SDTM-ETL 4.1 User Manual and Tutorial

Author: Jozef Aerts, XML4Pharma

Last update: 2022-09-24

Generating datasets in the CDISC Dataset-JSON format - the submission format of the future

SAS Transport 5 ("XPT format") is a very old format, one generation after the punch card, and thus completely outdated. Still the regulatory authorities such as FDA, PMDA and NMPA require the use of the SAS Transport 5 format for regulatory submissions. As other industries have already since a long time given up the use of SAS Transport (5 or 8) format, the Clinical Research industry is the only industry that is still using this outdated format.

A few of the most important disadvantages of SAS Transport are:

- only US-ASCII is supported.
  This is especially important for submissions to PMDA and NMPA (no support for Japanese and Chinese characters)
- 8-, 40- and 200-character (bytes) limitations for variable names, variable labels and variable values.
- not vendor-neutral
- not very suitable for modern RESTful web services

And in the scope of CDISC standards

- 8-character (bytes) limitation for variables values that might be used in transposal such as --TESTCD, QNAM, ...
- 40-characters (bytes) limitation for variables values that might be used as labels in transposal such as --TEST, QLABEL, ...

Some have proposed the use of SAS Transport ***8*** as an solution for the character length limitations but this essentially is a fake solution, as it doesn't solve the ASCII-only problems[1] and vendor-neutrality issues.

Already many years ago, CDISC developed Dataset-XML as a modern alternative for SAS Transport. The FDA then conducted a pilot with a number of pharma companies and with the support of CDISC. Essentially however, the pilot approach was not very good, essentially testing features that were already proven many times before. For example, the feature that Dataset-XML fully supports Unicode (which supports all written languages in the world) was not taken into account at all.
The FDA report (unfortunately not available anymore) mentioned that the FDA would expect problems with file sizes, as in some cases, Dataset-XML files can be considerably larger in size than SAS-XPT files. We think however that this was just a cheap excuse to not have to change anything.

Recently, CDISC developed the Dataset-JSON format, which is similar to Dataset-XML, but then in ... JSON. Major advantages of Dataset-JSON are:

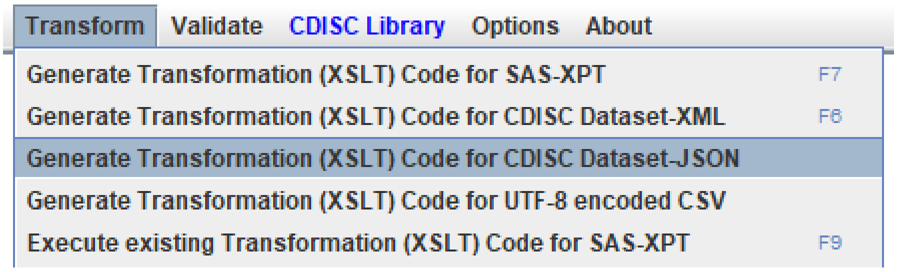
[1] Using SAS Transport 8, it is possible to have non-ASCII characters using some "tricks", but when inspecting such files, it is impossible to find out whether, and which of these "tricks" have been used.

- All advantages of Dataset-XML: no character limitations, Unicode "out of the box"
- Much smaller file sizes than SAS Transport 5 (even by a factor of 10)
- For basic consumption, no use of the define.xml is necessary
- Extremely suitable for use with RESTful web services and APIs

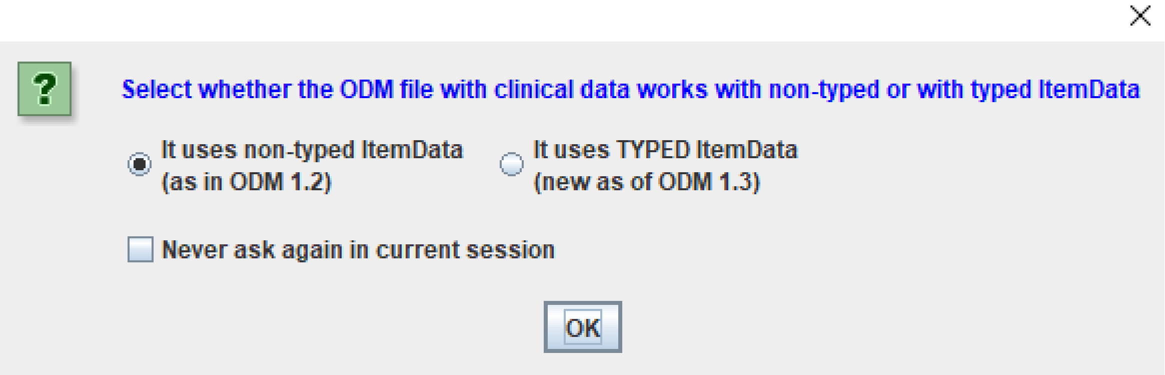
The latter would be extremely worthful once regulatory authorities and the pharma industry come to a "neutral zone", very probably in the cloud, where submissions can essentially be started once the first data point has been collected, and which is governed by RESTful web services, i.e. having the ability to work without the concept of "files".

SDTM-ETL now fully supports the generation of SDTM and SEND files in the new CDISC Dataset-JSON format, including the visualization with the free and open-source "Smart Submission Dataset Viewer".

