SDTM-ETL 4.2: Validating SDTM/SEND datasets Using CORE (CDISC Open Rules Engine)

Author: Jozef Aerts, XML4Pharma

Last update: **2023-04-10**

Table of Contents

Introduction

CDISC CORE (CDISC Open Rules Engine) is currently revolutionizing the submission validation world.
Until recently, sponsors were dependent on software from a vendor who had a quasi-monopoly, but delivered validation software of pretty bad quality that was not very user-friendly and that produced large amounts of "false positives", for which it was notorious.
CORE has been developed by CDISC by CDISC coworkers and a large number (>100) of volunteers, especially from the standards developments teams, i.e. persons who know the standards very well, and do know the validation rules very well, and understand how to implement them. This in contrast to the usual vendor, who mostly implemented their own interpretation of the CDISC and FDA rules.
The great advantage of CORE is that all the rules are completely transparent, and can be inspected by the users. So in contrast to the old validation software from the mentioned vendor, it is not "black box" at all.
CORE also is "open source", i.e. everyone can download it from the project's GitHub site, and can implement it or integrate it in their own tools. We therefore expect that CORE will be implemented in many applications and by a good number of vendors: these will then compete on software quality, user friendliness, cost of ownership etc.. However, all these vendors will use the same basis validation engine, so that the validation results will essentially be independent of from which vendor one uses the validation software.

As one of the first vendors, CORE has now also be integrated in our SDTM-ETL software.
At the moment of writing (April 2023), CORE is already in an advanced state for the CDISC SDTM rules, less for SEND, but also still in development, and new versions released and new rules added almost every week.
Therefore, we have implemented CORE in SDTM-ETL in such a way that new versions of CORE can be installed without needing an SDTM-ETL software update: new versions of CORE that have been tested by us, will be made available to our customers through a website for plugging in into their existing SDTM-ETL installation.

Features of CORE

Besides that all CORE rules are fully transparent, and the CORE engine is free and open source, some features of CORE that are of enormous importance, and a huge improvement over existing validation software are:

- One can select individual rules (one, or a set of) for execution of the validation. With the existing software, it was an "all or nothing", i.e. one could not exclude specific rules for execution. This was especially nerve-wracking during development of the SDTM/SEND datasets: when some datasets were failing, one would get thousands of error messages. With CORE, each individual rule can be switched off or on for execution.
- One can select individual datasets to be involved in the validation, even when one has generated more than one dataset. This allows to concentrate on a single or a few datasets only. With the old software, it was again an "all or nothing". When for example the TS dataset or DM dataset was failing (maybe as it was not developed yet), one would even get "rejection messages", usually causing a lot of panic and/or confusion.
- Some of the rules developed in the past are more like "guidance" or "best practice". So, in some cases, only part of the rule can be made executable. When applicable, CORE always prefers to be rather "under-reporting" than "over-reporting". This means one should not get "false positives" (over-reporting), but that it can happen that an issue is not reported (false negative - under-reporting). CORE has the principle that false positives are worse than false negatives.
- In future, sponsors will be able to extend CORE with their own, internal SDTM/SEND/ADaM rules. This is especially important for quality assurance during the development of the datasets. Exending the validation with own, internal rules was not possible with the old software that is being used by many sponsors, at least not at a reasonable cost

With all these advantages, it is expected that regulatory authorities will soon switch to CORE as their major validation mechanism. CDISC is now already implementing the FDA rules in CORE, and is in close contact with the FDA and discussing shared governance over the FDA validation rules.

