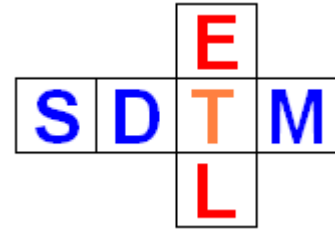


SDTM-ETL 4.0: Working with Define-XML 2.1

Author: Jozef Aerts, XML4Pharma

Last update: 2020-05-25



Introduction


Define-XML v.2.1 was published by CDISC exactly one year ago. It has some major improvements of v.2.0 which is currently also the version accepted by the FDA. It is however clear that the FDA will soon start accepting v.2.1 (see further). Therefore, and for many other obvious reasons, we have implemented Define-XML v.2.1 in SDTM-ETL

How to start with v.2.1 of the Define-XML standard?

After having loaded an ODM (source) file with the metadata of your study, you will usually start loading an SDTM template for the SDTM- or SEND-IG version of your choice. This is done using the menu "File – Create define.xml" (or use CTRL-n). The following dialog appears¹:

¹ The dialog may slightly be different depending on which standards to be used have been defined in (or added) to your SDTM_SEND_standards.xml file.

SDTM/SEND Version ×

 Do you want to work with the SDTM-IG or SEND-IG CDISC Standard?

SDTM-IG Standard SEND-IG Standard

Which version of the Standard would you like to work with?

SDTM-IG 3.1.2
 SDTM-IG 3.1.3
 SDTM-IG 3.2
 SDTM-IG 3.3
 SDTM-IG MD.1.0
 SDTM-IG MD.1.1
 SDTM-IG AP.1.0
 SDTM-IG PGx.1.0

Define.xml version:

define.xml 1.0 define.xml 2.0 define.xml 2.1

Controlled Terminology Version:

2019-06-28
2019-09-27
2019-12-20
2020-03-27
2020-05-08

You can select a SDTM or SEND version. In order to generate a define.xml 2.1 that will be used to store all your mappings, select the radiobutton "define.xml 2.1", then select a codelist version (the list is automatically generated from the contents of your "CDISC-CT" directory). After having clicked, the corresponding template for the selected SDTMIG/SENDIG in define.xml v.2.1 form will be loaded.

This may take a few minutes, ideal for a fresh cup of coffee ...

Once the template loaded:

Domain	Variable	Variable	Variable	Variable	Variable	Variable	Variable	Variable	Variable
CO	STUDYID	DOMAIN	RDOMAIN	USUBJID	USUBJID	CO.COSEQ	CO.IDIVAR	CO.IDIVARVAL	CO.COREP
DM	STUDYID	DOMAIN	USUBJID	SUBJID	DM.RFSTDTCTC	DM.RFENDTCTC	DM.RFXSTDTCTC	DM.RFXENDTCTC	DM.RFV
SE	STUDYID	DOMAIN	USUBJID	SE.SESEQ	SE.ETCD	SE.ELEMENT	SE.TAETORD	SE.EPOCH	SE.SEEST
SM	STUDYID	DOMAIN	USUBJID	SM.SMSEQ	SM.MIDSD	SM.MIDSTYPE	SM.SMSTDTCTC	SM.SMENDTCTC	SM.SMS
SV	STUDYID	DOMAIN	USUBJID	SV.VISITNUM	SV.VISIT	SV.VISITDY	SV.SYSTDTCTC	SV.VSNTCTC	SV.SVST
AG	STUDYID	DOMAIN	USUBJID	AG.AGSEQ	AG.AGGRPID	AG.AGSPID	AG.AGLNKID	AG.AGLNKGRP	AG.AGT
CM	STUDYID	DOMAIN	USUBJID	CM.CMSEQ	CM.CMGRPID	CM.CMSPID	CM.CMTRT	CM.CMMODIFY	CM.CMTD
EX	STUDYID	DOMAIN	USUBJID	EX.EXSEQ	EX.EXGRPID	EX.EXREFID	EX.EXSPID	EX.EXLNKID	EX.EXLN
EC	STUDYID	DOMAIN	USUBJID	EC.ECSEQ	EC.ECGRPID	EC.ECREFID	EC.ECSPID	EC.ECLNKID	EC.ECLN
ML	STUDYID	DOMAIN	USUBJID	ML.MLSEQ	ML.MLGRPID	ML.MLSPID	ML.MLTRT	ML.MLCAT	ML.MLS
PR	STUDYID	DOMAIN	USUBJID	PR.PRSEQ	PR.PRGRPID	PR.PRSPID	PR.PRLNKID	PR.PRLNKGRP	PR.PRTI
SU	STUDYID	DOMAIN	USUBJID	SU.SUSEQ	SU.SUGRPID	SU.SUSPID	SU.SUTRT	SU.SUMODIFY	SU.SUD
AE	STUDYID	DOMAIN	USUBJID	AE.AESEQ	AE.AEGRPID	AE.AEREFID	AE.AESPID	AE.AETERM	AE.AEM
CE	STUDYID	DOMAIN	USUBJID	CE.CESEQ	CE.CEGRPID	CE.CEREFID	CE.CESPID	CE.CETERM	CE.CED
DS	STUDYID	DOMAIN	USUBJID	DS.DSSEQ	DS.DSGRPID	DS.DSREFID	DS.DSSPID	DS.DSTERM	DS.DSD
DV	STUDYID	DOMAIN	USUBJID	DV.DVSEQ	DV.DVREFID	DV.DVSPID	DV.DVTERM	DV.DVDECOD	DV.DVC
HO	STUDYID	DOMAIN	USUBJID	HO.HOSEQ	HO.HOGRPID	HO.HOREFID	HO.HOSPID	HO.HOTERM	HO.HOC
MH	STUDYID	DOMAIN	USUBJID	MH.MHSEQ	MH.MHGRPID	MH.MHREFID	MH.MHSPID	MH.MHTERM	MH.MHM
DA	STUDYID	DOMAIN	USUBJID	DA.DASEQ	DA.DAGRPID	DA.DAREFID	DA.DASPID	DA.DATERM	DA.DATE
DD	STUDYID	DOMAIN	USUBJID	DD.DDSEQ	DD.DDTESTCD	DD.DDTEST	DD.DDORRES	DD.DDSTRESC	DD.DDR
EG	STUDYID	DOMAIN	USUBJID	EG.SPDEVID	EG.EGSEQ	EG.EGGRPID	EG.EGREFID	EG.EGTERM	EG.EGT
IE	STUDYID	DOMAIN	USUBJID	IE.IESEQ	IE.IESPID	IE.IESTESTCD	IE.IESTEST	IE.IECAT	IE.IESCA
IS	STUDYID	DOMAIN	USUBJID	IS.ISSEQ	IS.ISGRPID	IS.ISREFID	IS.ISSPID	IS.ISTESTCD	IS.ISTES
LB	STUDYID	DOMAIN	USUBJID	LB.LBSEQ	LB.LBGRPID	LB.LBREFID	LB.LBSPID	LB.LBTESTCD	LB.LBTE
MB	STUDYID	DOMAIN	USUBJID	MB.MBSEQ	MB.MBGRPID	MB.MBREFID	MB.MBSPID	MB.MBTERM	MB.MBL
MS	STUDYID	DOMAIN	USUBJID	MS.MSSEQ	MS.MSGRPID	MS.MSREFID	MS.MSSPID	MS.MSTERM	MS.MSL
MI	STUDYID	DOMAIN	USUBJID	MI.MISEQ	MI.MIGRPID	MI.MIREFID	MI.MISPID	MI.MITESTCD	MI.MITES
MO	STUDYID	DOMAIN	USUBJID	MO.MOSEQ	MO.MOGRPID	MO.MOREFID	MO.MOSPID	MO.MOLNKID	MO.MOT
CV	STUDYID	DOMAIN	USUBJID	CV.CVSEQ	CV.CVGRPID	CV.CVREFID	CV.CVSPID	CV.CVLNKID	CV.CVLN
MK	STUDYID	DOMAIN	USUBJID	MK.MKSEQ	MK.MKGRPID	MK.MKREFID	MK.MKSPID	MK.MKTERM	MK.MKL
NV	STUDYID	DOMAIN	USUBJID	NV.NVSEQ	NV.NVGRPID	NV.NVREFID	NV.NVSPID	NV.NVTERM	NV.NVLT
OE	STUDYID	DOMAIN	USUBJID	OE.OESEQ	OE.OEGRPID	OE.OEREFID	OE.OESPID	OE.OELNKID	OE.OELN
RP	STUDYID	DOMAIN	USUBJID	RP.RPSEQ	RP.RPGRPID	RP.RPREFID	RP.RPSPID	RP.RPLNKID	RP.RPLN
RE	STUDYID	DOMAIN	USUBJID	RE.RESEQ	RE.REGRPID	RE.REFERID	RE.RESPID	RE.RETERM	RE.RELI
UR	STUDYID	DOMAIN	USUBJID	UR.URSEQ	UR.URGRPID	UR.URREFID	UR.URSPID	UR.URLNKID	UR.URLN
PC	STUDYID	DOMAIN	USUBJID	PC.PCSEQ	PC.PCGRPID	PC.PCREFID	PC.PCSPID	PC.PCTESTCD	PC.PCTI
PP	STUDYID	DOMAIN	USUBJID	PP.PPSEQ	PP.PPGRPID	PP.PPREFID	PP.PPSPID	PP.PPTESTCD	PP.PPTI
PE	STUDYID	DOMAIN	USUBJID	PE.PESEQ	PE.PEGRPID	PE.PEREFID	PE.PESPID	PE.PETERM	PE.PETI