In order to generate datasets in the new Dataset-JSON, instead of using the menu "Transform - Generate Transformation (XSLT) Code for SAS-XPT", the menu "Transform - Generate Transformation (XSLT) Code for CDISC Dataset-JSON":


Transform
Validate
CDISC Library
Options
About
Generate Transformation (XSLT) Code for SAS-XPT F7
Generate Transformation (XSLT) Code for CDISC Dataset-XML F6
Generate Transformation (XSLT) Code for CDISC Dataset-JSON
Generate Transformation (XSLT) Code for UTF-8 encoded CSV
Execute existing Transformation (XSLT) Code for SAS-XPT F9

After a few moments, the following dialog then shows up:

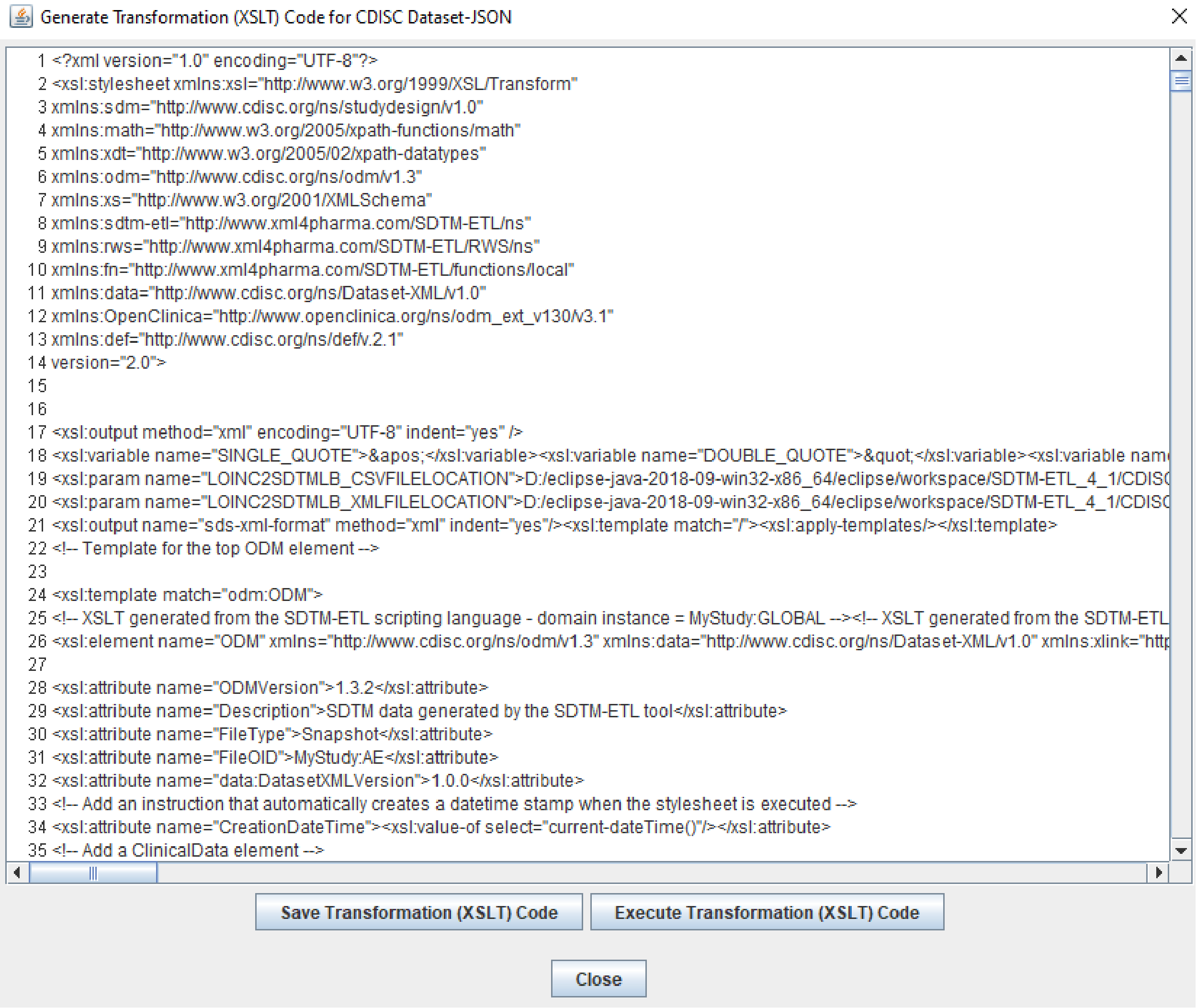


Select whether the ODM file with clinical data works with non-typed or with typed ItemData

● It uses non-typed ItemData (as in ODM 1.2)　○ It uses TYPED ItemData (new as of ODM 1.3)

