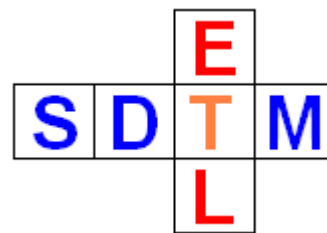


SDTM-ETL 3.1 User Manual and Tutorial

Author: Jozef Aerts, XML4Pharma

Last update: 2015-09-13



Creating and editing Trial Design datasets

When starting from an ODM file with metadata including SDM-XML ([Study Design Model in XML](#)), it is pretty straightforward to generate trial design datasets from the SDM-XML using the SDTM-ETL software (see separate manual).

However, not everyone is using SDM-XML yet, and most EDC systems do even not export SDM-XML. So, how can one generate the trial design datasets when the information is not, or only limited, in the ODM file with the study metadata?

We have recognized this problem, and added a new module to the SDTM-ETL software: a "trial design dataset editor". This editor does not only allow to generate trial design datasets from scratch, but also allows to edit existing trial design in Dataset-XML format¹.

The editor can also be run in standalone method (so without using the SDTM-ETL software). This is explained later in this document.

In this tutorial, we will demonstrate the use of the "trial design dataset editor" using the TA (Trial Arms) as an example.

Starting the editor from within SDTM-ETL

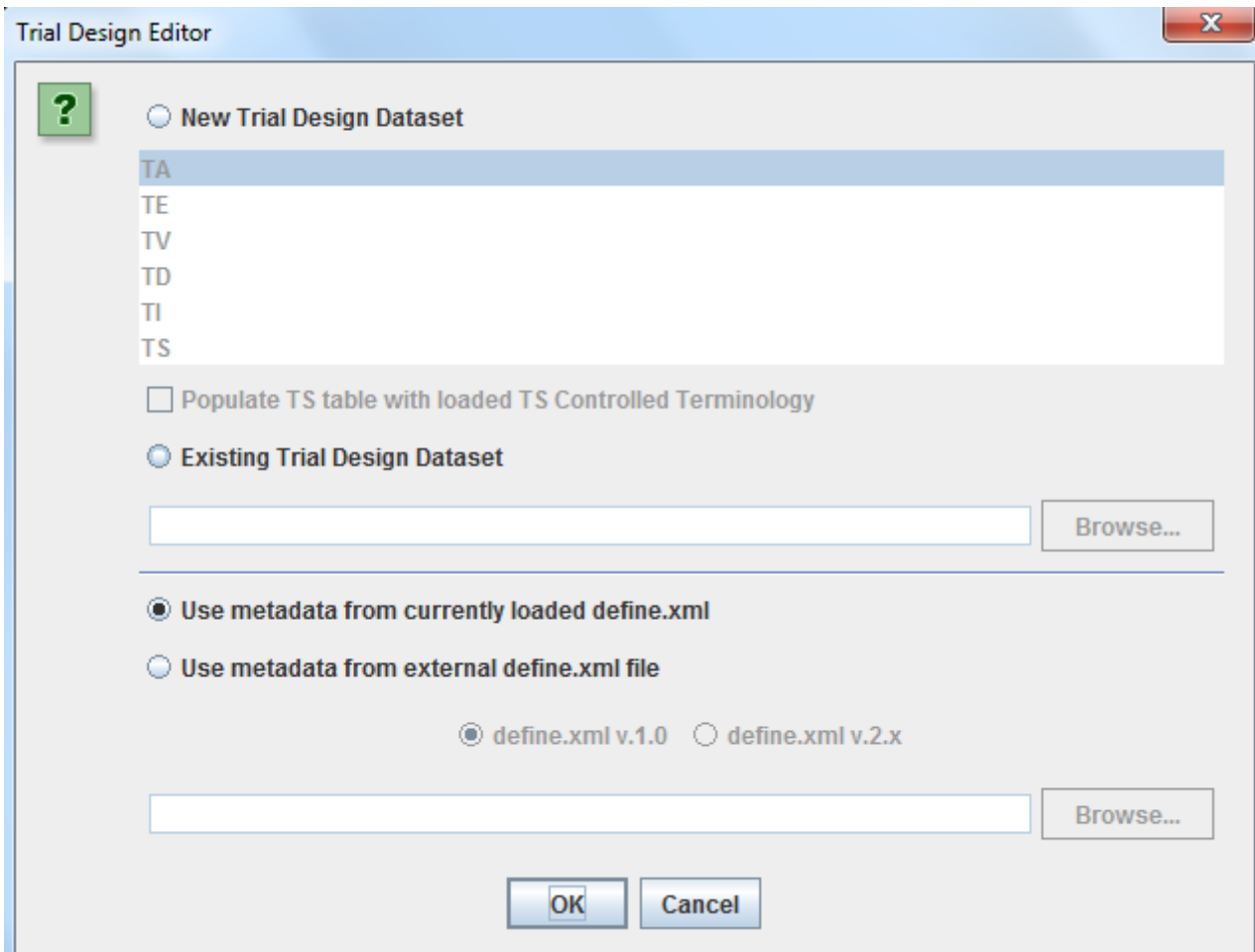
After having loaded an SDTM template or existing define.xml with SDTM-ETL mappings, create a study-specific instance of the desired domain, in this case the TA domain. Do so by dragging the "TA" row from the template rows to the bottom of the table. This e.g. results in:

SC	STUDYID	DOMAIN	USUBJID	SC.SCSEQ
SS	STUDYID	DOMAIN	USUBJID	SS.SSSEQ
TU	STUDYID	DOMAIN	USUBJID	TU.TUSEQ
TR	STUDYID	DOMAIN	USUBJID	TR.TRSEQ
RS	STUDYID	DOMAIN	USUBJID	RS.RSSEQ
VS	STUDYID	DOMAIN	USUBJID	VS.VSSEQ
FA	STUDYID	DOMAIN	USUBJID	FA.FASEQ
SR	STUDYID	DOMAIN	USUBJID	SR.SRSEQ
RELREC	STUDYID	RDOMAIN	USUBJID	IDVAR
SUPPQUAL	STUDYID	RDOMAIN	USUBJID	IDVAR
CES:TA	STUDYID	DOMAIN	TA.ARMCD	TA.ARM

This ensures that the metadata for your TA dataset will be included in the define.xml file. You will also be able to use the trial design dataset editor without having dragged-and-dropped the TA row

¹ At this moment, using SAS Transport 5 format is not foreseen. One can however transform Dataset-XML files into SAS Transport 5 files using other tools. See <http://wiki.cdisc.org/display/PUB/CDISC+Dataset-XML+Resources> for a list of such tools.

following start window is displayed:

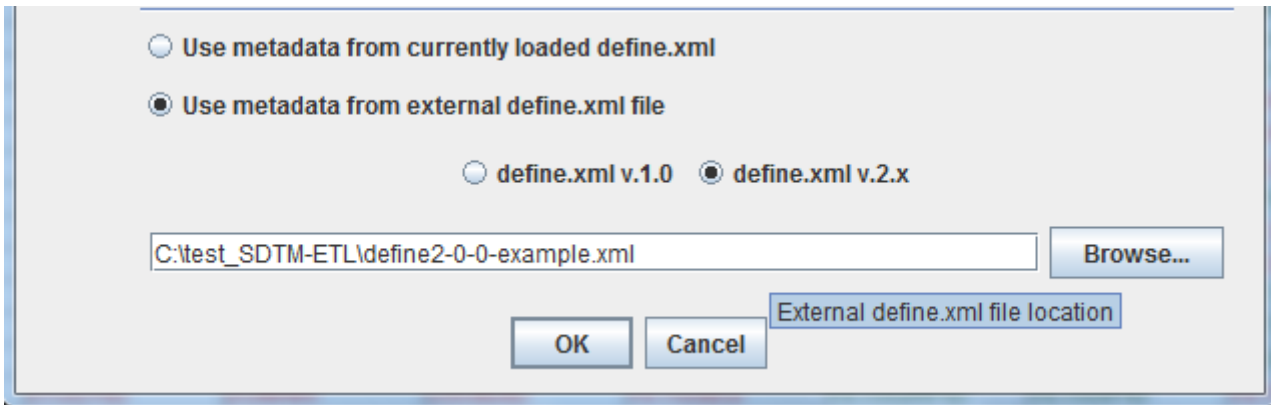


In case the software was started from within SDTM-ETL, the radiobutton "Use metadata from currently loaded define.xml" is preselected. This means that the define.xml from the SDTM-ETL will be used (in its current state). One can however then also choose to choose another define.xml file containing the metadata for the trial design dataset(s).

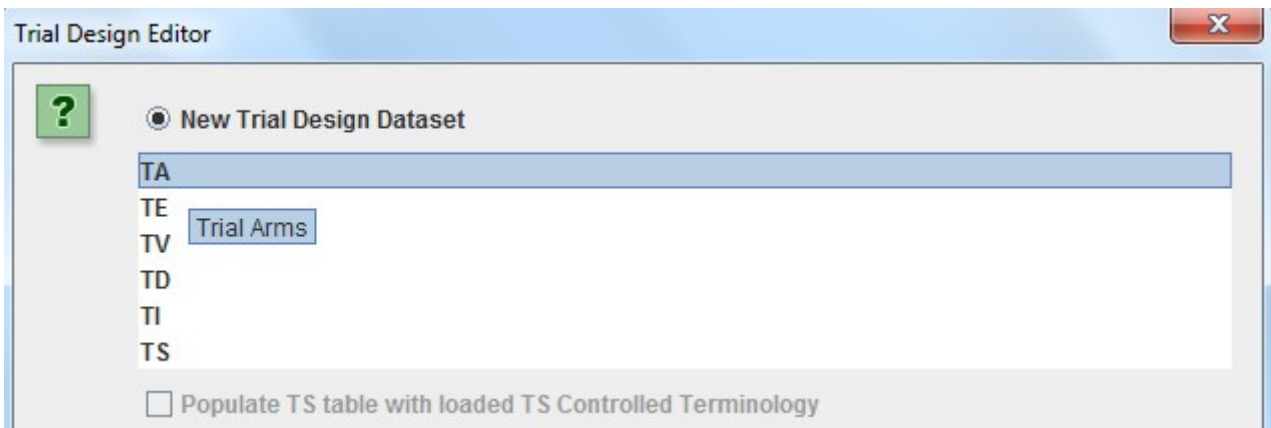
In case the software was started in standalone, the radiobutton "Use metadata from external define.xml file" is preselected, and the radiobutton "Use metadata from currently loaded define.xml" is disabled. So in the latter case, it is always necessary to provide a define.xml file containing the trial design metadata.

In this tutorial we will continue with the option "Use metadata from external define.xml file".

First select the correct version of define.xml that you are working with. This is important as otherwise the file will not be parsed. In our case, we use a define.xml v.2.0 file. Select it using the "Browse" button on the right lower corner of the window. For example:



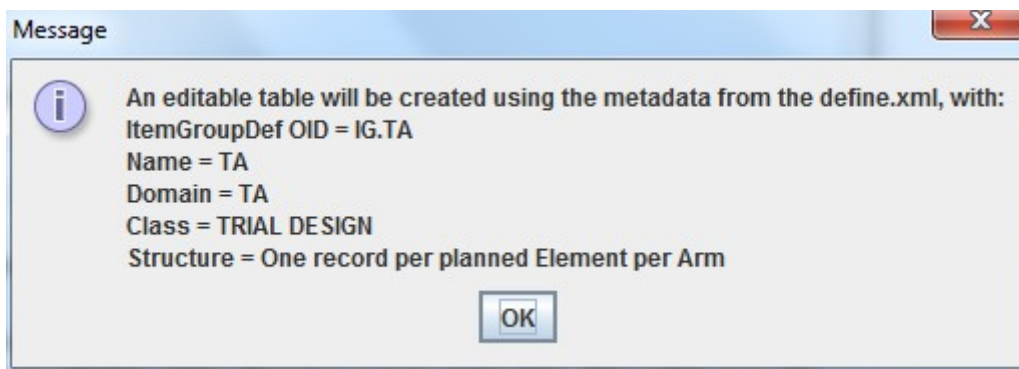
Now you need to decide whether you want to create a new trial design dataset (from scratch) or that you want to work on an already existing trial design dataset in Dataset-XML format. We will first work with the case that one wants to start a completely new trial design dataset. In order to do so, select the radiobutton "New Trial Design Dataset" and select a trial design domain from the list. For example:



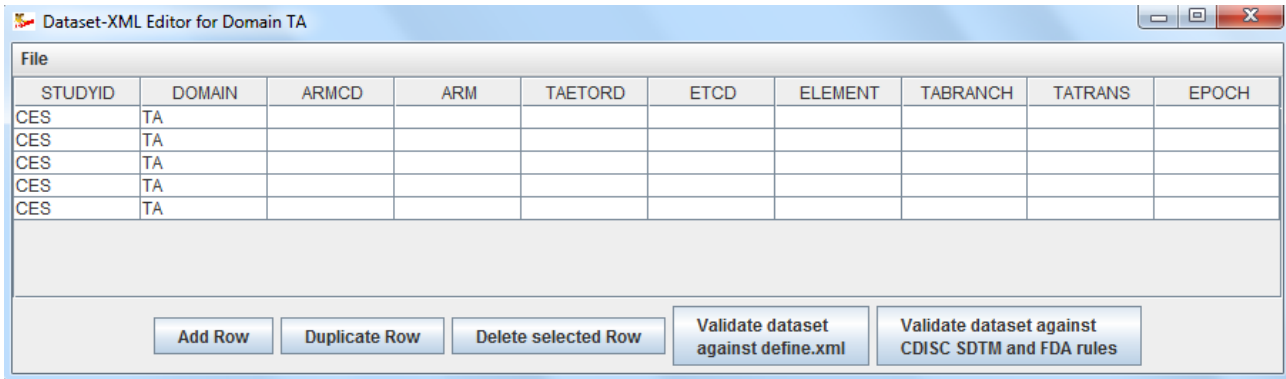
The tooltip on the selection shows the full name of the dataset.