Very often, one will limit the number of domains that is visible to the ones one want to start with. This can easily be done using the menu "View – View/Hide Domains".

Just for the tutorial, we will only work with the domains DM (Demographics), EX (Exposure), AE (Adverse Events), LB (Laboratory) and VS (Vital Signs), so we hide all other ones. After this has been done, we get a much better oversight:

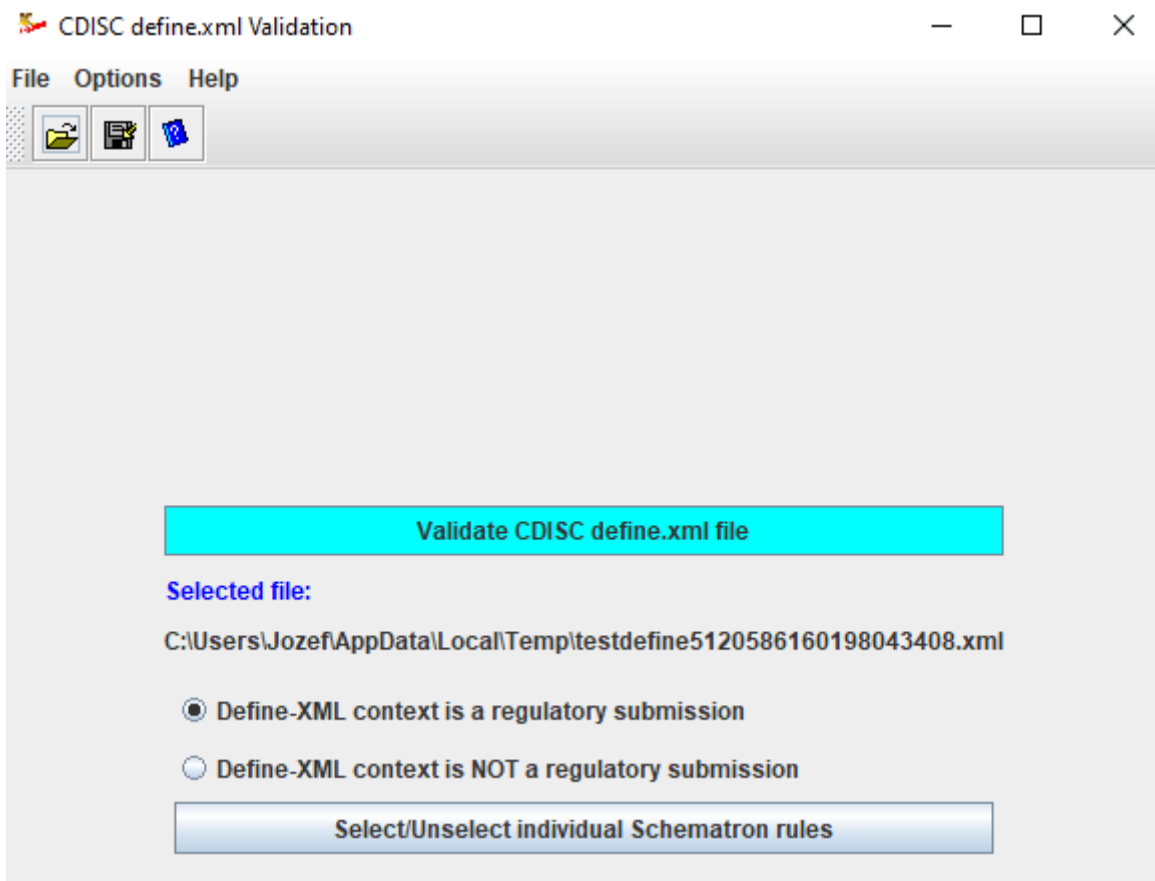
Domain	Variable	Variable	Variable	Variable	Variable	Variable	Variable	Variable
DM	STUDYID	DOMAIN	USUBJID	SUBJID	DM.RFSTDTCTC	DM.RFENDTCTC	DM.RFXSTDTCTC	DM.RFXENDTCTC
EX	STUDYID	DOMAIN	USUBJID	EX.EXSEQ	EX.EXGRPID	EX.EXREFID	EX.EXSPID	EX.EXLNKID
AE	STUDYID	DOMAIN	USUBJID	AE.AESEQ	AE.AEGRPID	AE.AEREFID	AE.AESPID	AE.AETERM
LB	STUDYID	DOMAIN	USUBJID	LB.LBSEQ	LB.LBGRPID	LB.LBREFID	LB.LBSPID	LB.LBTESTCD
VS	STUDYID	DOMAIN	USUBJID	VS.VSSEQ	VS.VSGRPID	VS.VSREFID	VS.VSSPID	VS.VSTESTCD