□ Never ask again in current session

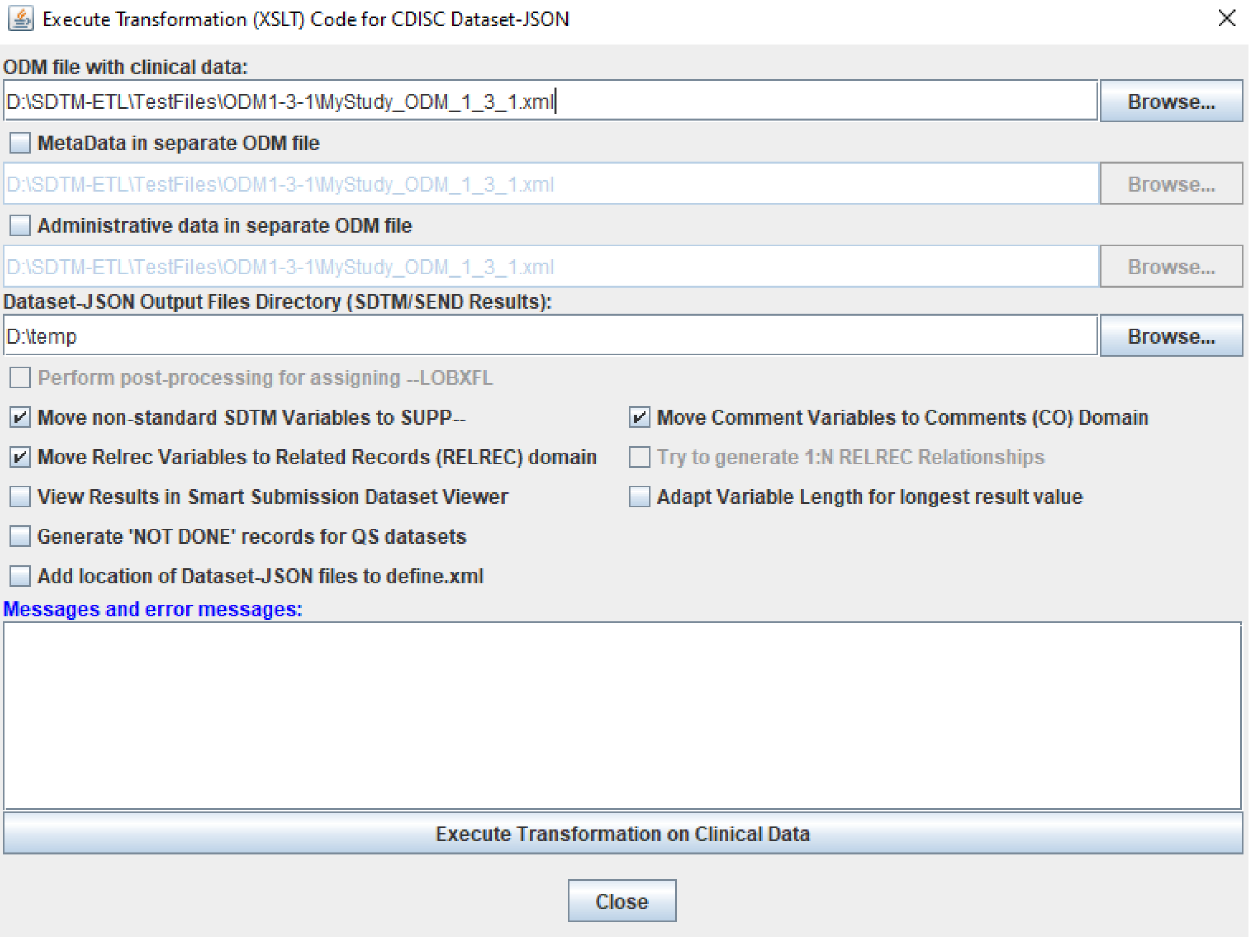
Allowing to choose between the two "tastes" of ODM files with clinical data, and then followed by a new window:



```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xmlns:sdm="http://www.cdisc.org/ns/studydesign/v1.0"
xmlns:math="http://www.w3.org/2005/xpath-functions/math"
xmlns:xdt="http://www.w3.org/2005/02/xpath-datatypes"
xmlns:odm="http://www.cdisc.org/ns/odm/v1.3"
xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:sdtm-etl="http://www.xml4pharma.com/SDTM-ETL/ns"
xmlns:rws="http://www.xml4pharma.com/SDTM-ETL/RWS/ns"
xmlns:fn="http://www.xml4pharma.com/SDTM-ETL/functions/local"
xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0"
xmlns:OpenClinica="http://www.openclinica.org/ns/odm_ext_v130/v3.1"
xmlns:def="http://www.cdisc.org/ns/def/v.2.1"
version="2.0">


<xsl:output method="xml" encoding="UTF-8" indent="yes" />
<xsl:variable name="SINGLE_QUOTE">&apos;</xsl:variable><xsl:variable name="DOUBLE_QUOTE">&quot;</xsl:variable><xsl:variable nam
<xsl:param name="LOINC2SDTMLB_CSVFILELOCATION">D:/eclipse-java-2018-09-win32-x86_64/eclipse/workspace/SDTM-ETL_4_1/CDIS
<xsl:param name="LOINC2SDTMLB_XMLFILELOCATION">D:/eclipse-java-2018-09-win32-x86_64/eclipse/workspace/SDTM-ETL_4_1/CDIS
<xsl:output name="sds-xml-format" method="xml" indent="yes"/><xsl:template match="/"><xsl:apply-templates/></xsl:template>
<!-- Template for the top ODM element -->

<xsl:template match="odm:ODM">
<!-- XSLT generated from the SDTM-ETL scripting language - domain instance = MyStudy:GLOBAL --><!-- XSLT generated from the SDTM-ETL
<xsl:element name="ODM" xmlns="http://www.cdisc.org/ns/odm/v1.3" xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0" xmlns:xlink="http

<xsl:attribute name="ODMVersion">1.3.2</xsl:attribute>
<xsl:attribute name="Description">SDTM data generated by the SDTM-ETL tool</xsl:attribute>
<xsl:attribute name="FileType">Snapshot</xsl:attribute>
<xsl:attribute name="FileOID">MyStudy:AE</xsl:attribute>
<xsl:attribute name="data:DatasetXMLVersion">1.0.0</xsl:attribute>
<!-- Add an instruction that automatically creates a datetime stamp when the stylesheet is executed -->
<xsl:attribute name="CreationDateTime"><xsl:value-of select="current-dateTime()"/></xsl:attribute>
<!-- Add a ClinicalData element -->
```