The use of the checkbox "Populate TS table ..." will be explained later. It is disabled for TA. Now that a trial design domain and a define.xml containing the metadata has been selected, use the "OK" button to start generating an editable table allowing you to enter data. This table uses the metadata from the define.xml.

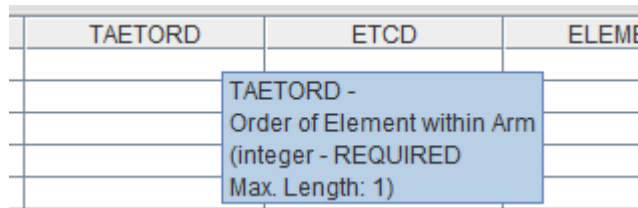
First a dialog is shown containing the most important information:



Click "OK" to continue. The editable table is displayed:

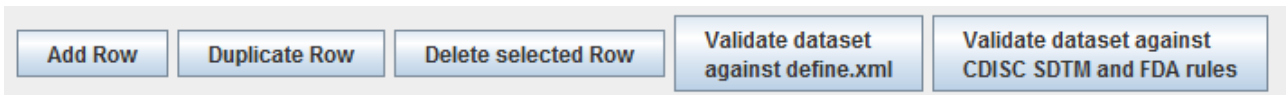


Hovering the mouse over a column header displays some of the most important metadata:

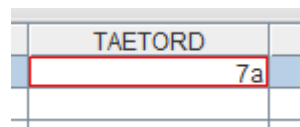


(in case you see things that you don't like, not a bad idea to quit and correct your define.xml ...)

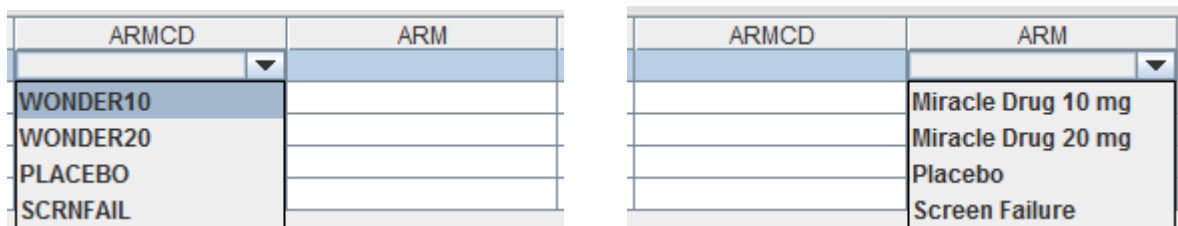
Near the bottom, some buttons are provided to add or delete rows, duplicate a row, start a simple validation against the define.xml, and start validation against CDISC SDTM and FDA SDTM validation rules:



The table editor is pretty "idiot proof". For example, if you try to enter text in a cell for which the datatype is "integer", the cell will be marked by a red border, and you will not be able to move to another cell before the error is corrected:

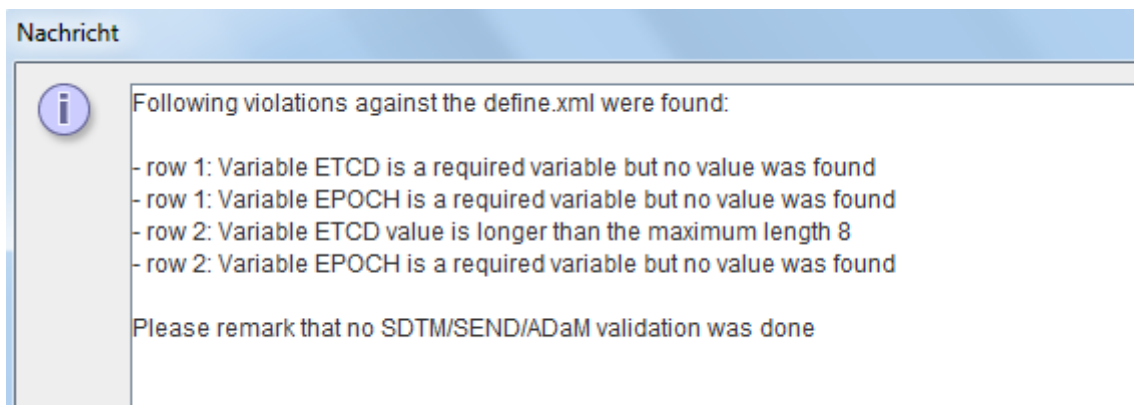


In case a codelist is given for a specific variable, you cannot enter any text, but you need to select a value from the combobox that is shown when trying to edit the cell. For example:



It is also always advised to validate the table against the define.xml, using the button "Validate Dataset". A list is then show with information about violations in the table against the metadata in

the define.xml. For example:

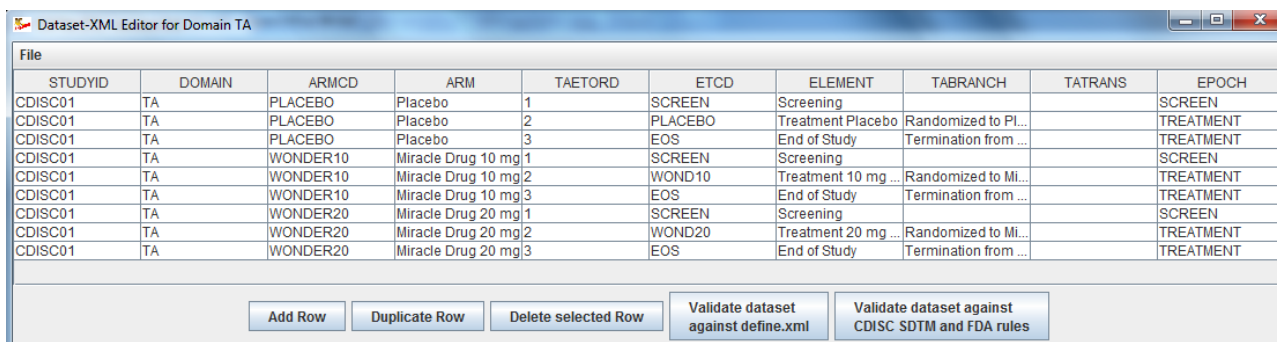


Also the cells that violate the definitions in the define.xml will then be colored red.