Variables that are colored red are "required", the blue ones "expected" and the green ones "permissible".

It is of course also always possible to add additional allowed variables (such as additional timing variables) and "non-standard" variables (NSVs), the latter will finally be "banned" to SUPPQUAL datasets. How to add additional variables is explained in another tutorial.

The first new important feature in the Define-XML 2.1 standard is the possibility to set whether the define.xml is meant to be in the context of a regulatory submission (def:Context attribute in the define.xml). In the SDTM-ETL software, this will each time be asked for when either validating the define.xml, or when writing the define.xml to file.

For example, when using the menu "Validate – Validate define.xml structure", we get:



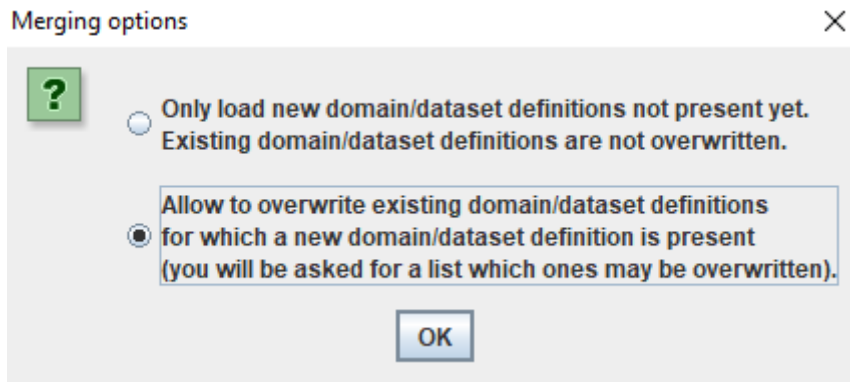
When the radiobutton "Define-XML context is a regulatory submission", the underlying "def:Context" in the define.xml will be set to "Submission". If the second radiobutton is selected, it will be set to "Other". See the ["Define-XML v.2.1 specification"](#) for more details.

An important new feature in the Define-XML 2.1 standard is to "merge" versions of the SDTM-IG or SEND-IG. This is especially useful when new domains have been developed, e.g. as part of a "Therapeutic Area User Guide" (TAUG). These are usually not immediately available as an official SDTM-IG domain, but often appear in the draft version of the next SDTMIG version.