Save Transformation (XSLT) Code
Execute Transformation (XSLT) Code
Close

The generated transformation code can then be saved to file for later execution, e.g. when batch generation of SDTM/SEND datasets are planned (e.g. using a job-scheduling mechanism), or executed when clicking the "Execute Transformation (XSLT) Code". The latter then leads to:



Execute Transformation (XSLT) Code for CDISC Dataset-JSON

ODM file with clinical data: D:\SDTM-ETL\TestFiles\ODM1-3-1\MyStudy_ODM_1_3_1.xml

□ MetaData in separate ODM file

D:\SDTM-ETL\TestFiles\ODM1-3-1\MyStudy_ODM_1_3_1.xml

□ Administrative data in separate ODM file

D:\SDTM-ETL\TestFiles\ODM1-3-1\MyStudy_ODM_1_3_1.xml

Dataset-JSON Output Files Directory (SDTM/SEND Results): D:\temp

□ Perform post-processing for assigning --LOBXFL

☑ Move non-standard SDTM Variables to SUPP--

☑ Move Comment Variables to Comments (CO) Domain

☑ Move Relrec Variables to Related Records (RELREC) domain

□ Try to generate 1:N RELREC Relationships

□ View Results in Smart Submission Dataset Viewer

□ Adapt Variable Length for longest result value

□ Generate 'NOT DONE' records for QS datasets

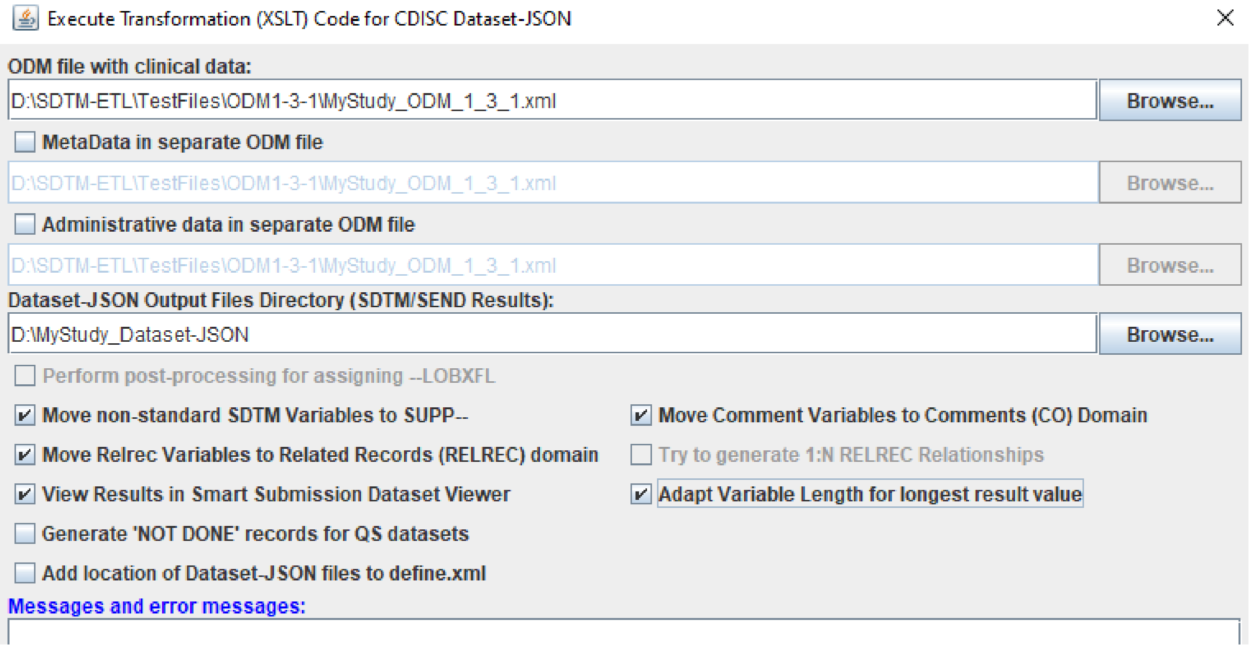
□ Add location of Dataset-JSON files to define.xml

Messages and error messages:

The window looks very similar to the one when generating SAS-XPT files with some major differences:

- There is no checkbox "Split records > 200 characters to SUPP-- records"
  The reason is that, as a modern format, Dataset-JSON does not have any character length limitations, so that no splitting is at all necessary
- The checkbox "View Result SDTM tables" is replaced by a checkbox "View Results in Smart Submission Dataset Viewer.
  If this checkbox is checked, after the datasets are generated, the "Smart Submission Dataset Viewer" will be started, allowing to inspect the datasets, and much much more (as it is "smart").
- The field for providing the location where the generated datasets need to be written to (a folder) is provided higher up.
- There checkbox "Adapt Variable Length for longest result value" is still present. In the case of SAS-XPT it is necessary for making the XPT files as compact as possible (SAS-XPT is an extremely inefficient format). In the case of Dataset-JSON, it is still there to set the "Length" attribute on the variable definition in the define.xml, as this is still required by the Define-XML standard. In Dataset-JSON, adding the maximal length is optional, but in our case, it is still added, in order to provide backward compatibility for some tools.