CORE Implementation in SDTM-ETL

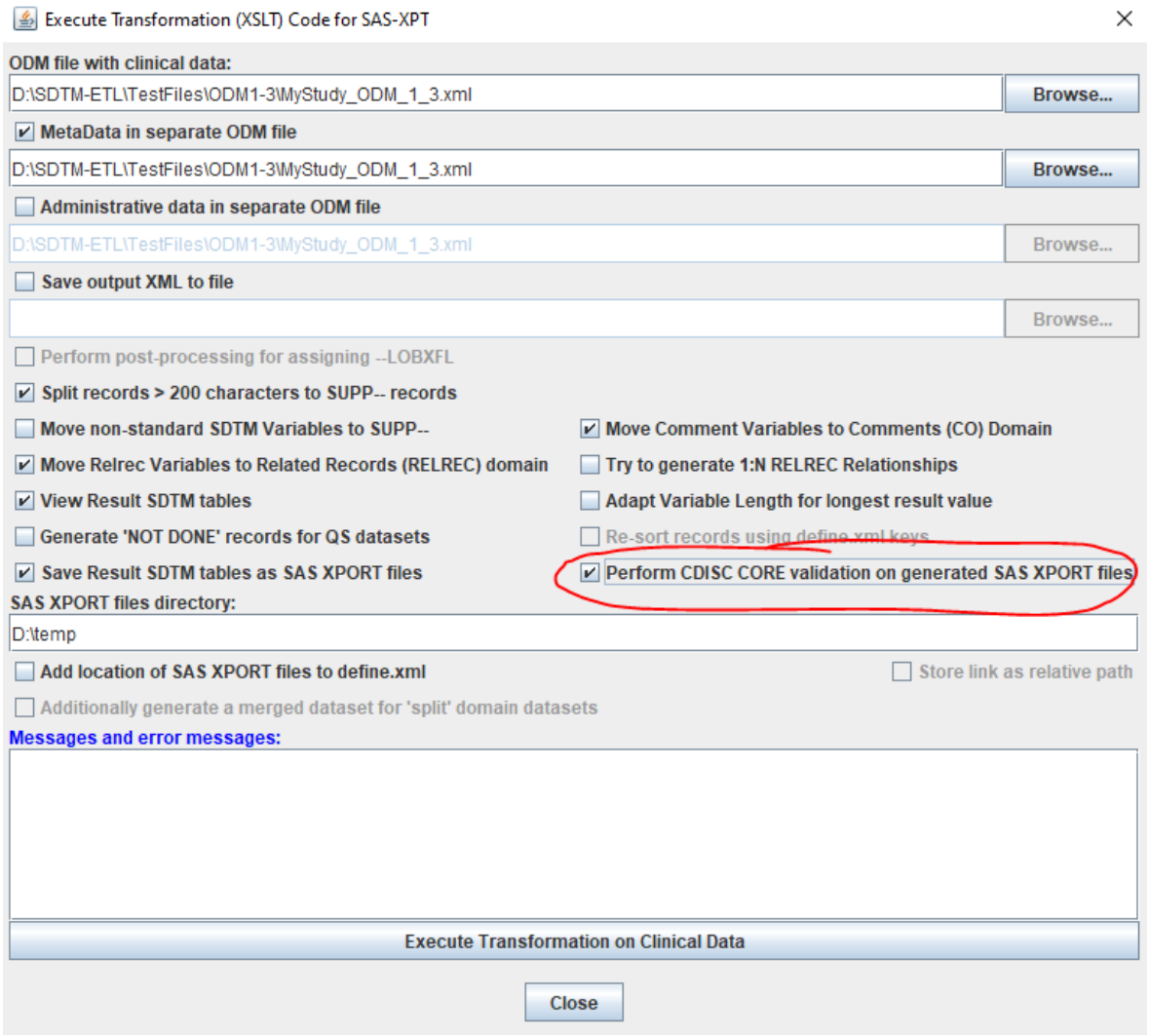
As of version 4.2 of SDTM-ETL, the CORE engine software and files are located in the directory "CDISCCORE". There is also a directory "CDISCCOREFiles" which is used during validation to store a copy of the generated XPT files that need to be validated.

This also means that CORE can be run from CLI (command line) starting from the "CDISCCORE" directory. For information how to run CORE from the command line, see the CDISC-CORE GitHub website explaining this (under "Running a validation").

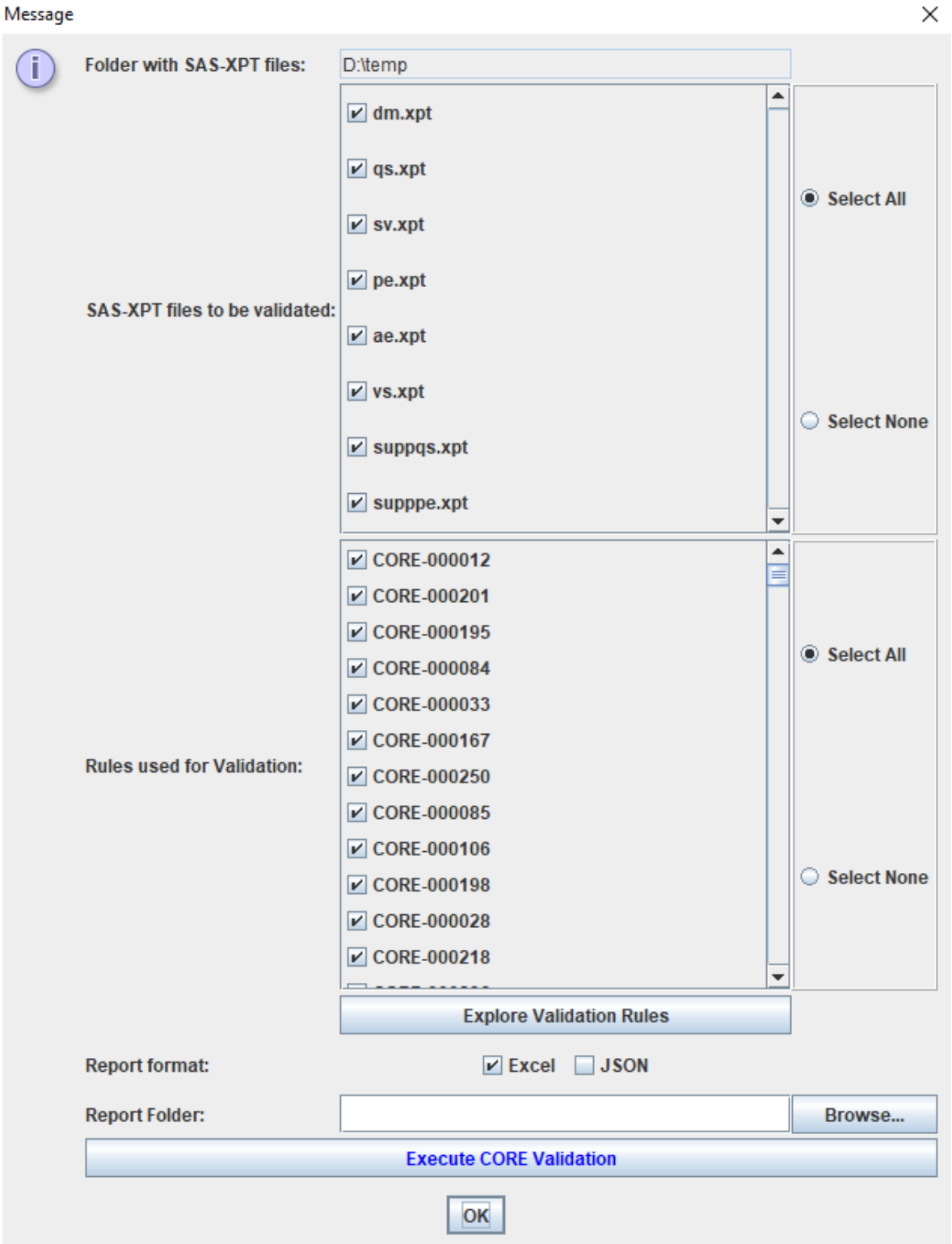
In SDTM-ETL, CORE can be run immediately after the generation of the SDTM/SEND SAS-XPT files. In near future, also an interface with the generation of SDTM or SEND files in Dataset-JSON format will be implemented.

