Paper SA04

Do's and Don'ts of Define.xml

David Roulstone, Pinnacle 21

ABSTRACT

The FDA's Technical Conformance Guide states that the define.xml "is considered arguably the most important part of the electronic dataset submission for regulatory review" [1]. Often, it can be difficult to determine what content to include versus what content is out of scope for a define file. References to raw database tables and variables often appear in define.xml comments, however these have no value to anyone without a deep understanding of the raw database. Some sponsors become very familiar with their own internal standards and fail to fully document how they populate sponsor defined variables (e.g. SPID, GRPID, REFID, etc.). This paper will aim to help people understand what content is suitable for a define file and show how to link the define.xml to external documents to further aid in documenting complicated data derivations.

INTRODUCTION

The define.xml describes the data within a data package and is both machine and human readable. CDISC standards are open to interpretation and sponsors often have their own internal standards that build upon CDISC theories and standards.

When creating the define.xml, careful consideration must be so that all the information within the define is clear and concise.

Someone attempting to review the data package, who is not only unfamiliar with the study, but also unfamiliar with the sponsors own internal standards, will need a define.xml that correctly and clearly describes all origins and derivations. If the define.xml is lacking or missing information, this will increase the amount of time it takes someone to become familiar with the data package.

In practice, the define.xml is created just prior to a submission, generally by someone who is very close to the data and derivations. In these situations, people can sometimes leave important information out of the define, because they forget that certain conventions are not governed by CDISC but are sponsor specific.

For the rarer cases where the define.xml is created while the study is still ongoing, often data and derivations change as protocol amendments are made and people focus on ensuring the data is correct and has been QC'ed, while the define.xml is neglected and the outdated origins and/or derivations remain in place.

Another common mistake is including references to raw data variables. Generally, references to raw data variables are only meaningful to certain people working for the sponsor. References to raw data variables should not be made in any submissions to the FDA or PMDA. Having references to raw data variables makes any derivations hard to read and clutters the define.xml with unnecessary information.

Even when correct information is provided in the define.xml, there can also be a problem with long comments and/or derivations, as text formatting is not supported within the define.xml. Items like new line characters, numbered lists, bullet points, etc. can be put into the define.xml, however it will not be displayed to the user the same as it would if the same text was in a format like a word document.

DERIVED VARIABLES LISTED AS ASSIGNED

Validation tools check that for all variables and value level items with Origin = Derived, there is a corresponding derivation. Unfortunately, there are occasions where the person creating the define.xml takes shortcuts and for variables that are truly derived, they list the origin as Assigned. This is done purely to save the time it takes to properly document all derivations. In reality, this makes the define.xml file almost worthless and brings into question the validity of the other information within the define.xml.

MISSING INFORMATION

When the define.xml is missing information, it often appears to be due to the define.xml creator not having a good understanding of the differences between sponsor and CDISC conventions, and the sponsor specific conventions are not properly documented.

For example, consider a sponsor that uses the xxSPID variable to provide traceability between the raw data, study data (SDTM/SEND) and analysis data (ADaM). The xxSPID value is a combination of raw database values that are concatenated together. These xxSPID values may be used daily for QC purposes and because everyone working with the data is so familiar with the xxSPID values, nobody remembers to document the purpose of xxSPID in the define.xml and reviewer's guides. Anyone new to the study must then waste time researching and coming to their own conclusions as to the meaning of xxSPID, or worse still, not realize how xxSPID could be used. This could be avoided with a one or two sentence explanation in the define.xml.

RAW DATA REFERENCES

Raw data is not submitted to the FDA or PMDA and therefore no references to raw database tables, variables, or values, should be made in the define.xml.

Derivations and comments within the define.xml are often extracted directly from the mapping specification being used for the study. The extra step must be taken where someone reviews all the comments and derivations, to ensure any raw data references are removed and replaced with meaningful text referencing the data package variable names and values.

Looking at Figure 1 below, AESPID should have an origin of eDT and a more meaningful comment that does not reference raw data variables.

The comment for AETERM is also confusing and reference raw data variables. The comment for AETERM should be re-written or removed.

The comments for AELLT, AELLTCD, AEDECOD and AEPTCD are useless. Technically they are referencing raw data variables, however the dataset and variable names exactly match those in SDTM. Overall it would be quicker to review the define.xml when it is not full of redundant information and references to raw data.