We add some information and select some of the choices, e.g.:



Execute Transformation (XSLT) Code for CDISC Dataset-JSON

ODM file with clinical data:

D:\SDTM-ETL\TestFiles\ODM1-3-1\MyStudy_ODM_1_3_1.xml

□ MetaData in separate ODM file

D:\SDTM-ETL\TestFiles\ODM1-3-1\MyStudy_ODM_1_3_1.xml

□ Administrative data in separate ODM file

D:\SDTM-ETL\TestFiles\ODM1-3-1\MyStudy_ODM_1_3_1.xml

Dataset-JSON Output Files Directory (SDTM/SEND Results):

D:\MyStudy_Dataset-JSON

□ Perform post-processing for assigning --LOBXFL

☑ Move non-standard SDTM Variables to SUPP--

☑ Move Comment Variables to Comments (CO) Domain

☑ Move Relrec Variables to Related Records (RELREC) domain

□ Try to generate 1:N RELREC Relationships

☑ View Results in Smart Submission Dataset Viewer

☑ Adapt Variable Length for longest result value

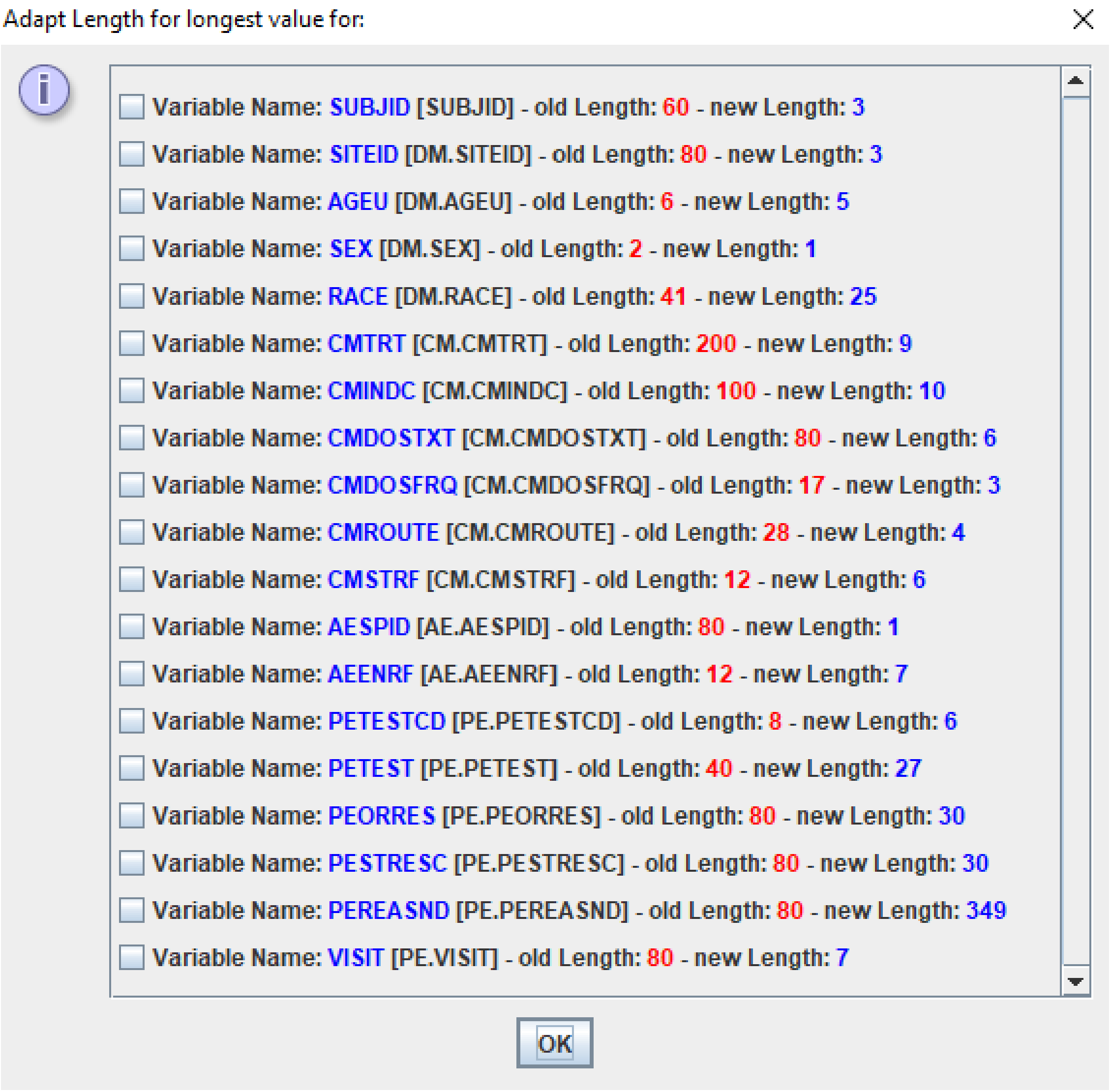
□ Generate 'NOT DONE' records for QS datasets

□ Add location of Dataset-JSON files to define.xml

Messages and error messages:

asking to put the results in the folder "D:\MyStudy_Dataset-JSON", and to open the "Smart Submission Dataset Viewer" once the generated files are ready.