Running CORE in SDTM-ETL

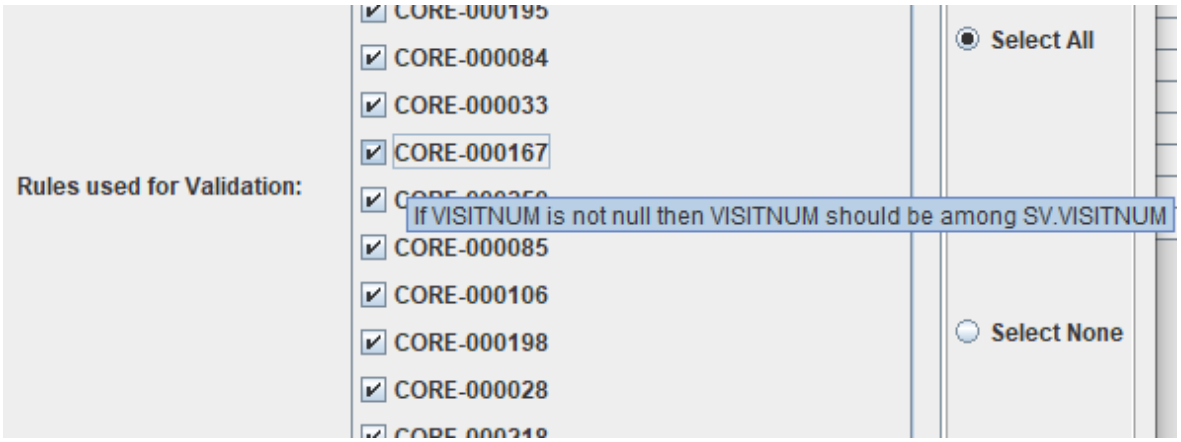
When generating SDTM or SEND files in SDTM-ETL, the following dialog is displayed:



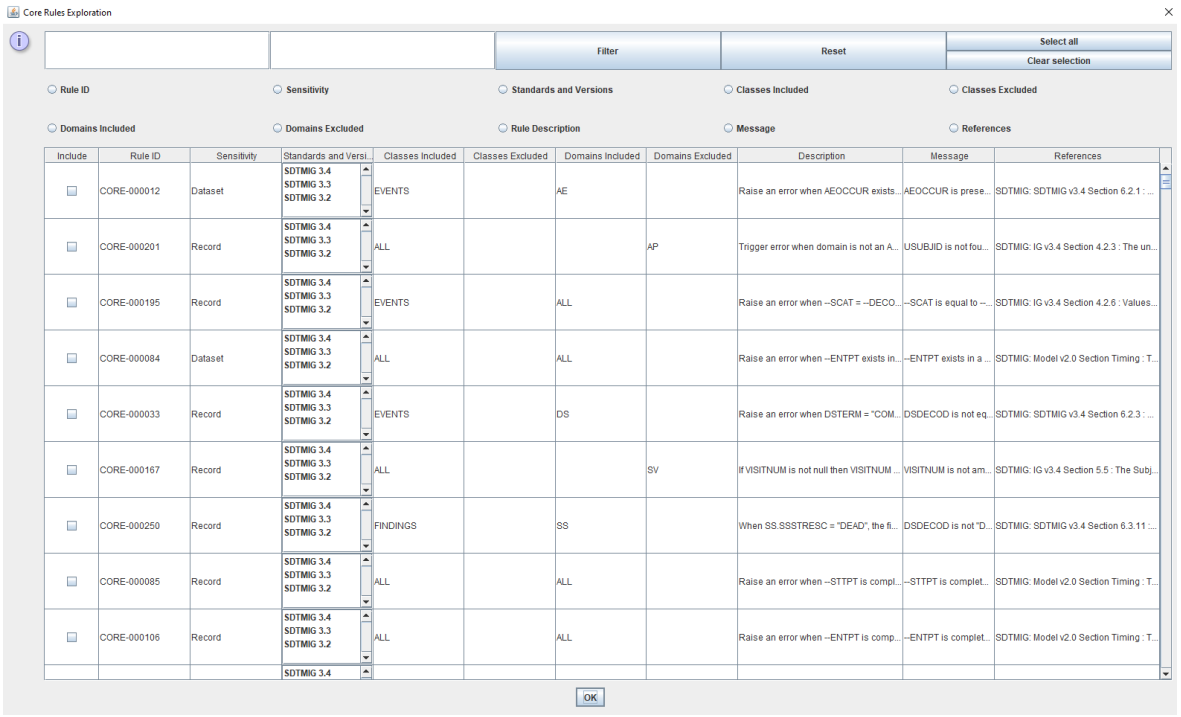
Remark the new checkbox "Perform CDISC CORE validation on generated SAS XPORT files". When it checked, the user will get the opportunity to perform CORE validation on the generated SAS-XPT files.
After generating the SDTM/SEND XPT files, the following dialog will then be shown:



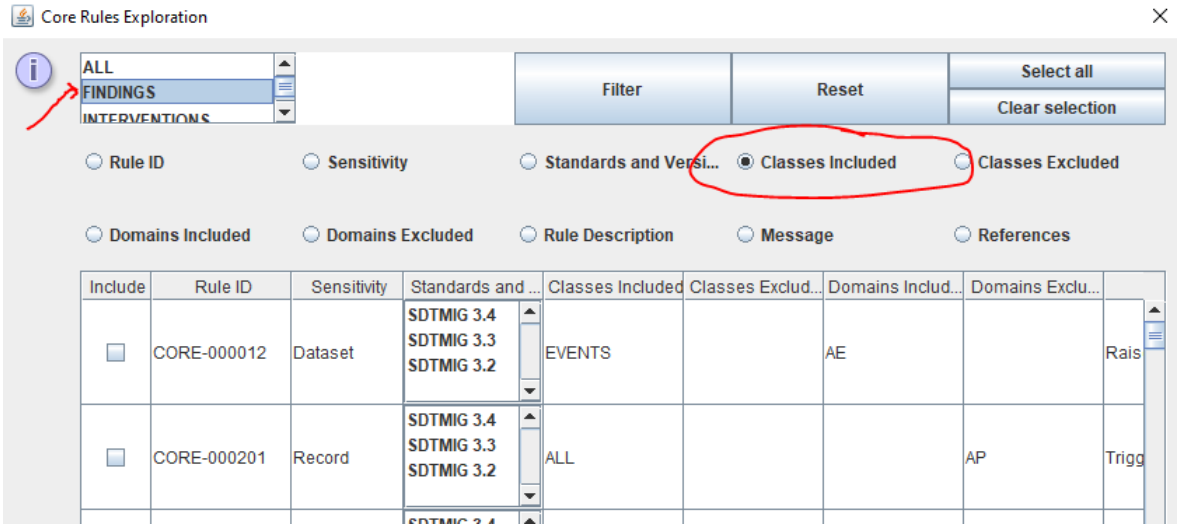
allowing to select which of the generated XPT files must be validated (default: all), and which of the rules must be executed (default: all available).
One can also choose whether the report needs to be generated in Excel format (default) or in JSON format. The latter will become very useful in future for use of the report in other applications, such as "smart" SDTM/SEND viewers.
When one moves the mouse over one of the CORE rules, a tooltip is displayed with some information about the rule:



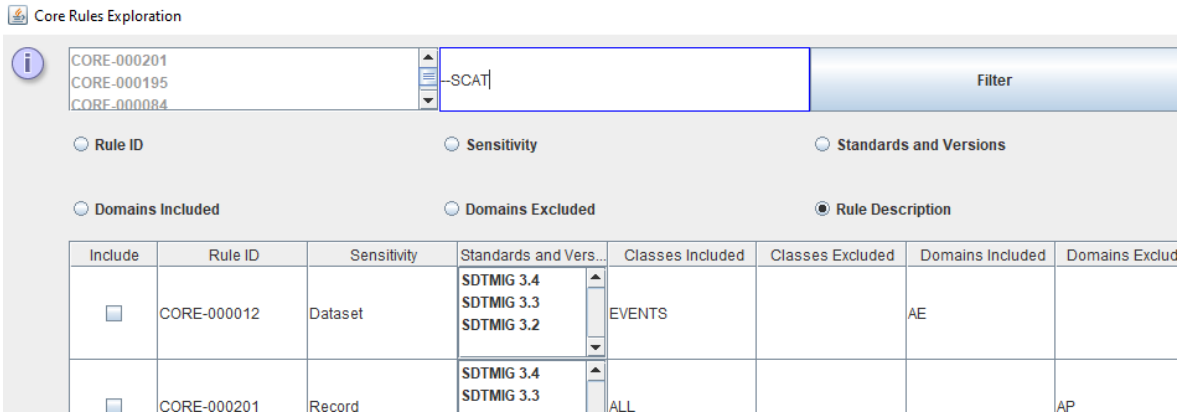
Even better is to use the button "Explore Validation Rules". This leads to a new window allowing to search for and select rules:



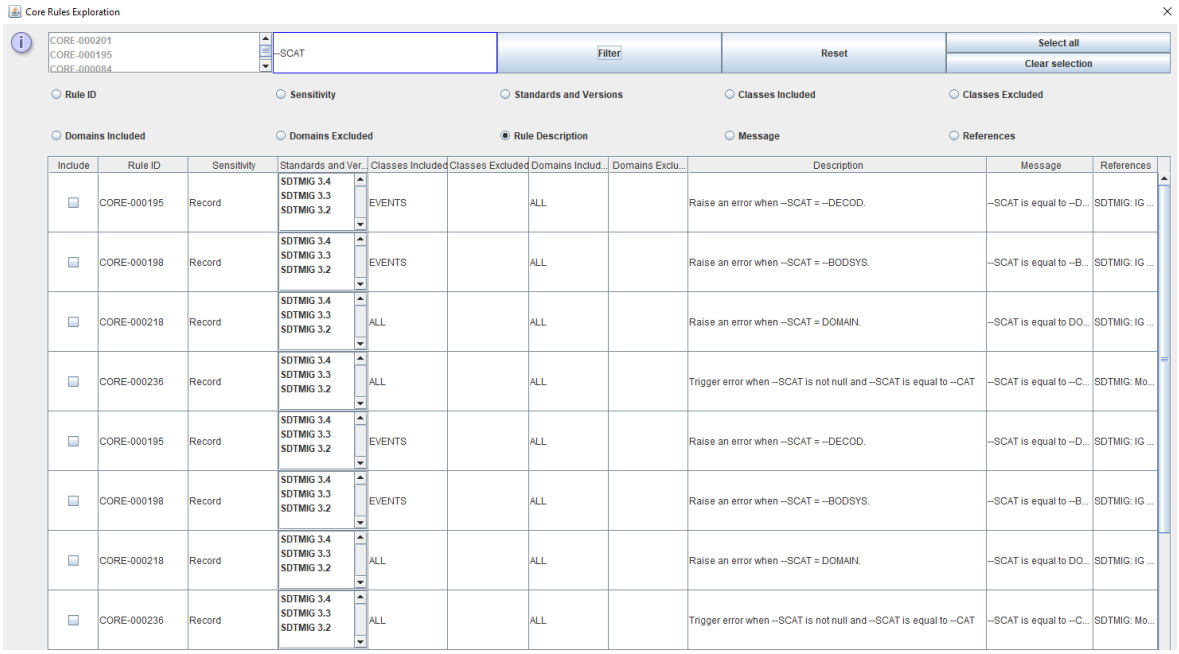
For example, if one only wants to have the rules executed that apply on "Findings" datasets, one first selects the checkbox "Classes Included" and then select "Findings" in the list on the left side (multiple selection allowed):



Like this, one can also include/exclude domains, or filter on a word in the rule description or message. For example, to only select rules about the --SCAT variable:

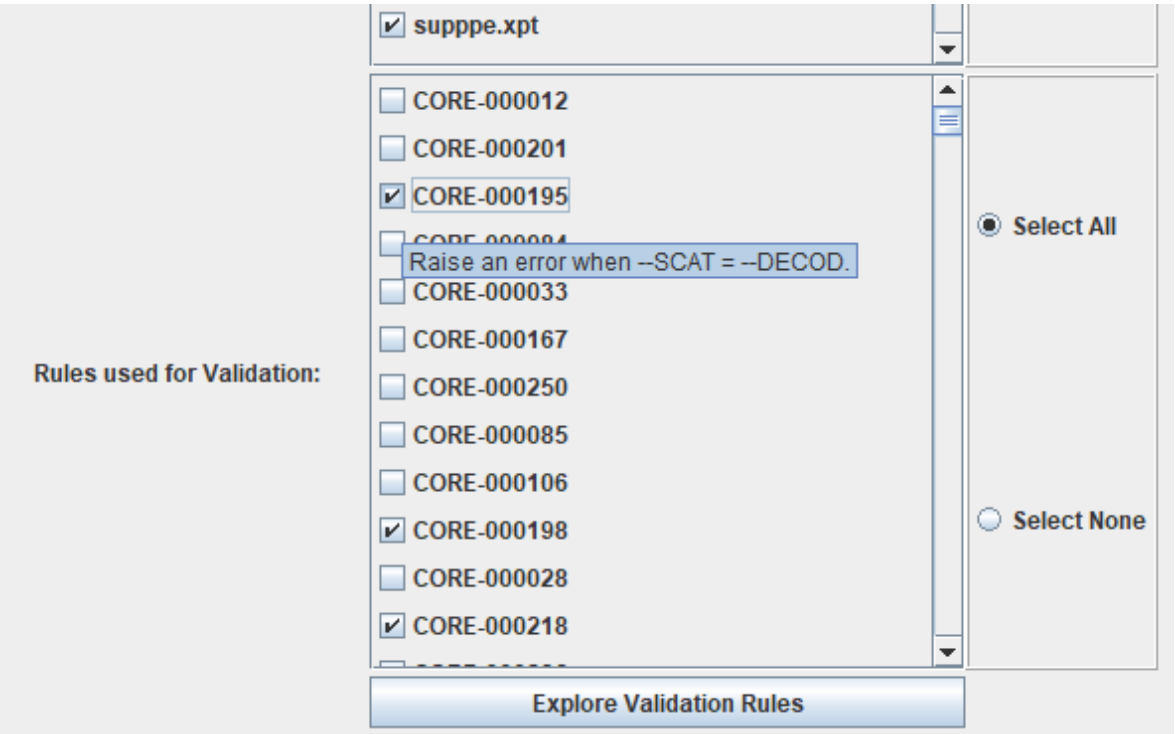


and then clicking the "Filter" button leads to:



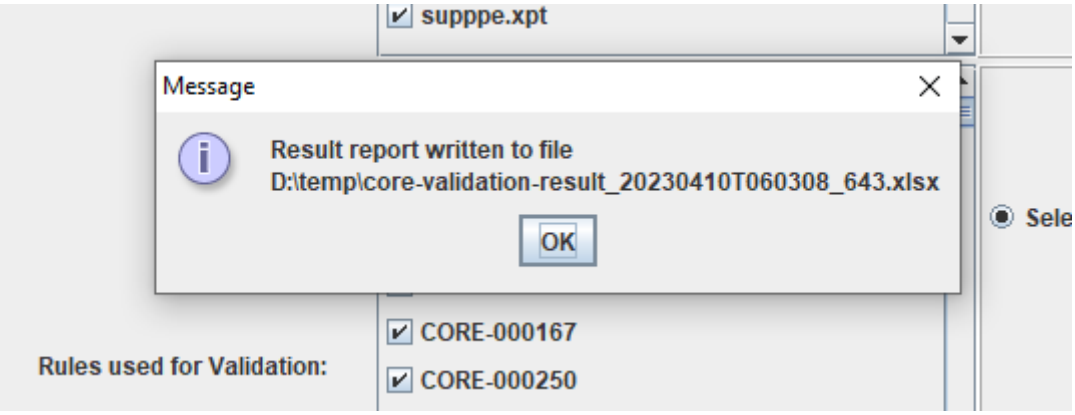
One can then select the rules about --SCAT that one want to be executed by checking the checkbox on the left side.

Clicking the "OK" button then confirms the selection. This e.g. leads to:

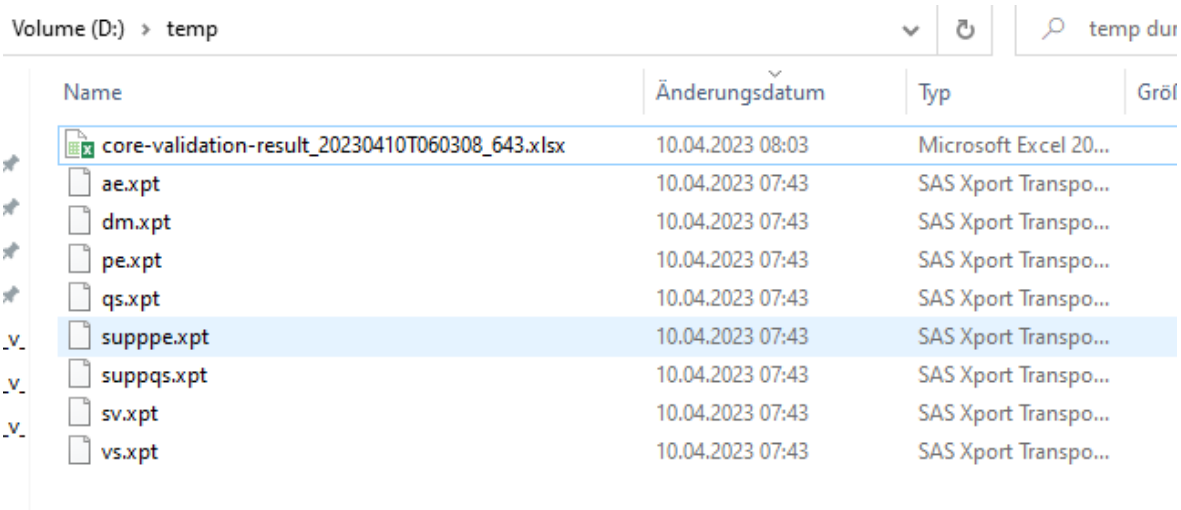


But now let us do a validation with all rules on all generated files.