Adverse Events (AE) [Location: ae.xpt]

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin	Derivation/Comment
STUDYID	Study Identifier		text	200		Derived	SDTM.DM.STUDYID. Merge by Site Number and SUBJID.
DOMAIN	Domain Abbreviation		text	2	DOMAIN	Assigned	Set as AE
USUBJID	Unique Subject Identifier	1	text	200		Derived	SDTM.DM.STUDYID || "-" || DM.SITEID || "-" || DM.SUBJID
AESEQ	Sequence Number		integer	8		Derived	Sort by Primary Keys and assign sequence number in the ascending order. Restart numbering with each new subject.
AESPID	Sponsor-Defined Identifier		text	200		CRF Page	AE.RECORDPOSITION
AETERM	Reported Term for the Adverse Event	3	text	200		CRF Pages 140 148	1) AE.AETERM. 2) If YR.YRAEYN = Y then set YR.YRCOM and generate an alert message for this case.
AELLT	Lowest Level Term		text	200		Assigned	AE.AELLT
AELLTCD	Lowest Level Term Code		integer	8		Assigned	AE.AELLTCD
AEDECOD	Dictionary-Derived Term		text	200		Assigned	AE.AEDECOD
AEPTCD	Preferred Term Code		integer	8		Assigned	AE.AEPTCD

Figure 1. Derivations and Comments with references to raw data

INCORRECT INFORMATION

Throughout a clinical trials lifecycle, protocol amendments might result in changes to the data collection and/or migration. For sponsors who create the define.xml prior to study completion, these data updates will also need to be reflected in the define.xml. Unfortunately, people far too often focus on getting the data correct and past QC, while the define.xml file is never updated.

At the very least, comments and derivations that are very important and specific to the study, should be checked prior to the submission. For example, if a derivation for EXDOSE mentions three separate treatments, then the latest version of the protocol and the data should be checked to ensure that additional treatments are documented.

FORMATTING ISSUES

The define.xml does not support any formatting like bullet points, numbered lists, or new line characters. Most comments and derivations do not take up too much text and therefore the lack of formatting is not an issue. Large derivations with multiple conditions are difficult to read when formatting has been removed. In these instances, it's often better to create a separate PDF document to make use of formatting options.

Figure 2 shows the original derivation in the define.xml. It's extremely hard to read as all the text runs together. Figure 3 shows the same derivation after a quick update. The start of the derivation states to refer to the Complex Algorithms document and the hyperlink is provided at the very end of the derivation. Figure 4 is of the complexalgorithms.pdf document and has the same derivation, however now we can make use of formatting, making the derivation much easier to understand.

EXDOSE	Dose per Administration		float	5		Derived	1. For EX.EXCAT=For records with STUDY DRUG ADMINSITRATION': a. EXTRT=ABC123 check the following: If SUPPEX.QVAL WHERE QNAM=DISPAMT is equal to 0 then EXDOSE=0. If SUPPEX.QVAL WHERE QNAM=DISPAMT is greater than 0 then check : When EX.EXTRT=ABC123 and EXROUTE=ORAL, EX.EXSPID= 1 then EXDOSE=(EXENDTC-EXSTDTC)*(SUPPEX.QVAL WHERE QNAM=DISPAMT)*0.5, If EXSPID=2 then EXDOSE=(EXENDTC-EXSTDTC)* (SUPPEX.QVAL WHERE QNAM=DISPAMT)*0.75. If EXSPID >2 and DM.ACTARM=contains 1 mg then EXDOSE=(EXENDTC-EXSTDTC)* (SUPPEX.QVAL WHERE QNAM=DISPAMT)*1 If EXSPID >2 and DM.ACTARM=contains 1 mg then EXDOSE=(EXENDTC-EXSTDTC)* (SUPPEX.QVAL WHERE QNAM=DISPAMT)*1 b. When EXTRT=Placebo and (SUPPEX.QVAL WHERE QNAM=DISPAMT) >0 then EXDOSE=0 2. EXCAT='ON SITE DRUG ADMINSTRATION' and EXSTDTC non missing a. if VISIT=BASELINE/V1M1 then EXDOSE=0.5 b.if ACTARM contains 1 mg otherwise check c. If ACTARM contains 0.75 mg then EXDOSE=0.75