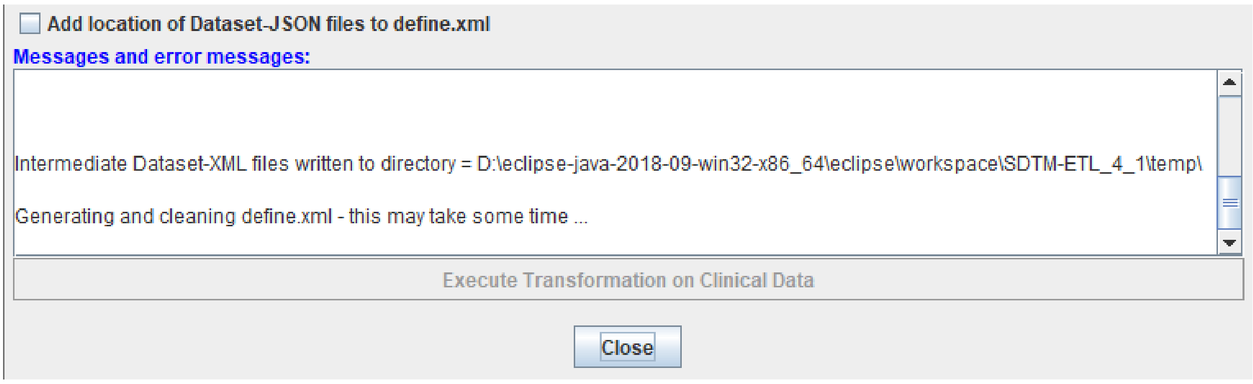
Then clicking "Execute Transformation on Clinical Data" starts the transformation.
As we have checked the checkbox "Adapt Variable Length", first a window is displayed, allowing us to select for which variables we want to have the maximal length adapted:



Adapt Length for longest value for:

- □ Variable Name: SUBJID [SUBJID] - old Length: 60 - new Length: 3
- □ Variable Name: SITEID [DM.SITEID] - old Length: 80 - new Length: 3
- □ Variable Name: AGEU [DM.AGEU] - old Length: 6 - new Length: 5
- □ Variable Name: SEX [DM.SEX] - old Length: 2 - new Length: 1
- □ Variable Name: RACE [DM.RACE] - old Length: 41 - new Length: 25
- □ Variable Name: CMTRT [CM.CMTRT] - old Length: 200 - new Length: 9
- □ Variable Name: CMINDC [CM.CMINDC] - old Length: 100 - new Length: 10
- □ Variable Name: CMDOSTXT [CM.CMDOSTXT] - old Length: 80 - new Length: 6
- □ Variable Name: CMDOSFRQ [CM.CMDOSFRQ] - old Length: 17 - new Length: 3
- □ Variable Name: CMROUTE [CM.CMROUTE] - old Length: 28 - new Length: 4
- □ Variable Name: CMSTRF [CM.CMSTRF] - old Length: 12 - new Length: 6
- □ Variable Name: AESPID [AE.AESPID] - old Length: 80 - new Length: 1
- □ Variable Name: AEENRF [AE.AEENRF] - old Length: 12 - new Length: 7
- □ Variable Name: PETESTCD [PE.PETESTCD] - old Length: 8 - new Length: 6
- □ Variable Name: PETEST [PE.PETEST] - old Length: 40 - new Length: 27
- □ Variable Name: PEORRES [PE.PEORRES] - old Length: 80 - new Length: 30
- □ Variable Name: PESTRESC [PE.PESTRESC] - old Length: 80 - new Length: 30
- □ Variable Name: PEREASND [PE.PEREASND] - old Length: 80 - new Length: 349
- □ Variable Name: VISIT [PE.VISIT] - old Length: 80 - new Length: 7

During mapping development, one will often choose to not adapt variable lengths, but it can be of importance when generating the final datasets.
In our case, remark the "highest" length being 349 for PEREASND. In the case of SAS-XPT, this would have led to two additional records in an artificially generated SUPPPE.xpt file. In the case of Dataset-JSON, everything can stay nicely together.
If we check all the checkboxes and then click "OK", some messages first pass by:

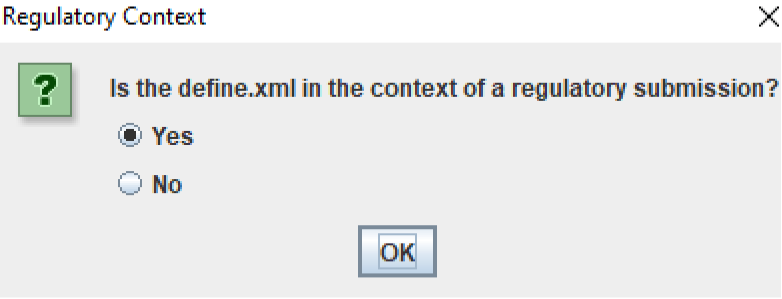


□ Add location of Dataset-JSON files to define.xml

Messages and error messages:

```
Intermediate Dataset-XML files written to directory = D:\eclipse-java-2018-09-win32-x86_64\eclipse\workspace\SDTM-ETL_4_1\temp\

Generating and cleaning define.xml - this may take some time ...
```

and some intermediate information and question messages, like:

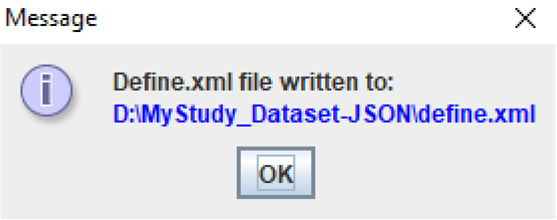


Is the define.xml in the context of a regulatory submission?