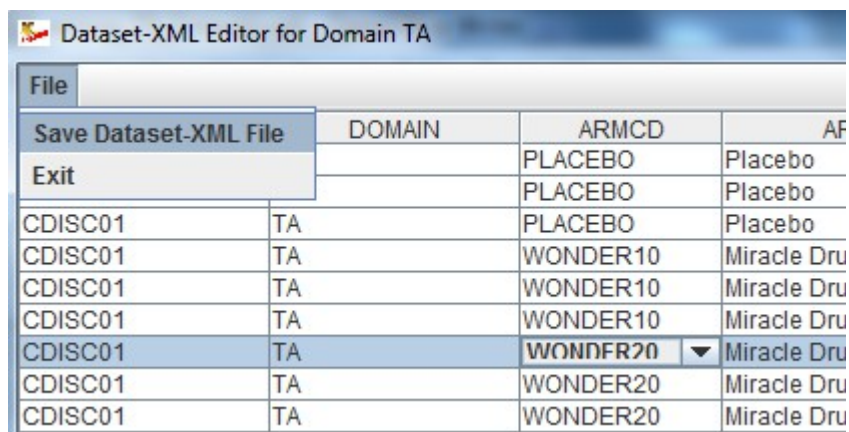
It is important to note that this is **NOT** an SDTM validation², only a validation against the define.xml. The SDTM validation will be explained later in the section "SDTM Validation".

To add a new row, use the button "Add Row", to delete a selected row, use the button "Delete Row". One can also use the button "Duplicate Row", which can speed up filling up - but be aware that it is your own responsibility to ensure that each row is unique as described by the domain rules.

The following figure shows a picture of a complete table for TA (trial arms) that is compliant towards its define.xml:

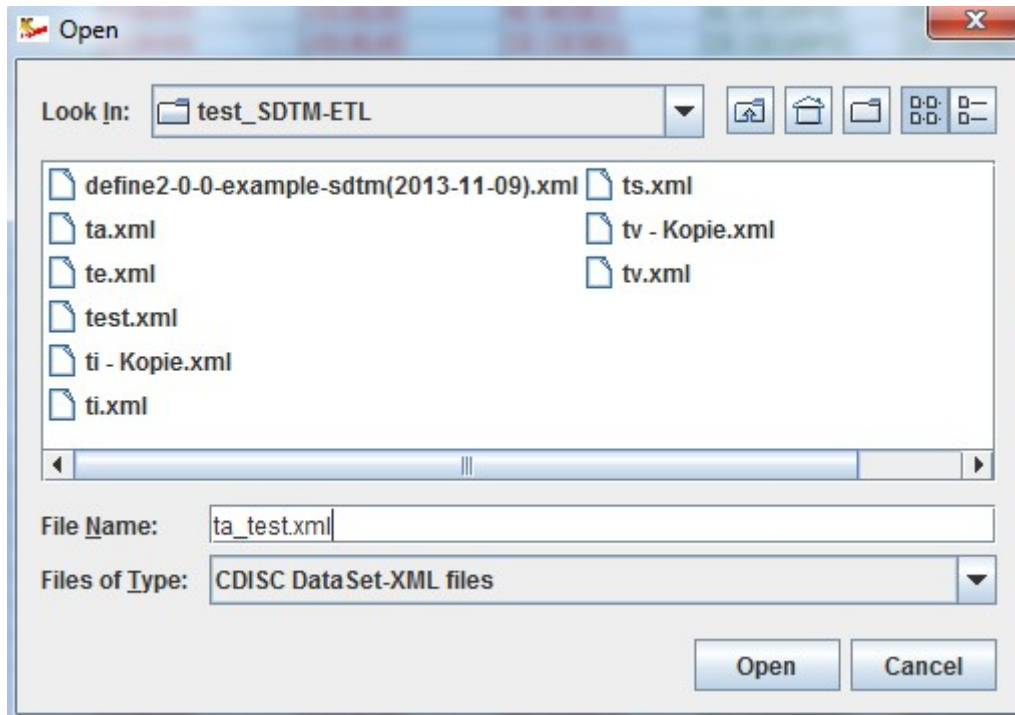


Finally, you can save the table to file in Dataset-XML format, using the menu "File - Save Dataset-XML File":



² Use other tools for SDTM validation, such as [OpenCDISC](#).

You will be prompted to provide a location to store the file. For example:

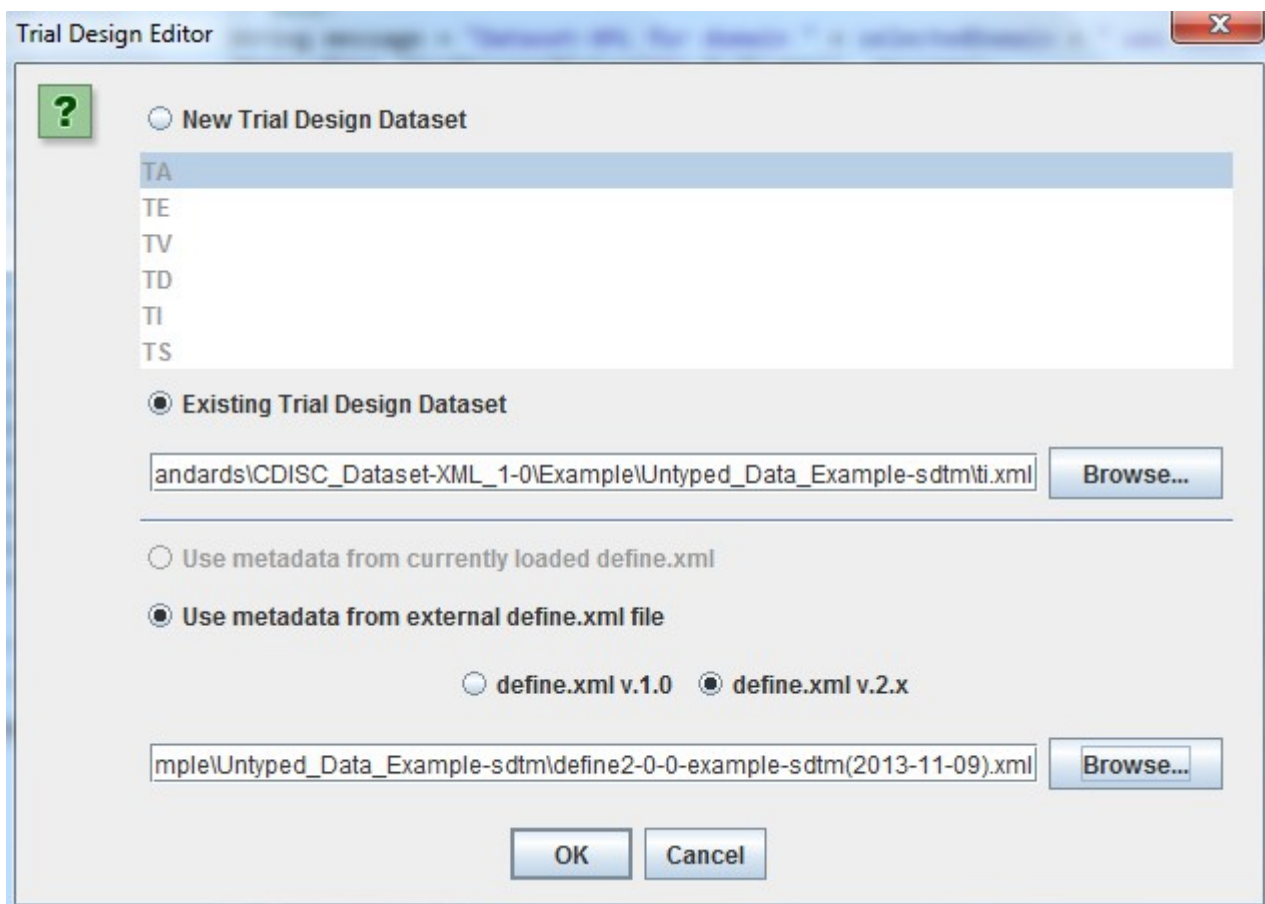


If the file already exists, the system than will ask for a confirmation to overwrite the file.

Editing an already existing trial design Dataset-XML file

To work on an existing trial design dataset in Dataset-XML format, use the radiobutton "Existing Trial Design Dataset", then use the "Browse" button to select a dataset.

Ensure that you also load a proper define.xml file. For example:



Clicking "OK" then loads the ti.xml file and the define.xml file containing the metadata for the TI dataset:

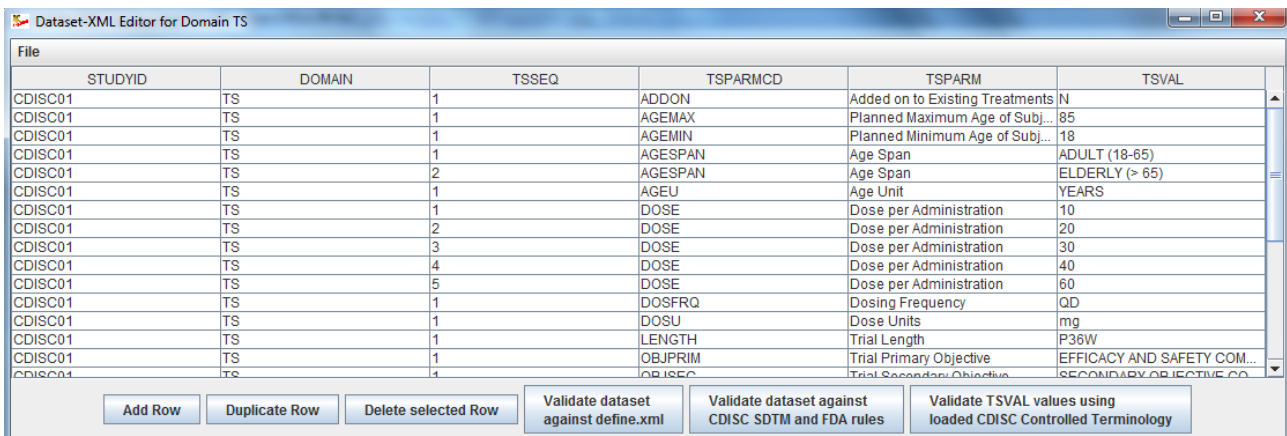
STUDYID	DOMAIN	IETESTCD	IETEST	IECAT
CDISC01	TI	EXCL01	Is pregnant, nursing, or planning to b...	EXCLUSION
CDISC01	TI	EXCL02	Is unable or unwilling to undergo mul...	EXCLUSION
CDISC01	TI	EXCL03	Is known to have had a substance ab...	EXCLUSION
CDISC01	TI	INCL01	Is age 18 - 85	INCLUSION
CDISC01	TI	INCL02	Has Xyz disease of at least 10 weeks...	INCLUSION
CDISC01	TI	INCL03	Did not respond to a standard course...	INCLUSION

One can then start editing the information, add or remove rows, and validate the dataset against the define.xml that was selected or against the CDISC or FDA SDTM rules . Do not forget to use the menu "File - Save Dataset-XML File" to save the results to disc.

To exit, use the menu "File - Exit".

Working with the TS (Trial Summary) Dataset

The "Trial Summary" (TS) dataset is somewhat special as it essentially is a "Entity - Attribute - Value" table³. An example is shown below:



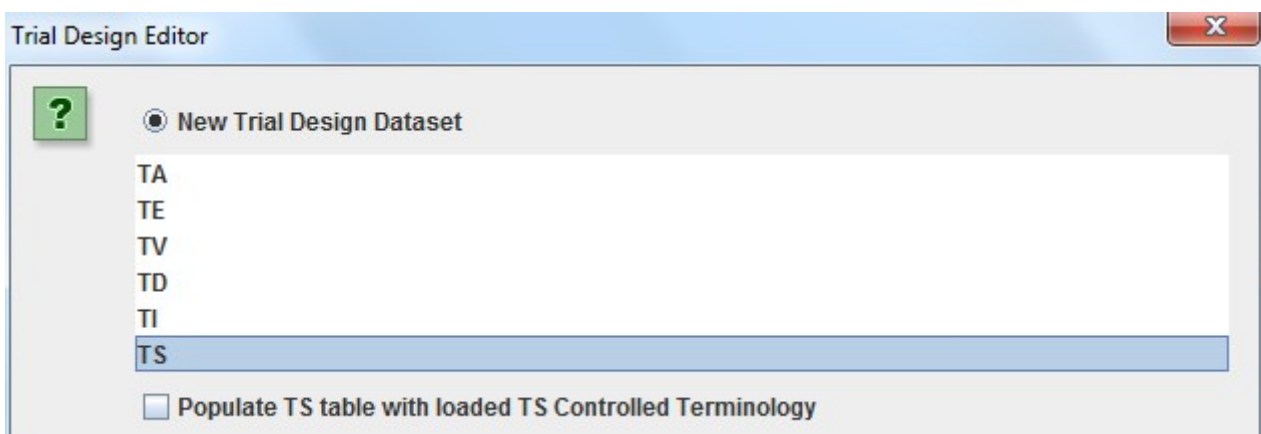
STUDYID	DOMAIN	TSSEQ	TSPARMCD	TSPARM	TSVAL
CDISC01	TS	1	ADDON	Added on to Existing Treatments	N
CDISC01	TS	1	AGEMAX	Planned Maximum Age of Subj...	85
CDISC01	TS	1	AGEMIN	Planned Minimum Age of Subj...	18
CDISC01	TS	1	AGESPAN	Age Span	ADULT (18-65)
CDISC01	TS	2	AGESPAN	Age Span	ELDERLY (> 65)
CDISC01	TS	1	AGEU	Age Unit	YEARS
CDISC01	TS	1	DOSE	Dose per Administration	10
CDISC01	TS	2	DOSE	Dose per Administration	20
CDISC01	TS	3	DOSE	Dose per Administration	30
CDISC01	TS	4	DOSE	Dose per Administration	40
CDISC01	TS	5	DOSE	Dose per Administration	60
CDISC01	TS	1	DOSFRQ	Dosing Frequency	QD
CDISC01	TS	1	DOSU	Dose Units	mg
CDISC01	TS	1	LENGTH	Trial Length	P36W
CDISC01	TS	1	OBJPRIM	Trial Primary Objective	EFFICACY AND SAFETY COM...
CDISC01	TS	1	OBJSEC	Trial Secondary Objective	SECONDARY OBJECTIVE CO...