Figure 2. Large derivation without formatting

EXDOSE	Dose per Administration		float	5		Derived	Please see the complex algorithms document, linked below, to see the derivation with formatting applied. 1. For EX.EXCAT=For records with STUDY DRUG DOSING': a. EXTRT=ABC123 check the following: If SUPPEX.QVAL WHERE QNAM=DISPAMT is equal to 0 then EXDOSE=0. If SUPPEX.QVAL WHERE QNAM=DISPAMT is greater than 0 then check : When EX.EXTRT=ABC123 and EXROUTE=ORAL, EX.EXSPID= 1 then EXDOSE=(EXENDTC-EXSTDTC)* (SUPPEX.QVAL WHERE QNAM=DISPAMT)*0.5, If EXSPID=2 then EXDOSE= (EXENDTC-EXSTDTC)*(SUPPEX.QVAL WHERE QNAM=DISPAMT)*0.75. If EXSPID >2 and DM.ACTARM=contains 1.25 mg then EXDOSE=(EXENDTC-EXSTDTC)*(SUPPEX.QVAL WHERE QNAM=DISPAMT)*1.25 If EXSPID >2 and DM.ACTARM=contains 1.25 mg then EXDOSE=(EXENDTC-EXSTDTC)* (SUPPEX.QVAL WHERE QNAM=DISPAMT)*1.25 b. When EXTRT=Placebo and (SUPPEX.QVAL WHERE QNAM=DISPAMT) >0 then EXDOSE=0 2. EXCAT='ON SITE DRUG DOSING' and EXSTDTC non missing a. if VISIT=BASELINE/V1M1 then EXDOSE=0.5 b.if ACTARM contains 1.25 mg otherwise check c. If ACTARM contains 0.75 mg then EXDOSE=0.75 Complex Algorithms

Figure 3. Large derivation with hyperlink to external document

Dataset	Variable	Derivation

EX	EXDOSE	1. If DM.ACTARM = "0.75 mg ABC123" and SUPPEX.DISPAMT > 0 then set EXDOSE as follows:
		a. If EXSTDY < 8 then EXDOSE = SUPPEX.DISPAMT * 0.5;
		b. If EXSTDY > 7 then EXDOSE = SUPPEX.DISPAMT * 0.75;
		
		2. Else If DM.ACTARM = "1.25 mg ABC123" and SUPPEX.DISPAMT > 0 then set EXDOSE as follows:
		a. IF EXSTDY < 8 EXDOSE = SUPPEX.DISPAMT * 0.5;
		b. IF EXSTDY > 7 EXDOSE = SUPPEX.DISPAMT * 1.25.
		
		3. For records where EXTRT = "Placebo":
		a. If SUPPEX.DISPAMT or SUPPEX.TAKEDOSE = "Yes" then set EXDOSE equal to "0";
		b. Else set EXDOSE to Null.
		
		4. For records where EXTRT = "COMP":
		a. If SUPPEX.TAKEDOSE = "Yes" then set EXDOSE equal to "30";
		b. Else set EXDOSE to Null.
		
		5. For records where EXCAT = "ON SITE DOSING":
		a. If EX.VISIT = "Month 1" and EX.EXSTDTC ^= null then set EXDOSE equal to "0.5".
		b. For subjects records where EX.VISIT ^= "Month 1" and EX.EXSTDTC ^= null and DM.ACTARM = "0.75 mg ABC123", set EXDOSE equal to "0.75".
		c. For subjects records where EX.VISIT ^= "Month 1" and EX.EXSTDTC ^= null and DM.ACTARM = "1.25 mg ABC123", set EXDOSE equal to "1.25".
		d. For subjects records where EX.VISIT ^= "Month 1" and EX.EXSTDTC ^= null and DM.ACTARM = "Comparator", set EXDOSE equal to "0".
		
		6. For records where EXTRT = "Placebo":
		a. If SUPPEX.DISPAMT > 0 or SUPPEX.TAKEDOSE = "Yes" then Set EXDOSE equal to "0";
		b. Else set EXDOSE to Null.
		
