADAE Join ADAE with ADSL based on the unique subject identifier (USUBJID) keeping only records in ADAE for the numerator.	SPECIFIED IN SAP	PRIMARY OUTCOME MEASURE	ADAE	AEB0005T AEB0000D		Unique count of subjects that experienced an Adverse Event by Preferred Term, System Organ Class, and Treatment Group and percent each based on the number of subjects in the safety population within each treatment group. The total number of times an event occurred is recorded by Preferred Term, System Organ Class, and Treatment Group. Fisher's exact test was used for treatment comparison of event rates.	LF Table 14-5.02.zw	

Figure 7. ARM.XML Analysis

WhereClause

- Table Number:
- Analysis : Type of analysis to be performed
- Dataset: Dataset name
- Variable : Variable names
- Value: Value for each variable

PhUSE US Connect 2018

Table Number	Analysis	Dataset	Variable	Value
Table 14.2.1.1.2.1	Difference between treatment by week	ADKNEE	PARAMCD	MIKPW1
Table 14.2.1.1.2.1	Difference between treatment by week	ADKNEE	FASFL	Y
Table 14.2.1.1.2.1	Difference between treatment by week	ADKNEE	AVISITN	3
Table 14.2.1.1.2.1	Difference between treatment by week	ADKNEE	AVISITN	4
Table 14.2.1.1.2.1	Difference between treatment by week	ADKNEE	AVISITN	5
Table 14.2.1.1.2.1	Difference between treatment by over 4 weeks	ADKNEE	PARAMCD	MIKPW1
Table 14.2.1.1.2.1	Difference between treatment by over 4 weeks	ADKNEE	FASFL	Y
Table 14.2.1.1.2.1	Difference between treatment by over 4 weeks	ADKNEE	AVISITN	5
Table 14.2.1.1.2.1	Difference between treatment by over 4 weeks	ADKNEE	AVISITN	6
Table 14.2.1.1.2.1	Difference between treatment by over 4 weeks	ADKNEE	AVISITN	7
Table 14-5.02	Incidence of Treatment Emergent Serious Adverse Events by Treatment Group	ADSL	SAFFL	Y
Table 14-5.02	Incidence of Treatment Emergent Serious Adverse Events by Treatment Group	ADAE	TEAE	Y
Table 14-5.02	Incidence of Treatment Emergent Serious Adverse Events by Treatment Group	ADAE	AESER	Y

Figure 8. ARM.XML WhereClause

Documents

This section collects the following information about each external documentation that is being referred in ARM.XML specifications:

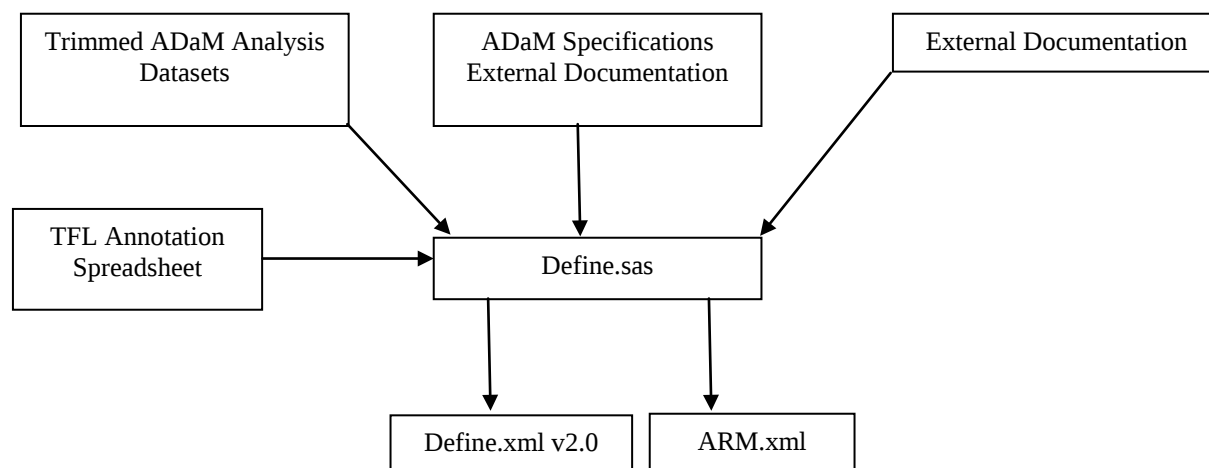
- Document ID: A unique short name with a prefix of "LF.", given to each external document.
- Title: Table Number that will be displayed in ARM.xml
- HRef: Physical location of the file with the actual file name and file extension.

ID	Title	href
LF.SAP-SEC-10.4	SAP Section 10.4	sap.pdf
LF.Table 14.2.1.1.2.1	Table 14.2.1.1.2.1	csr.pdf
LF.Table 14-5.02.sas	Table 14-5.02.sas	at14-5-02-sas.txt

Figure 9. ARM.XML Documets

PhUSE US Connect 2018

PROCESS FLOW OF GENREATING DEFINE.XML AND ARM.XML



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

Date of document generation: 2015-08-20T09:22:12
Stylesheet version: 2013-04-24

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 10. Define.xml v2.0 viewed via Microsoft Internet Explorer

PhUSE US Connect 2018

Table 14.2.1.1.2.1

Display	Table 14.2.1.1.2.1 Statistical Analysis (Repeated Measures Model) of Change from Baseline in Mean Daily Average Index Knee Pain While Walking (24h recall) by Week up to Week 4 - Primary Analysis
Analysis Result	Change from Baseline in Mean Daily Average Index Knee Pain While Walking
Analysis Parameter(s)	PARAMCD = "MIKPW1"
Analysis Variable(s)	CHG (Change from Baseline)
Analysis Reason	SPECIFIED IN SAP
Analysis Purpose	PRIMARY OUTCOME MEASURE
Data References (incl. Selection Criteria)	ADKNEE [PARAMCD = "MIKPW1" and AVISITN IN (4, 5, 6, 7) and FASFL = "Y"]
Documentation	Adjusted least squares mean SAP Section 10
Programming Statements	[SAS version 9.4] PROC MIXED DATA=adknee; CLASS avisitn trt01pn usubjid country ; MODEL chg = trt01pn avisitn base trt01pn*avisitn country / DDFM=RR; REPEATED avisitn/TYPE=UN SUBJECT= USUBJID; LSMEANS trt01pn*avisitn / PDIFF CL ; LSMEANS trt01pn / PDIFF CL; ESTIMATE "100 mg vs. Placebo" trt01pn -1 1 0 / CL ALPHA=0.05; ESTIMATE "200 mg vs. Placebo" trt01pn -1 0 1 / CL ALPHA=0.05; ESTIMATE "200 mg vs. 100 mg" trt01pn 0 -1 1 / CL ALPHA=0.05; ODS OUTPUT LSMEANS=lsm DIFFS=diff ESTIMATES=esti ; RUN;

Figure 11. ARM.xml

CONCLUSION

This data driven approach using ADaM Specifications, ADaM Datasets and TFL Annotation is a time-saving, straightforward and easy way of generating Define.xml v2.0 and ARM.xml. Programmers don't need to have XML knowledge for generating Define.xml or ARM.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 2018

CONTACT INFORMATION

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

Mark Crangle

ICON Clinical Research

Email:mark.crangle@iconplc.com

Hima Bhatia

ICON Clinical Research

Email:hima.bhatia@iconplc.com

Brand and product names are trademarks of their respective companies.