● Yes

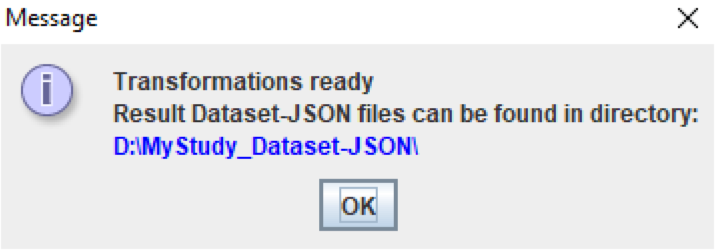
○ No

which is mostly important in the case of Define-XML 2.1.
Followed by:



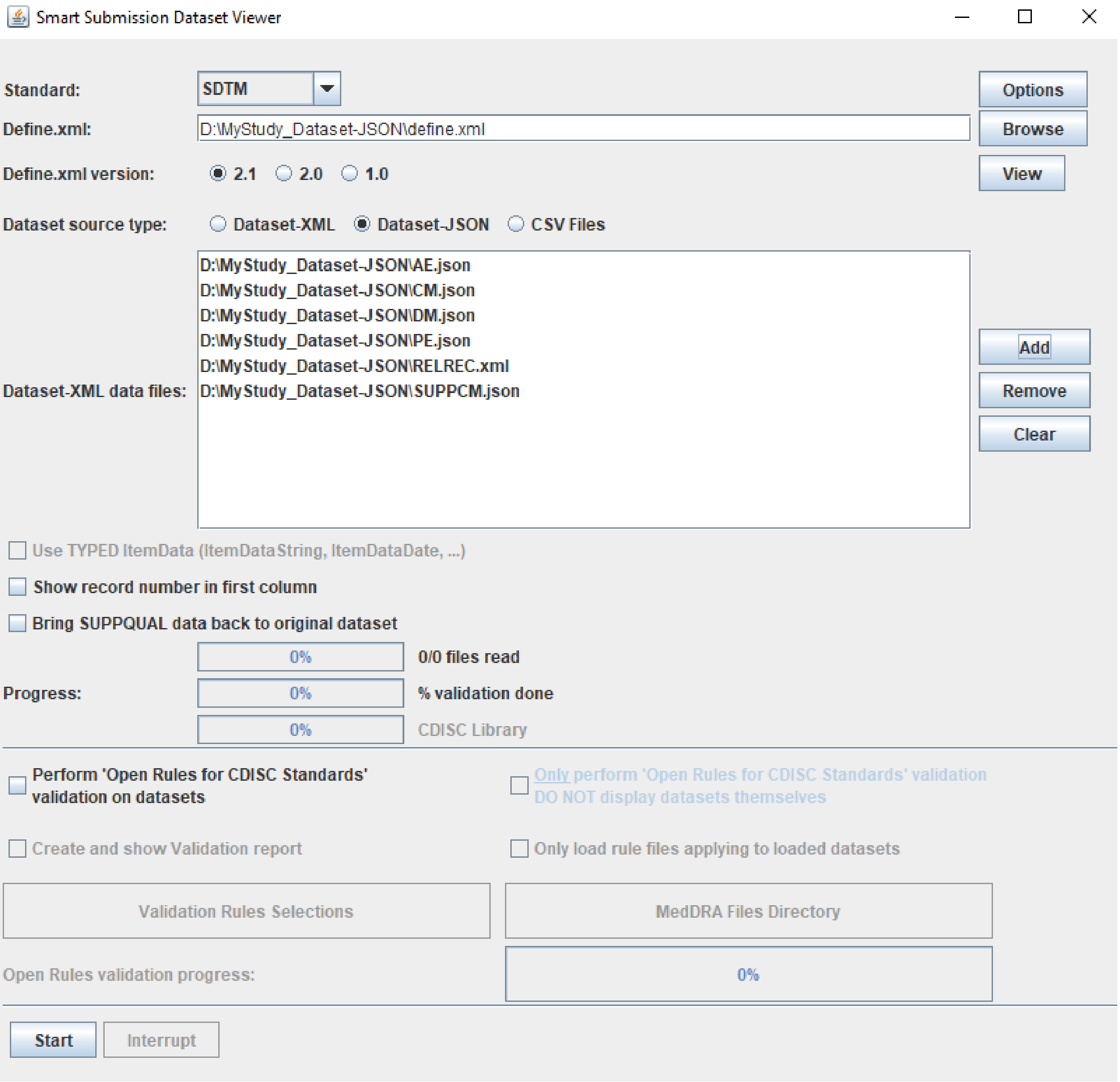
Define.xml file written to:
D:\MyStudy_Dataset-JSON\define.xml

and:



Transformations ready
Result Dataset-JSON files can be found in directory:
D:\MyStudy_Dataset-JSON\

The "Smart Submission Dataset Viewer" then opens with:



Smart Submission Dataset Viewer

Standard: SDTM

Define.xml: D:\MyStudy_Dataset-JSON\define.xml

Define.xml version: ● 2.1 ○ 2.0 ○ 1.0

Dataset source type: ○ Dataset-XML ● Dataset-JSON ○ CSV Files

Dataset-XML data files:
D:\MyStudy_Dataset-JSON\AE.json
D:\MyStudy_Dataset-JSON\CM.json
D:\MyStudy_Dataset-JSON\DM.json
D:\MyStudy_Dataset-JSON\PE.json
D:\MyStudy_Dataset-JSON\RELREC.xml
D:\MyStudy_Dataset-JSON\SUPPCM.json

□ Use TYPED ItemData (ItemDataString, ItemDataDate, ...)