When then clicking the button "**Execute CORE Validation**", after about 20 seconds, the user is informed that the validation is ready. For example:



with the files in the D:\temp directory in our case:



and then opening the Excel report with you favorite spreadsheet program, and navigating to the tab "Issue Summary":

	A	B	C	D
1	Dataset	CORE-ID	Message	Issues
2	AE	CORE-000022	At least one of the Seriousness criteria (AESCAN, AESCONG, AESDISAB, AESDTH, AESHOSP, AESLIFE, AESOD or AESMIE) = 'Y', but AESER = 'N' or empty.	1
3	AE	CORE-000264	Primary analysis used but AEBODSYS and AESOC are not equal	6
4	DM	CORE-000191	RFENDTC is missing when ARM is provided.	1
5	PE	CORE-000021	PESTRESC should not be blank when either PEORRES is provided or PEDRVFL = 'Y' .	27
6	PE	CORE-000099	Value for PESTAT is populated, when PEORRES is populated	1
7				
8				
9				
10				
11				

or the tab "Issue Details":

	A	B	C	D	E	F	G	H	
21	CORE-000021	PESTRESC should not be blank when either PEORRES is provided or PEDRVFL = 'Y' .		PE	008	192	10	PEDRVFL, PEORRES, PESTRESC	Not in dataset, Abnormal,
22	CORE-000021	PESTRESC should not be blank when either PEORRES is provided or PEDRVFL = 'Y' .		PE	008	195	13	PEDRVFL, PEORRES, PESTRESC	Not in dataset, Not Done,
23	CORE-000021	PESTRESC should not be blank when either PEORRES is provided or PEDRVFL = 'Y' .		PE	009	203	8	PEDRVFL, PEORRES, PESTRESC	Not in dataset, Not Done,
24	CORE-000021	PESTRESC should not be blank when either PEORRES is provided or PEDRVFL = 'Y' .		PE	009	216	21	PEDRVFL, PEORRES, PESTRESC	Not in dataset, Not Done,
25	CORE-000021	PESTRESC should not be blank when either PEORRES is provided or PEDRVFL = 'Y' .		PE	010	229	8	PEDRVFL, PEORRES, PESTRESC	Not in dataset, Not Done,
26	CORE-000021	PESTRESC should not be blank when either PEORRES is provided or PEDRVFL = 'Y' .		PE	010	242	21	PEDRVFL, PEORRES, PESTRESC	Not in dataset, Not Done,
27	CORE-000021	PESTRESC should not be blank when either PEORRES is provided or PEDRVFL = 'Y' .		PE	011	255	8	PEDRVFL, PEORRES, PESTRESC	Not in dataset, Not Done,
28	CORE-000021	PESTRESC should not be blank when either PEORRES is provided or PEDRVFL = 'Y' .		PE	012	268	8	PEDRVFL, PEORRES, PESTRESC	Not in dataset, Not Done,
29	CORE-000022	At least one of the Seriousness criteria (AESCAN, AESCONG, AESDISAB, AESDTH, AESHOSP, AESLIFE, AESOD or AESMIE) = 'Y', but AESER = 'N' or empty.	fully executable	AE	004	5	1	AESCAN, AESCONG, AESDISAB, AESDTH, AESER, AESHOSP, AESLIFE, AESMIE, AESOD, AESTDTC, AETERM	Not in dataset, Not in dataset, Not dataset, 2006-04-22, MYOCARDIAL
30	CORE-000099	Value for PESTAT is populated, when PEORRES is populated		PE				PEORRES, PESTAT	Not Done, NOT DONE
31	CORE-000191	RFENDTC is missing when ARM is provided.		DM	007	7		ARM, RFENDTC	Treatment A,
32	CORE-000264	Primary analysis used but AEBODSYS and AESOC are not equal	fully executable	AE	001	1	1	AEBODSYS, AESOC, AETERM	, Not in dataset, HEADACHE
33	CORE-000264	Primary analysis used but AEBODSYS and AESOC are not equal	fully	AE	001	2	2	AEBODSYS, AESOC, AETERM	, Not in dataset, CONGESTION
34	CORE-000264	Primary analysis used but AEBODSYS and AESOC are not equal	fully	AE	001	3	3	AEBODSYS, AESOC, AETERM	, Not in dataset, TOOTHACHE
35	CORE-000264	Primary analysis used but AEBODSYS and AESOC are not equal	fully	AE	002	4	1	AEBODSYS, AESOC, AETERM	, Not in dataset, HEARTH FAILURE
36	CORE-000264	Primary analysis used but AEBODSYS and AESOC are not equal	fully	AE	004	5	1	AEBODSYS, AESOC, AETERM	, Not in dataset, MYOCARDIAL ISCH
37	CORE-000264	Primary analysis used but AEBODSYS and AESOC are not equal	fully	AE	008	6	1	AEBODSYS, AESOC, AETERM	, Not in dataset, GASTROENTERITIS
38									
39									
40									