		7. For records where EXTRT = "COMP":
		a. If SUPPEX.TAKEDOSE = "Yes" then Set EXDOSE equal to "30";
		b. Else set EXDOSE to Null.

Figure 4. Large derivation with formatting applied

SCATTERED INFORMATION

Hyperlinks within the define.xml make it easy and quick to navigate to different sections. Sometimes all of this information remains separated, when it would be better to be grouped together. Figure 5 shows an example where all comments and derivations are lumped together as computational methods. Figure 6 shows the computational algorithms section of the define.xml. The problem is that the comments for EXDOSE and EXDOSU cannot be viewed together. First one must read one comment, then navigate to the other comment. Additionally, the naming convention for the comments is not user friendly. Looking purely at the comments/derivations in Figure 6, there's no way to tell DER107 corresponds to EXDOSFRM while DER108 corresponds to EXDOSFRQ.

Figure 7 has all comments/derivations present in the "Derivation/Comment" column on the right, meaning the user does not need to navigate through the define.xml as much, and also allows for multiple, related, comments/derivations to be displayed on the screen at the same time.

Exposure Dataset (EX)						ex.xpt
Variable	**Label**	**Type**	**Controlled Terms or Format**	**Origin**	**Role**	**Comment**
STUDYID	Study Identifier	text		ASSIGNED	Identifier	See Computational Method: DER229
DOMAIN	Domain Abbreviation	text	CODELISTC7	ASSIGNED	Identifier	See Computational Method: DER104
USUBJID	Unique Subject Identifier	text		DERIVED	Identifier	See Computational Method: DER271
EXSEQ	Sequence Number	integer		DERIVED	Identifier	See Computational Method: DER220
EXTRT	Name of Treatment	text		CRF Page 59	Topic	See Computational Method: DER301
EXDOSE	Dose	float		CRF Page 59	RecordQualifier	See Computational Method: DER302
EXDOSU	Dose Units	text	CODELISTC43	CRF Page 59	VariableQualifier	See Computational Method: DER109
EXDOSFRM	Dose Form	text	CODELISTC14	ASSIGNED	VariableQualifier	See Computational Method: DER107
EXDOSFRQ	Dosing Frequency per Interval	text	CODELISTC13	CRF	VariableQualifier	See Computational Method: DER108
EXROUTE	Route of Administration	text	CODELISTC37	CRF Page 59	VariableQualifier	See Computational Method: DER111
EXADJ	Reason for Dose Adjustment	text		CRF Page 59	RecordQualifier	See Computational Method: DER105
VISITNUM	Visit Number	float		DERIVED	Timing	See Computational Method: DER114
VISIT	Visit Name	text		CRF	Timing	See Computational Method: DER272

Figure 5. Variable level with hyperlinks to derivations and comments

Computational Algorithms (DER104)	
Reference Name	**Computation Method**
DER104	Set as EX
Computational Algorithms (DER105)	
Reference Name	**Computation Method**
DER105	EX.EXADJ
Computational Algorithms (DER302)	
Reference Name	**Computation Method**
DER302	EX.EXDOSE. Amount of EXTRT when numeric.
Computational Algorithms (DER107)	
Reference Name	**Computation Method**
DER107	Set the value using protocol info if its not available directly from CRF.
Computational Algorithms (DER108)	
Reference Name	**Computation Method**
DER108	Set the value using protocol info if its not available directly from CRF.
Computational Algorithms (DER109)	
Reference Name	**Computation Method**
DER109	EX.EXDOSU.
Computational Algorithms (DER110)	
Reference Name	**Computation Method**
DER110	Set using SDTM.EX.EXSTDTC.
Computational Algorithms (DER111)	
Reference Name	**Computation Method**
DER111	EX.EXROUTE

