SDTM-ETL 5.0 User Manual and Tutorial

Author: Jozef Aerts, XML4Pharma

Last update: 2025-01-09

Tutorial: Merging datasets

Table of Contents

Introduction

SDTM and regulatory authorities don't make it easy to us when it comes to generating and submitting comprehensive and especially logical, sets of data.
For example, even now in 2025, it forces us to "ban" non-standard variables (i.e. sponsor-defined variables) to a supplemental qualifier (SUPPxx) dataset. Also, when a data point value exceeds 200 characters, it must be split into chunks of not more than 200 characters and the second and further chunk must be banned to SUPPxx. Also, there is the famous 5GB file size limit, forcing us to "split" datasets[1] when the XPT file size grows beyond that limit.
Essentially, there are three reasons for all this:
The first is the mandated use of the outdated SAS Transport 5 (XPT) format, which is extremely inefficient in byte storage. Selecting XPT 30 years ago was a major error, even simple CSV would have been a better choice.
The second is the lack of SDTM understanding at the FDA: many of the reviewers are not able to distinguish between standard and non-standard SDTM variable, so CDISC decided that these need to go into SUPPxx datasets, which reviewers are supposed to then re-merge into the "parent" dataset. This could easily be solved e.g. by color-coding of non-standard variable columns in modern viewers (as the information about standard and non-standard is in the define.xml), but the relative primitive tools (third reason) at the regulatory authorities do not support this.

Modern viewers such as the "Smart Submission Dataset Viewer" have such features, e.g.:

[1] The interesting thing is that in such case, one still needs to submit the >5GB dataset, although FDA claims that it cannot read these ...