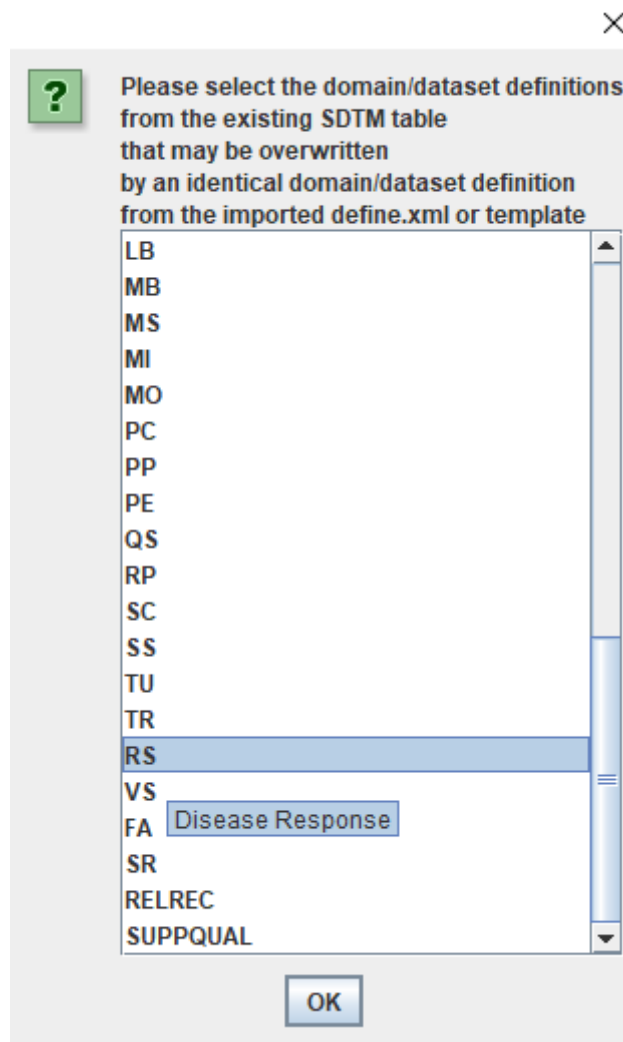
For example, the TAUG "Hepatitis-C" describes a CC domain "Clinical Classifications" which is not described present in the SDTMIG v.3.2. In version 3.3 of the SDTM-IG however, it is described as the domain "Disease Response and Clin Classification" with the domain code "RS". So when a sponsor is using SDTM-IG 3.2, and wants to implement the TAUG "Hepatitis-C", it can set up a "sponsor-defined" CC domain, using the menu "Insert – Sponsor defined SDTM domain" and then add SDTM v.1.4 variables to it, and probably also some "non-standard" variables (NSVs) and then stating in the define.xml that this domain comes from the TAUG "Hepatitis-C". The second possibility is to use the SDTMIG v.3.3 domain RS (Disease Response and Clin Classification) and make the statement in the define.xml that this is not an SDTMIG 3.2 domain, but an SDTMIG 3.3 domain. Remark that "RS" in SDTMIG 3.2 was named "Disease Response", had only 26 variables (in 3.3 it has 45 variables), so it looks as the scope of "RS" was extended from moving from SDTMIG 3.2 to 3.3.

When merging the SDTMIG 3.3 template (using "File – Load Template define.xml") into an already loaded SDTMIG 3.2 template, by default, only the domains that were not already in SDTMIG 3.2 will be added (i.e. the "new" domains), so in case one wants the "new" RS domain definition (v.3.3) to replace the "old" RS domain definition (v.3.2) one will need to

indicate this by clicking the radiobutton "Allow to overwrite existing domain/dataset definitions":



After which a list appears of the existing domain/dataset definitions, and one selects the ones that one want to be overwritten (in our case by the SDTMIG 3.3 versions). In our case, we want to use the RS domain of v.3.3 (and remove the 3.2 version):



After clicking OK, the 3.3 datasets are loaded, and the merging starts. This can take 1-2 minutes (you can follow the progress in the console). As a result, we get:

VS	STUDYID	DOMAIN	USUBJID	VS.VSSEQ	VS.VSGRPID	VS.VSSPID	VS.VSTESTCD	VS.VSTEST
FA	STUDYID	DOMAIN	USUBJID	FA.FASEQ	FA.FAGRPID	FA.FASPID	FA.FATESTCD	FA.FATEST
SR	STUDYID	DOMAIN	USUBJID	SR.SRSEQ	SR.SRGRPID	SR.SRREFID	SR.SRSPID	SR.SRTEST
RELREC	STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	RELTYPE	RELID	
SUPPQUAL	STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL
SM	STUDYID	DOMAIN	USUBJID	SM.SMSEQ	SM.MIDS	SM.MIDSTYPE	SM.SMSTDTC	SM.SMENDT
AG	STUDYID	DOMAIN	USUBJID	AG.AGSEQ	AG.AGGRPID	AG.AGSPID	AG.AGLNKID	AG.AGLNKG
ML	STUDYID	DOMAIN	USUBJID	ML.MLSEQ	ML.MLGRPID	ML.MLSPID	ML.MLTRT	ML.MLCAT
CV	STUDYID	DOMAIN	USUBJID	CV.CVSEQ	CV.CVGRPID	CV.CVREFID	CV.CVSPID	CV.CVLNKID
MK	STUDYID	DOMAIN	USUBJID	MK.MKSEQ	MK.MKGRPID	MK.MKREFID	MK.MKSPID	MK.MKLNKID
NV	STUDYID	DOMAIN	USUBJID	NV.FOCID	NV.NVSEQ	NV.NVGRPID	NV.NVREFID	NV.NVSPID
OE	STUDYID	DOMAIN	USUBJID	OE.FOCID	OE.OESEQ	OE.OEGRPID	OE.OELNKID	OE.OELNKG
RE	STUDYID	DOMAIN	USUBJID	RE.SPDEVID	RE.RESEQ	RE.REGRPID	RE.REFERID	RE.RESPID
UR	STUDYID	DOMAIN	USUBJID	UR.URSEQ	UR.URGRPID	UR.URREFID	UR.URSPID	UR.URLNKID
FT	STUDYID	DOMAIN	USUBJID	FT.FTSEQ	FT.FTGRPID	FT.FTREFID	FT.FTSPID	FT.FTTESTCD
RS	STUDYID	DOMAIN	USUBJID	RS.RSSEQ	RS.RSGRPID	RS.RSREFID	RS.RSSPID	RS.RSLNKID
TM	STUDYID	DOMAIN	USUBJID	TM.TMDEF	TM.TMRPT			
RELSUB	STUDYID	DOMAIN	USUBJID	RELSUB.RELSUB...	RELSUB.SREL			
OI	STUDYID	DOMAIN	USUBJID	OI.OISEQ	OI.OIPARMCD	OI.OIPARM	OI.OIVAL	

define.xml information:

OID: RS

Name: RS

Domain: RS

SAS Dataset Name: RS

Purpose: Tabulation

Repeating: Yes

Is Reference Data: No

Label: Disease Response and Clin Classification

Class: FINDINGS

Structure: One record per RS.RSTESTCD per USUBJID

Archive Location ID: Location.RS

Standard-OID: STD.SDTMIG-3.3

All the SDTMIG 3.3 domains that are not in SDTMIG 3.2 have been loaded (starting from SM), plus the SR domain of v.3.3, replacing the one from version 3.2, which has been removed from the template table.

As we are working with define.xml 2.1, one also sees in the tooltip that the RS domain comes from STMIG v.3.3, whereas e.g. the VS domain still comes from the SDTMIG 3.2:

TR	STUDYID	DOMAIN	USUBJID	TR
VS	STUDYID	DOMAIN	USUBJID	VS
FA	STUDYID	DOMAIN	USUBJID	FA
SR	STUDYID	DOMAIN	USUBJID	SR
RELREC	STUDYID	RDOMAIN	USUBJID	RELREC
SUPPQUAL	STUDYID	RDOMAIN	USUBJID	SUPPQUAL
SM	STUDYID	DOMAIN	USUBJID	SM
AG	STUDYID	DOMAIN	USUBJID	AG
ML	STUDYID	DOMAIN	USUBJID	ML
CV	STUDYID	DOMAIN	USUBJID	CV
MK	STUDYID	DOMAIN	USUBJID	MK
NV	STUDYID	DOMAIN	USUBJID	NV
OE	STUDYID	DOMAIN	USUBJID	OE
RE	STUDYID	DOMAIN	USUBJID	RE
UR	STUDYID	DOMAIN	USUBJID	UR
FT	STUDYID	DOMAIN	USUBJID	FT
RS	STUDYID	DOMAIN	USUBJID	RS
TM	STUDYID	DOMAIN	USUBJID	TM

define.xml information:

OID: VS

Name: VS

Domain: VS

SAS Dataset Name: VS

Purpose: Tabulation

Repeating: Yes

Is Reference Data: No

Label: Vital Signs

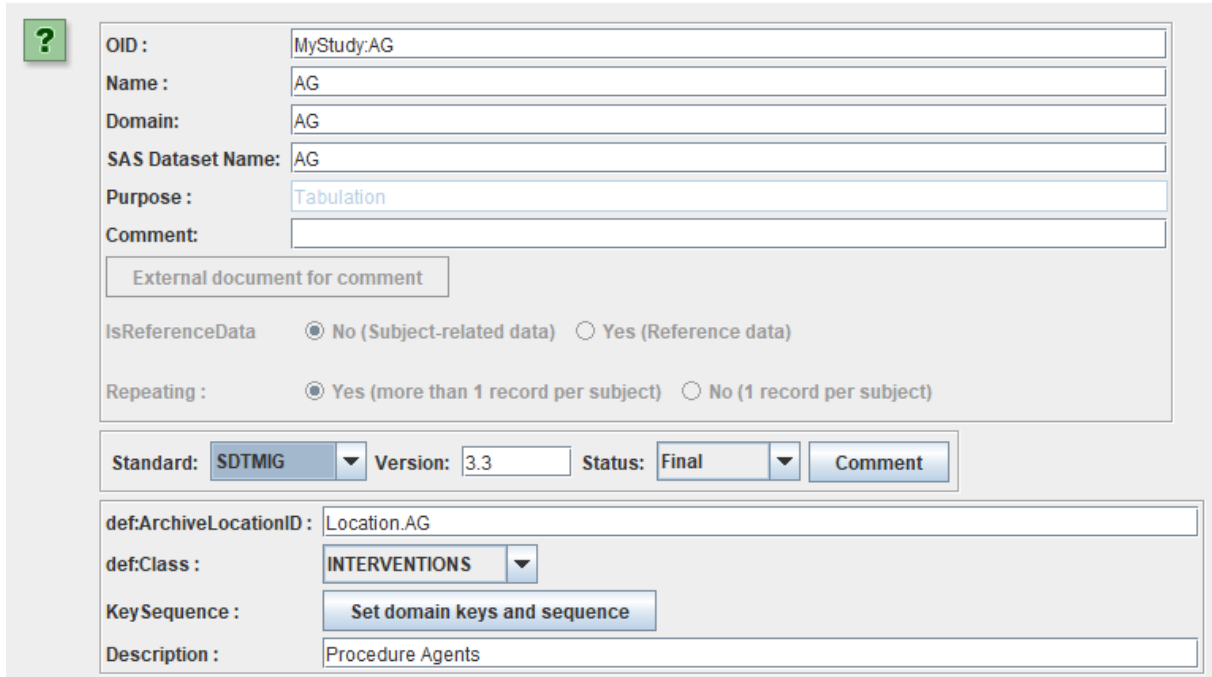
Class: FINDINGS

Structure: One record per VS.VSTESTCD per USUBJID

Archive Location ID: Location.VS

Standard-OID: STD.SDTMIG-3.2

When now selecting one of the "new" (SDTMIG 3.3) domains, like AG (Procedure Agents) and then using "Edit – SDTM dataset properties" after having created (by drag-and-drop) a study-specific instance, one gets:



OID : MyStudy:AG

Name : AG

Domain: AG

SAS Dataset Name: AG

Purpose : Tabulation

Comment:

External document for comment

IsReferenceData No (Subject-related data) Yes (Reference data)

Repeating : Yes (more than 1 record per subject) No (1 record per subject)

Standard: SDTMIG Version: 3.3 Status: Final Comment

def:ArchiveLocationID : Location.AG

def:Class : INTERVENTIONS

KeySequence : Set domain keys and sequence

Description : Procedure Agents

Changing the SDTMIG version manually, when loaded from a template, is of course discouraged, but it is possible. Setting the version manually can however be a good choice when e.g. having created a domain/dataset definition from a draft SDTMIG.

TIP: This also works with define.xml 2.0, but then the information whether the domain belongs to SDTMIG 3.2 or 3.3 is lost. FDA does not accept define.xml 2.1 nor SDTMIG 3.3 (status May 2020), so we often get the question what to do when one would like to use one of the new 3.3 domains anyway. The answer is pretty simple: use the 3.3 domain and declare it as a "sponsor-defined domain". When doing so with define.xml 2.0, this should be documented in the SDRG.

We do however expect that the FDA will start accepting SDTMIG 3.3 and define.xml 2.1 very soon, as the "conformance rules" for them have already or will soon be published².

² This was a requirement from the FDA towards CDISC for starting accepting any new standard.