The "entity" is TSPARMCD ("trial summary parameter code") defining "what it is about". The "value" is TSVAL ("trial summary value"). All other columns contain "attributes".

TSPARMCD is coded, i.e. a list of allowed values for TSPARMCD has been published by CDISC. The same applies to TSPARM ("trial summary parameter name"). Furthermore, some of the TSVAL values are coded, depending on the value for TSPARMCD. Also the expected data types can be different

For example, when TSPARMCD is AGEMAX, then a non-negative integer is expected. However, for TSPARMCD=TBLIND ("Blinding Type"), the only allowed values are: "SINGLE BLIND", "DOUBLE BLIND" and "OPEN LABEL".

Let us start a new TS dataset. To do so, use the menu "Edit - Trial Design Dataset" and select "New Trial Design Dataset", and then select "TS".



If the checkbox "Populate TS table with loaded TS Controlled Terminology" is not checked, an empty table will be generated, to which you will need to add TSPARMCD and TSPARM values yourself. As these are coded, this can easily be done using the dropdown. For example:

³ See e.g.: <http://ycmi.med.yale.edu/nadkarni/Introduction%20to%20EAV%20systems.htm>

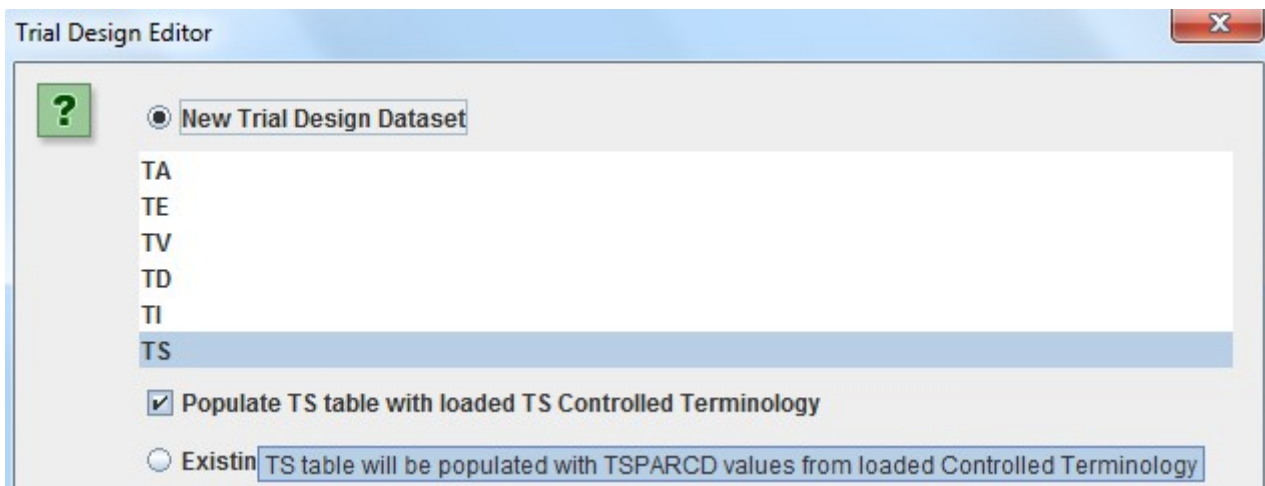
TSGRPID	TSPARMCD	TSPARM	TSVAL	TSVALNF
	ADAPT			
	ACTSUB			
	ADAPT			
	ADDON			
	AGEMAX			
	AGEMIN			
	COMPTRT			
	CRMDUR			
	CURTRT			

The same applies for TSPARM.

This is a suitable way to produce a TS table and dataset where the number of parameters is relative low.

The second strategy is to load all possible values for TSPARMCD, and generate a row for each of them, and then remove the ones that are not needed. This is usually the best strategy when you want to be sure that all by the FDA required "trial design parameters" are provided.

In order to do so, ensure that the checkbox "" is selected when starting to generate the TS table:



This will then result in the following table:

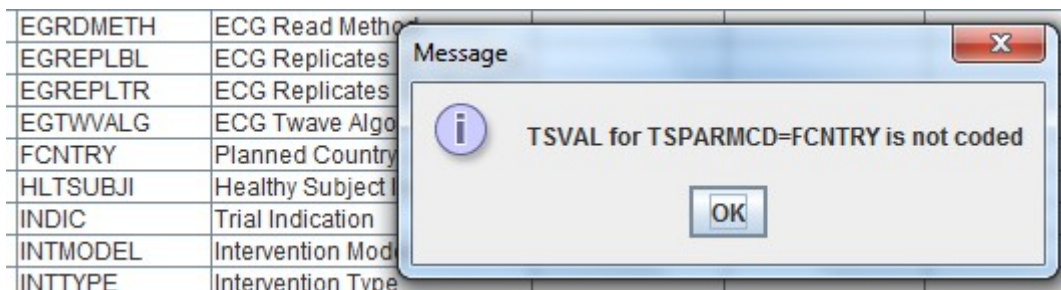
STUDYID	DOMAIN	TSSEQ	TSGRPID	TSPARMCD	TSPARM	TSVAL	TSVALNF	TSVALCD	TSVCDREF	TSVCDVER
CES	TS			ACTSUB	Actual Number of Subjects					
CES	TS			ADAPT	Adaptive Design					
CES	TS			ADDON	Added on to Existing Treat...					
CES	TS			AGEMAX	Planned Maximum Age of ...					
CES	TS			AGEMIN	Planned Minimum Age of S...					
CES	TS			COMPTRT	Comparative Treatment Na...					
CES	TS			CRMDUR	Confirmed Response Mini...					
CES	TS			CURTRT	Current Therapy or Treatm...					
CES	TS			DCUTDESC	Data Cutoff Description					
CES	TS			DCUTDTC	Data Cutoff Date					
CES	TS			DOSE	Dose per Administration					
CES	TS			DOSFRQ	Dosing Frequency					
CES	TS			DOSU	Dose Units					
CES	TS			DXCRIT	Diagnostic Criteria					
CES	TS			EGBLIND	ECG Reading Blinded					
CES	TS			EGCTMON	ECG Continuous Monitoring					
CES	TS			EGLEADPR	ECG Planned Primary Lead					
CES	TS			EGLEADSM	ECG Used Same Lead					
CES	TS			EGRDMETH	ECG Read Method					
CES	TS			EGREPLBL	ECG Replicates at Baseline					
CES	TS			EGREPLTR	ECG Replicates On-Treat...					
CES	TS			EGTWALG	ECG Twave Algorithm					
CES	TS			FCNTRY	Planned Country of Investi...					
CES	TS			HLTSUBJI	Healthy Subject Indicator					
CES	TS			INDIC	Trial Indication					