AGE	AGEU	SEX	RACE	ETHNIC	ARMCD	ARM	ACTARMCD	ACTARM	COUNTRY	DMDTC	DMDY	COMPLT16	COMPLT24	COMPLT8	EFFICACY	SAFETY	ITT
61	YEARS	F	WHITE	HISPANIC ...	Pbo	Placebo	Pbo	Placebo	USA	2013-12-26	-7	Y	Y	Y	Y	Y	Y
64	YEARS	M	WHITE	HISPANIC ...	Pbo	Placebo	Pbo	Placebo	USA	2012-07-22	-14						
71	YEARS	M	WHITE	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2013-07-11	-8	Y	Y				
74	YEARS	M	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2014-03-10	-8						
77	YEARS	F	WHITE	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2014-06-24	-7	Y	Y				
85	YEARS	F	WHITE	NOT HISPA...	Pbo	Placebo	Pbo	Placebo	USA	2013-01-22	-21						
59	YEARS	F	WHITE	HISPANIC ...	Scrnfail	Screen Fail...	Scrnfail	Screen Fail...	USA	2013-12-20							
68	YEARS	M	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2013-12-23	-9	Y	Y				
81	YEARS	F	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2012-08-25	-13						
84	YEARS	M	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2012-11-23	-7			Y	Y	Y	Y
52	YEARS	M	WHITE	NOT HISPA...	Pbo	Placebo	Pbo	Placebo	USA	2014-02-27	-13	Y	Y	Y	Y	Y	Y
84	YEARS	M	WHITE	NOT HISPA...	Pbo	Placebo	Pbo	Placebo	USA	2014-02-09	-6	Y	Y	Y	Y	Y	Y
81	YEARS	F	WHITE	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2012-10-23	-5	Y	Y	Y	Y	Y	Y
57	YEARS	F	WHITE	NOT HISPA...	Scrnfail	Screen Fail...	Scrnfail	Screen Fail...	USA	2013-09-05							
75	YEARS	F	WHITE	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2013-05-07	-13				Y	Y	Y
57	YEARS	M	WHITE	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2013-08-14	-9	Y	Y	Y	Y	Y	Y
79	YEARS	F	WHITE	NOT HISPA...	Pbo	Placebo	Pbo	Placebo	USA	2013-09-06	-17	Y	Y	Y	Y	Y	Y
82	YEARS	F	WHITE	NOT HISPA...	Scrnfail	Screen Fail...	Scrnfail	Screen Fail...	USA	2013-04-18							
62	YEARS	F	AMERICAN...	NOT HISPA...	Scrnfail	Screen Fail...	Scrnfail	Screen Fail...	USA	2012-09-30							
56	YEARS	M	WHITE	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2013-01-28	-15				Y	Y	Y
79	YEARS	F	WHITE	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Lo	Xanomelin...	USA	2013-11-26	-9				Y	Y	Y
71	YEARS	M	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2013-02-03	-12				Y	Y	Y
80	YEARS	F	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2012-07-08	-14	Y	Y	Y	Y	Y	Y
81	YEARS	F	BLACK OR ...	NOT HISPA...	Pbo	Placebo	Pbo	Placebo	USA	2013-01-25	-8	Y	Y	Y	Y	Y	Y
76	YEARS	F	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2012-10-30	-16			Y	Y	Y	Y
69	YEARS	M	WHITE	NOT HISPA...	Pbo	Placebo	Pbo	Placebo	USA	2013-03-20	-10	Y	Y	Y	Y	Y	Y
56	YEARS	M	WHITE	HISPANIC ...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2013-12-28	-14	Y	Y	Y	Y	Y	Y
57	YEARS	F	BLACK OR ...	NOT HISPA...	Scrnfail	Screen Fail...	Scrnfail	Screen Fail...	USA	2013-09-22							
61	YEARS	M	AMERICAN...	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2014-01-25	-13	Y		Y	Y	Y	Y
56	YEARS	F	WHITE	HISPANIC ...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2014-01-17	-8	Y	Y	Y	Y	Y	Y
67	YEARS	M	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2013-03-17	-7			Y	Y	Y	Y
61	YEARS	M	WHITE	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Hi	Xanomelin...	USA	2013-08-20	-9			Y	Y	Y	Y
80	YEARS	F	WHITE	NOT HISPA...	Scrnfail	Screen Fail...	Scrnfail	Screen Fail...	USA	2013-12-08							
68	YEARS	M	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2014-05-10	-12	Y	Y	Y	Y	Y	Y
79	YEARS	M	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2012-09-16	-16	Y	Y	Y	Y	Y	Y
51	YEARS	M	WHITE	NOT HISPA...	Xan_Lo	Xanomelin...	Xan_Lo	Xanomelin...	USA	2012-12-22	-14				Y	Y	Y
63	YEARS	F	WHITE	NOT HISPA...	Pbo	Placebo	Pbo	Placebo	USA	2013-10-01	-7	Y	Y	Y	Y	Y	Y
54	YEARS	F	WHITE	HISPANIC ...	Scrnfail	Screen Fail...	Scrnfail	Screen Fail...	USA	2014-04-01							
67	YEARS	M	WHITE	NOT HISPA...	Xan_Hi	Xanomelin...	Xan_Lo	Xanomelin...	USA	2013-07-24	-7				Y	Y	Y
81	YEARS	F	BLACK OR ...	NOT HISPA...	Pbo	Placebo	Pbo	Placebo	USA	2013-05-20	-10	Y	Y	Y	Y	Y	Y
74	YEARS	M	WHITE	NOT HISPA...	Scrnfail	Screen Fail...	Scrnfail	Screen Fail...	USA	2013-09-18							

Define-XML metadata:
Name: COMPLT24
Label: Completers of Week 24 Population Flag
Mandatory: No
Datatype: text
Length: 1
CodeList: Y_BLANK
Non-Standard Variable (NSV)

(the DM file is from the famous LZZT CDISC-FDA SDTM pilot)

SDTM-ETL and 'split' datasets

With SDTM-ETL, we never need to "split" datasets, we take care in advance that we never produce huge datasets (larger than 5GB) that we then later need to split. Even for large datasets smaller than 5GB, we need to think about whether such huge datasets are "usable" for reviewers. For example, an LB-dataset with millions of rows spread over different categories of lab tests will be hard to analyze for reviewers - they will need to start to filter to make these usable.
Therefore, the better strategy, followed by SDTM-ETL, is to develop different instances of the same domain, usually one dataset per category. This also makes the mapping considerably easier, as data for different categories can come from different sources, e.g. some from EDC, other from electronic data transfers (e.g. in CSV format).