Figure 6. Derivations and comments with confusing naming convention

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin	Derivation/Comment
STUDYID	Study Identifier		text	200			SDTM.DM.STUDYID. Merge by Site Number and SUBJID.
DOMAIN	Domain Abbreviation		text	2	DOMAIN		Set as EX
USUBJID	Unique Subject Identifier	1	text	200			SDTM.DM.STUDYID || "-" || DM.SITEID || "-" || DM.SUBJID
EXSEQ	Sequence Number		integer	8			Sort by Primary Keys and assign sequence number in the ascending order. Restart numbering with each new subject.
EXTRT	Name of Treatment	4	text	200		CRF Page 59	Set using EX.EXTRT
EXDOSE	Dose		float	8		CRF Page 59	EX.EXDOSE. Amount of EXTRT when numeric.
EXDOSU	Dose Units		text	200	UNIT	CRF Page 59	EX.EXDOSU.
EXDOSFRM	Dose Form		text	200	["SOLUTION" = "Solution"] <FRM>		Set the value using protocol info if its not available directly from CRF.
EXDOSFRQ	Dosing Frequency per Interval		text	200	["ONCE" = "Once"] <FREQ>	CRF Page	Set the value using protocol info if its not available directly from CRF.
EXROUTE	Route of Administration		text	200	["ORAL" = "Intraoral Route of Administration; PO"] <ROUTE>	CRF Page 59	EX.EXROUTE
EXADJ	Reason for Dose Adjustment		text	200		CRF Page 59	EX.EXADJ
VISITNUM	Visit Number	2	float	8			Set using SDTM.SV.VISITNUM. Numeric version of VISIT.
VISIT	Visit Name		text	200		CRF Page	Corresponding domain.VISIT

Figure 7. Derivations and comments shown

CODELIST ISSUES

ONE CODELIST FOR ALL UNIT VALUES ACROSS THE DATA PACKAGE

The FDA has stated that a "separate UNIT codelist not used for each variable" is a common issue that impacts review[2]. If the exposure domain has one unique unit, then the define.xml codelist associated with EXDOSU should only hold that one value.

Far too often, all units used across an entire study are lumped into one codelist. Looking just at the define.xml, people are unable to distinguish which units are used for laboratory results versus units used for exposure. Instead they must open the data to pull unique unit values, therefore making define.xml codelist useless.

CODELISTS MEANT FOR OTHER DATA PACKAGES

A define.xml for SDTM data, should only hold codelists related to the SDTM data for that study. There is no need to include codelists for analysis data (ADaM) in the define.xml for study data (SEND or SDTM). Nor is there a need to have codelists for study data, in an analysis data define.xml.

CODELISTS WITH ALL VALUES FROM THE CDISC CODELIST

A codelist in the define.xml should hold all possible values for the datapoint(s) assigned to it. If the Adverse Events CRF page has 5 options for 'Action Taken with Study Treatment', then all 5 options should be present in the define.xml codelist, even if only 3 of those options are present in the data.

For a variable like 'Action Taken with Study Treatment', since the CDISC codelist contains just 7 values, including all 7 values, when only 5 values are relevant to the study, shouldn't be too confusing for the end user. Laboratories are a different story. There are close to 2000 laboratory tests in the CDISC LBTESTCD and LBTEST codelists, while the unit codelist has over 500 values. Studies with around 50 different lab tests and around 30 different lab unit values are quite common, and in these cases including nearly 2000 lab tests in the define.xml codelists will only confuse the end user as it takes time to conclude that the codelists are useless. It would probably be easier and quicker to simply not create LBTESTCD and LBTEST codelists, rather than simply include all possible CDISC values.

CONCLUSION

The person creating the define.xml will most often have worked on the study for a longer time period of time and have a better understanding of the study/data/derivations than anyone else who will use the define.xml in the future. This must be kept in mind when creating the define.xml.

People should assume that the end consumers of the define.xml:

- have never seen the study before
- have not read the protocol in detail
- have no access to, or information about, the raw database
- have little understanding of CDISC standards
- are attempting to gain familiarity with the study, data, and documents, and are turning to the define.xml for explanations

REFERENCES

1. U.S. Food and Drug Administration, “Study Data Technical Conformance Guide” October 2017

   https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

2. DeYett Law, Crystal Allard, Mary Doi, Lilliam Rosario, Barbara Witczak, Jesse Anderson, Kathryn Matto, Austin Taylor, Jeno Pizarro, Margo Cohen, “Data Quality Findings from JumpStart” PhUSE CSS March 2017

   http://www.phusewiki.org/docs/2017_CSS_US/PP29_Draft.pdf

RECOMMENDED READING

- *CDISC Define-XML Specification Version 2.0*

CONTACT INFORMATION

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

David Roulstone
Pinnacle 21
droulstone@pinnacle21.net