Very important is that, under "Rules Report", the report also displays which rules have been executed, and which have been skipped, for example as the rule is only applicable to a domain for which there is no dataset in the set of generated XPT files (or selection of it):

	A	B	C	D	E	F	G
19	CORE-000018	1	CG0086			--OCCUR is blank when --PRESP is equal to "Y" and --STAT is not provided.	SKIPPED
20	CORE-000019	1	CG0311			Variable label length should be less than or equal to 40 characters	SKIPPED
21	CORE-000020	1	CG0206			TAETORD should be null when ETCD = 'UNPLAN'.	SKIPPED
22	CORE-000021	1	CG0397			--STRESC should not be blank when either --ORRES is provided or --DRVFL = 'Y' .	SUCCESS
23	CORE-000022	1	CG0041			At least one of the Seriousness criteria (AESCAN, AESCONG, AESDISAB, AESDTH, AESHOSP, AESLIFE, AESOD or AESMIE) = 'Y', but AESER = 'N' or empty.	SUCCESS
24	CORE-000023	1	CG0084			--TOX present in dataset even though --TOXGR is not	SUCCESS
25	CORE-000024	1	CG0082			--BODSYS is not empty and --BDSYCD is empty	SKIPPED
26	CORE-000025	1	CG0177			IESTRESC is not equal to IEORRES	SKIPPED
27	CORE-000026	1	CG0468			The --TPTNUM variable does not exist when --TPT does exist.	SUCCESS
28	CORE-000027	1	CG0328			At least one of TEENRL and TEDUR must be populated.	SKIPPED
29	CORE-000028	1	CG0008			--TPTREF is empty and --ELTM is not empty	SKIPPED
30	CORE-000029	1	CG0661			--TPTNUM exists in a dataset, but --TPT does not exist. When time points are represented in SDTMIG domains, both --TPT and --TPTNUM must be used.	SUCCESS
31	CORE-000030	1	CG0053			--REASND should not be present in dataset when --PRESP is not present in dataset	SUCCESS
32	CORE-000031	1	CG0659			--EVAL is present. --EVAL must not be used to model QRS data.	SUCCESS
33	CORE-000032	1	CG0660			--EVALID is present. --EVALID must not be used to model QRS data.	SUCCESS
34	CORE-000033	1	CG0065			DSDECOD is not equal to "COMPLETED" when DSTERM equals "COMPLETED".	SKIPPED
35	CORE-000034	1	CG0069			DSSTDTC does not equal DM.DTHDTC, when DSDECOD equals "DEATH".	SKIPPED
36	CORE-000035	1	CG0658			VISITDY is populated when SVPRESP is null. VISITDY is the Planned Study Day of a visit. It should not be populated for unplanned visits.	SKIPPED
37	CORE-000036	1	CG0657			Planned visit is not found in TV.	SKIPPED
38	CORE-000037	1	CG0653			SVPRESP is not null and not equal to "Y". Values should be "Y" or null.	SKIPPED
39	CORE-000038	1	CG0654			SVPRESP is not "Y" when SVOCCUR is populated.	SKIPPED
40	CORE-000039	1	CG0655			VISITNUM for planned visit is not in TV.	SKIPPED
41	CORE-000040	1	CG0656			VISITNUM for unplanned visit is present in TV.	SKIPPED
42	CORE-000041	1	CG0649			TSVALNF is populated when TSVAL is populated with a value that is not in the ISO 21090 null flavor codelist.	SKIPPED
43	CORE-000042	1	CG0647			TT dataset is present.	SKIPPED
44	CORE-000043	1	CG0648			TP dataset is present.	SKIPPED
45	CORE-000044	1	CG0646			SJ dataset is present.	SKIPPED
46	CORE-000045	1	CG0517			ARMNRS value is missing when ARMCD value is missing	SKIPPED
47	CORE-000046	1	CG0519			ARMNRS is missing when ARM value is missing	SKIPPED
48	CORE-000047	1	CG0518			ARM value in DM dataset is not among the values of ARM variable in TA dataset. This is allowed only in a multistage study with incomplete ARM assignment. Please confirm if your study is a multistage assignment study	SKIPPED
49	CORE-000048	1	CG0621			--METHOD is present in an interventions domain.	SKIPPED

Conformance Details | Dataset Details | Issue Summary | Issue Details | Rules Report

For example, the rule CORE-000028 (CDISC CG0008) was skipped, as none of our datasets contains the variable --TPTREF or --ELTM. Rules however that are applicable to at least one of our datasets and that ran successfully (not meaning whether a violation was found) are marked "Success".

One can now start using the results of this CORE validation to improve the mappings for the generation of SDTM or SEND datasets.

Conclusion

CDISC CORE is a revolution in the area of validation of CDISC datasets for submissions of datasets to the regulatory authorities (and beyond that).

As one of the first software vendors (if not the first), we have implemented CORE in our software, enabling to use CORE from within SDTM-ETL v.4.2. The implementation allows to select on datasets generated, as well as on rules to be executed during the validation process.