This strategy e.g. leads to different dataset definitions in SDTM-ETL, for example:

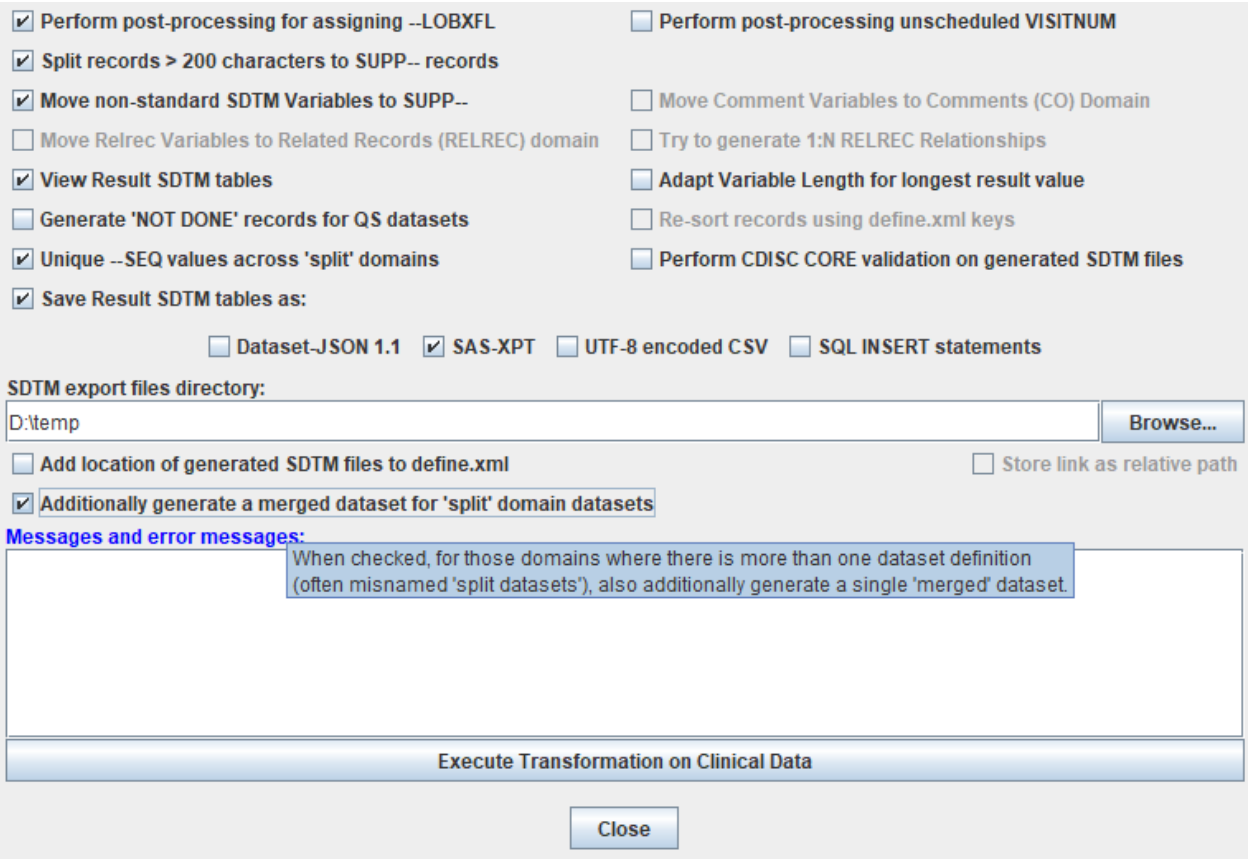
RELREC	STUDYID	RDOMAIN	USUBJID	IDVAR
SUPPQUAL	STUDYID	RDOMAIN	USUBJID	IDVAR
RELSUB	STUDYID	USUBJID	RELSUB.POOLID	RELSUB.RS
OI	STUDYID	DOMAIN	OI.NHOID	OI.OISEQ
[redacted]GLOBAL	STARTREF			
[redacted]LBUR	STUDYID	DOMAIN	USUBJID	LB.LBSEQ
[redacted]LBBL	STUDYID	DOMAIN	USUBJID	LB.LBSEQ
[redacted]LBCO	STUDYID	DOMAIN	USUBJID	LB.LBSEQ
[redacted]LBBR	STUDYID	DOMAIN	USUBJID	LB.LBSEQ