Buttons at the bottom: Add Row, Duplicate Row, Delete selected Row, Validate dataset against define.xml, Validate dataset against CDISC SDTM and FDA rules, Validate TSVAL values using loaded CDISC Controlled Terminology

Also notice the extra button that has appeared at the bottom at the right stating "Validate TSVAL values using loaded CDISC controlled terminology". Its function will be explained further on.

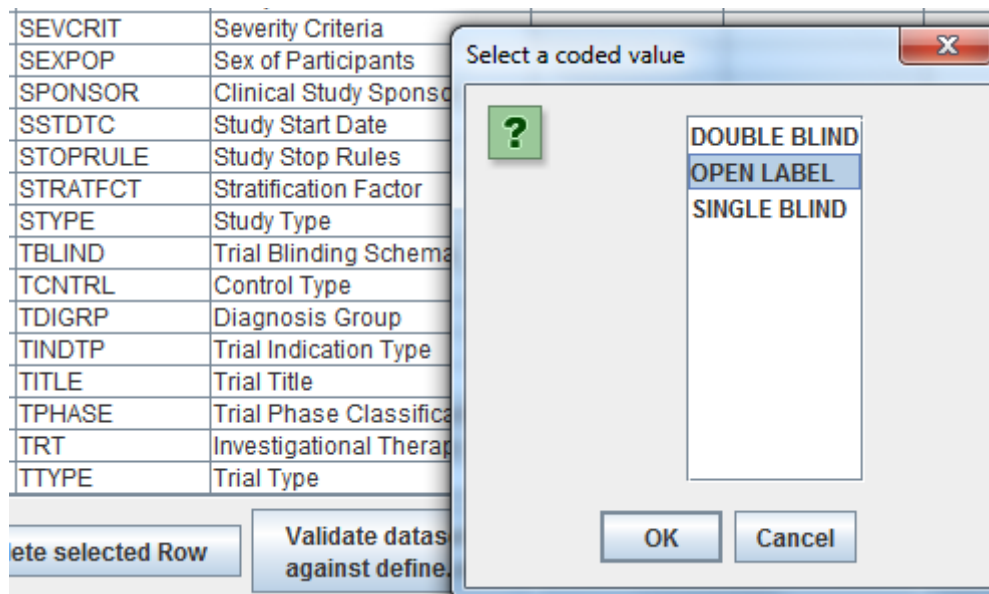
One can now start adding information using the buttons "Add Row" and "Duplicate Row" (the latter for trial design parameters for which there is more than one value - e.g. "FCNTRY" ("country of planned investigation")).

Some of the TSVAL values are coded. To find out, right-click the TSVAL cell. If the TSVAL for the TSPARMCD value is not coded, a message will displayed:



as no controlled terminology for "FCNTRY" has been defined by CDISC.

If however TSVAL is coded, for example for TBLIND ("Trial Blinding Scheme), the following dialog is popping up:



You can then select one of the values, and when clicking "OK", the value will be added to the TSVVAL cell. Even then, you can still edit that cell and put another value (discouraged).

STYPE	Study Type	
TBLIND	Trial Blinding Schema	OPEN LABEL
TCNTRL	Control Type	

Please note that there are many "hidden" rules for TS for which no machine-executable rules have been defined yet, so looking into the SDTM-IG when filling the TS table is not a bad idea.

It is also always a good idea to do use the button "Validate TSVVAL values using loaded CDISC controlled terminology". This will validate all TSVVAL values against the codelists provided by CDISC. It also validates whether the value for TSPARMCD and TSPARM match. For example:

TSPARMCD	TSPARM	TSVAL	TSVALNF	TSVALCD	TSVCDREF	TSVCDV
TTYPE	Trial Type	blinded				
TEST	Trial Type					

Value 'blinded' is an invalid value for TSPARMCD TTYPE
 Following values (case sensitive) are allowed:
 (right-click for adding a valid coded value)

- BIO-AVAILABILITY
- BIO-EQUIVALENCE
- EFFICACY
- FOOD EFFECT
- IMMUNOGENICITY
- PHARMACODYNAMIC
- PHARMACOECONOMIC
- PHARMACOGENOMIC
- PHARMACOKINETIC
- SAFETY
- TOLERABILITY

and:

TSPARMCD	TSPARM	TSVAL	TSVALNF	TSVALCD
TTYPE	Trial Type	blinded		
TEST	Trial Type			

Mismatch between TSPARMCD and TSPARM

OR:

TSVAL	TSVALNF	TSVALCD	TSVCDR
blinded			

Either one of TSVAL or TSVALNF must be populated

Also here, do not forget to do a validation against the define.xml to check for completeness.