where it was decided to develop mappings and generate 4 instances of the LB (Laboratory) domain: one for urinalysis, one for all all blood tests (chemistry and hematology), one for coagulation tests, and one for breath alcohol tests.
Of course the choice of categories is arbitrary. For example, we could also have chosen to generate separate datasets for hematology and blood chemistry.

With the above setup, developing the mappings will be considerably easy, and lead to 4 submission datasets. In case SAS Transport is used: LBUR.xpt, LBBL.xpt, LBCO.xpt and LBBR.xpt. These will be considerably easier to analyze by the regulatory reviewers than would be the case with one huge LB dataset.

However, these authorities (or in case work is done as a service provider, the sponsor) will also want to see a single LB dataset.

In SDTM-ETL, this can be simply realized during execution of the mappings (when theXPT datasets are generated), by checking the checkbox "*Additionally generate a merged dataset for 'split' domain datasets*".



When it is checked, also a single dataset named "LB.xpt" is generated in the output folder, containing the combined content of the separate LBUR, LBBL, etc. datasets.
The same is of course also possible when generating datasets in the modern CDISC Dataset-JSON 1.1 format, which is the successor for the outdated XPT format. In that case, splitting-off records with more than 200 characters doesn't make sense.

Remark that for QS (Questionnaires), one will always want to generate distinct datasets, one for each questionnaire, and will never merge these into a single dataset.

Merging datasets that were generated separately

The above approach works very well when the different instances of the same domain have been developed together, or are loaded by merging (using the menu "File - Load Study define.xml" followed by "I want to merge with the existing define.xml").

There may be circumstances however, that also this approach does not work.
One such case is that we have two (or more) types of data in SUPPxx:
- SUPPxx dataset with "non-standard variables" automatic split off during mapping execution, and maybe containing values beyond the 200 character limitation.
- custom SUPPxx datasets (derived from the "SUPPQUAL" in the template)

Another example is the case that we have generated different CO (Comments) datasets in different sessions, by automatic split of ("Comment variable in the dataset definition" generated by "Insert - New SDTM Variable for Comment" - see the tutorial "Auto-generation of comments and putting them in the Comments (CO) domain.

In the past, when it was supposed that FDA, and other regulatory would move to Dataset-XML format, we had an application "XML2SASDatasetMerger" allowing to do so without the need of (expensive) SAS software. Now that it is clear that the future is JSON, and it is expected that FDA and other regulatory agencies will soon accept submissions in CDISC Dataset-JSON-1.1 format, we will develop a separate application (and deliver it with SDTM-ETL) to merge Dataset-JSON-1.1 files, and also generate XPT files for these. The latter will however not be really necessary, as the regulatory agencies are expected to move away from XPT anymore.

Conclusions

The by FDA and other regulatory authorities mandated (but outdated) SAS-XPT format makes life unnecessary complicated when generating CDISC SDTM and SEND datasets.
Essentially, "splitting" and "merging" datasets should never have to be done, and if so anyway, it should be easy to accomplish. It is the SAS-XPT format that makes it complicated: when using XML (e.g. Dataset-XML) or JSON (e.g. Dataset-JSON), this is much easier to accomplish.

This tutorial demonstrated two ways of merging datasets for the case of XPT as the output format. For the simple case that several instances of the same domain are present in the same "working" define.xml, one can set the "*Additionally generate a merged dataset for 'split' domain datasets"* checkbox.

We expect that FDA will give green light for Dataset-JSON submissions pretty soon. As soon as that is clear, we will also provide functionalities for merging Dataset-JSON files. Developing these will however be much more straightforward than for SAS-XPT.