SDTM Validation

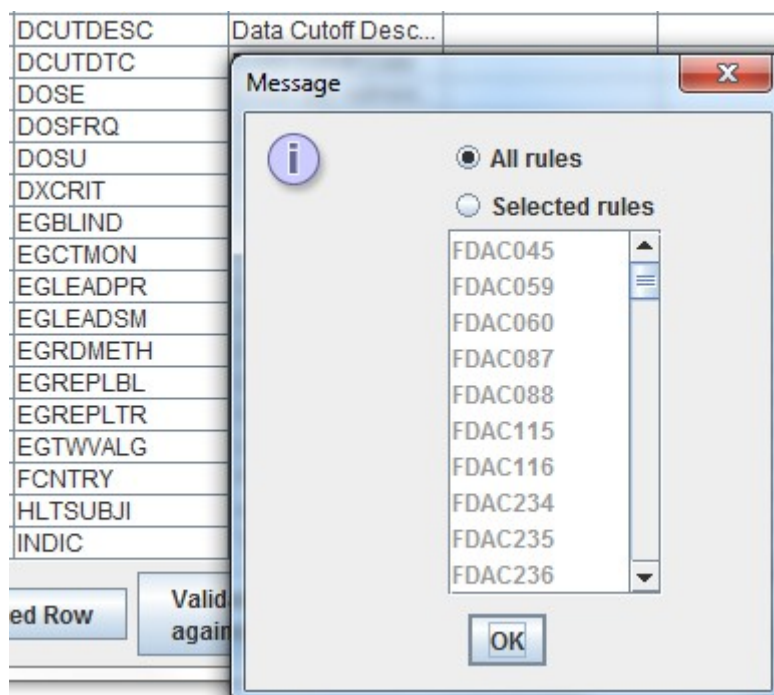
We have already seen the function "Validate dataset against define.xml", which does a simple validation on data types and whether a variable is a required variable or not.

However, trial design datasets have more complex rules that cannot be represented by define.xml alone. Even a simple rule that either TSVAl or TSVAlNF must be populated cannot be expressed by define.xml.

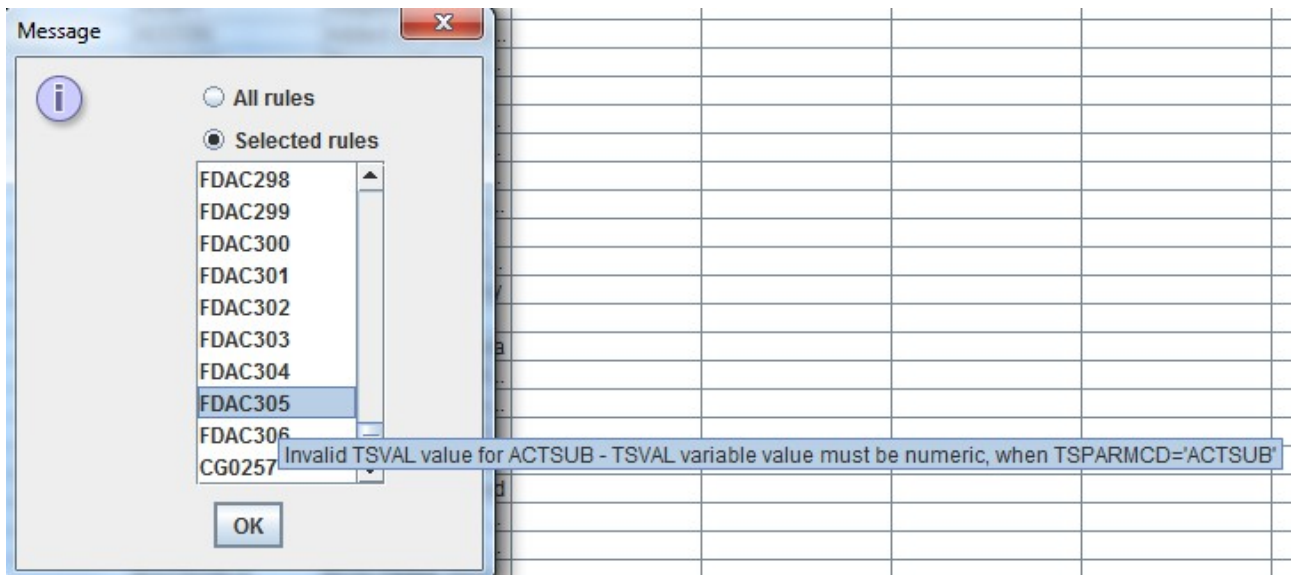
Recently, and in cooperation with CDISC, we started developing the [FDA SDTM validation rules](#) as open, human readable, machine-executable rules. They can be found at: http://xml4pharmaserver.com/WebServices/XQueryRules_webservices.html. These rules are later envisaged to become a part of [CDISC eSHARE](#).

The set of rules is permanently be extended in cooperation with the "CDISC SDTM Validation Subteam".

The software allows to use these rules for validation of the TS datasets. When the button "Validate dataset against CDISC SDTM and FDA rules" is clicked, the software looks up all the rules for the specific domain and presents these for selection to the user. For example for TS (Trial Summary):



One can either decide to run all rules applicable to TS, or to do a selection. Each rule has a tooltip on it, so that the rule is explained when one hovers the mouse over the rule identifier. For example:



When the OK button is then clicked, the validation engine starts, and each of the rules is executed subsequently on the dataset. After a few seconds, a report is then displayed. For example:

Validation Messages for DomainTS			
Validation messages for XQuery validation for domain			
TS			
Date and time: 2015-09-13T14:11:30			
Rule	Severity	Record number	Message
FDAC237	error		Invalid (non-ISO8601) TSVAl value=85 for AGEMAX in dataset=temp_TS.xml
FDAC234	error		Invalid (non-ISO8601) TSVAl value=18 for AGEMIN in dataset=temp_TS.xml
FDAC243	error		Invalid (non-ISO8601) TSVAl value=P36W for LENGTH in dataset=temp_TS.xml
FDAC252	error		Missing STOPRULE Trial Summary Parameter in dataset temp_TS.xml
FDAC255	error		Missing TDIGRP Trial Summary Parameter in dataset temp_TS.xml
FDAC260	error		Missing CURTRT Trial Summary Parameter in dataset temp_TS.xml
FDAC276	error		Missing REGID Trial Summary Parameter
FDAC277	error	27	Invalid TSVAl value for TRT in dataset temp_TS.xml: TSVAl for TRT record must be a valid preferred term from FDA Substance Registration System (SRS): value 'TEST PRODUCT XXX' is not a valid preferred term identifier from SRS
FDAC280	error	27	Invalid TSVCDREF value for TRT: TSVCDREF variable value must be 'UNII', when TSPARMCD='TRT', value found is "

Please note that some of these rules use a RESTful webservice provided by the US "National Library of Medicine". So to be able to use these, your computer will need an internet connection. If no internet connection is available, the engine will simply skip the rule.

A great advantage of this new mechanism for performing SDTM validation is that the rules are "really open": anyone can inspect them. This is different from other validation tools, where the rules are hidden in the software (so intransparent), so that one cannot inspect how they have been implemented.

Furthermore, it allows organizations to develop company-internal validation rules. These can then simply be added to the software by copying them into one of the XML files of the directory "Validation_Rules_XQuery"⁴.

[Rules for other domains](#) than the trial design datasets have already been developed, but they have not been implemented in the SDTM-ETL software. This will be started in the near future.

⁴ There is no additional license necessary to add company-specific rules to the software.