□ Show record number in first column

□ Bring SUPPQUAL data back to original dataset

Progress:
0% 0/0 files read
0% % validation done
0% CDISC Library

□ Perform 'Open Rules for CDISC Standards' validation on datasets

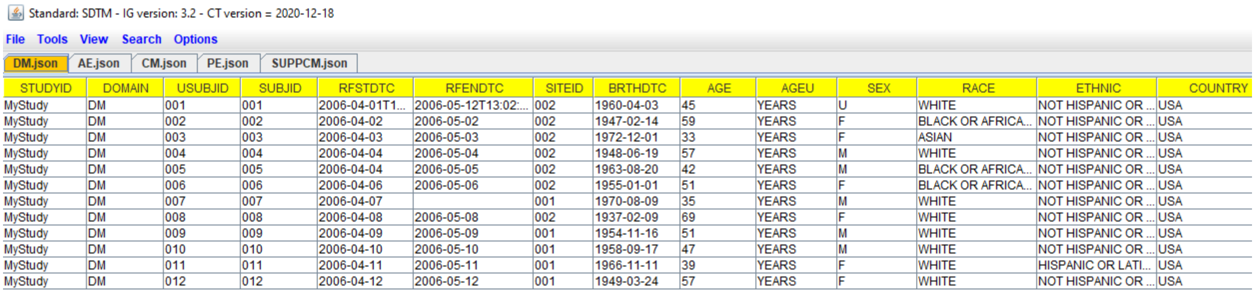
□ Only perform 'Open Rules for CDISC Standards' validation DO NOT display datasets themselves

□ Create and show Validation report

□ Only load rule files applying to loaded datasets

Open Rules validation progress: 0%

Listing the generated datasets and the location of the define.xml.
One can then switch on many of the "smart" features of the "Smart Submission Dataset Viewer" by clicking the "Options" button. During generation of the mappings and testing, this will however usually not be necessary.
What is often interesting however is to check the checkbox "Bring SUPPQUAL data back to original dataset", which will generate a "view" where SUPPQUAL records with NSVs (Non-Standard-Variables) are recombined with the original records.
Then clicking "Start" starts the generation and then leads to a set of tables like:



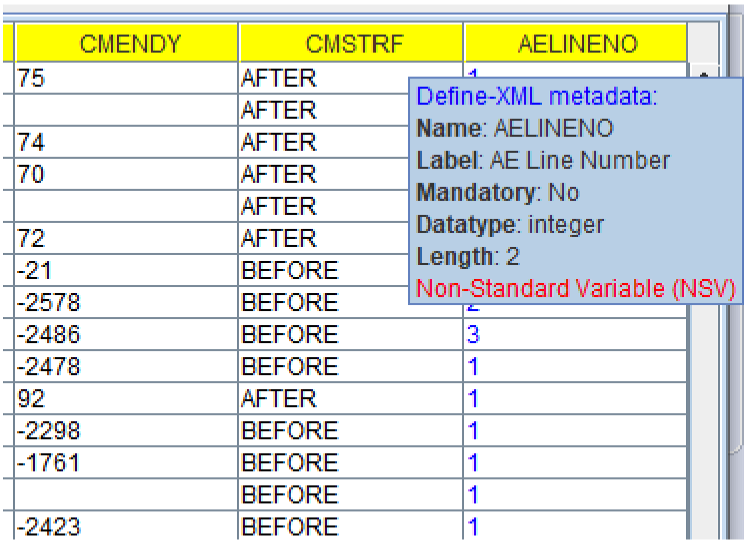
Standard: SDTM - IG version: 3.2 - CT version = 2020-12-18

STUDYID	DOMAIN	USUBJID	SUBJID	RFSTDTC	RFENDTC	SITEID	BRTHDTC	AGE	AGEU	SEX	RACE	ETHNIC	COUNTRY
MyStudy	DM	001	001	2006-04-01T1...	2006-05-12T13:02:...	002	1960-04-03	45	YEARS	U	WHITE	NOT HISPANIC OR ...	USA
MyStudy	DM	002	002	2006-04-02	2006-05-02	002	1947-02-14	59	YEARS	F	BLACK OR AFRICA...	NOT HISPANIC OR ...	USA
MyStudy	DM	003	003	2006-04-03	2006-05-03	002	1972-12-01	33	YEARS	F	ASIAN	NOT HISPANIC OR ...	USA
MyStudy	DM	004	004	2006-04-04	2006-05-04	002	1948-06-19	57	YEARS	M	WHITE	NOT HISPANIC OR ...	USA
MyStudy	DM	005	005	2006-04-04	2006-05-05	002	1963-08-20	42	YEARS	M	BLACK OR AFRICA...	NOT HISPANIC OR ...	USA
MyStudy	DM	006	006	2006-04-06	2006-05-06	002	1955-01-01	51	YEARS	F	BLACK OR AFRICA...	NOT HISPANIC OR ...	USA
MyStudy	DM	007	007	2006-04-07		001	1970-08-09	35	YEARS	M	WHITE	NOT HISPANIC OR ...	USA
MyStudy	DM	008	008	2006-04-08	2006-05-08	002	1937-02-09	69	YEARS	F	WHITE	NOT HISPANIC OR ...	USA
MyStudy	DM	009	009	2006-04-09	2006-05-09	001	1954-11-16	51	YEARS	M	WHITE	NOT HISPANIC OR ...	USA
MyStudy	DM	010	010	2006-04-10	2006-05-10	001	1958-09-17	47	YEARS	M	WHITE	NOT HISPANIC OR ...	USA
MyStudy	DM	011	011	2006-04-11	2006-05-11	001	1966-11-11	39	YEARS	F	WHITE	HISPANIC OR LATI...	USA
MyStudy	DM	012	012	2006-04-12	2006-05-12	001	1949-03-24	57	YEARS	F	WHITE	NOT HISPANIC OR ...	USA

Some of the features that are always interesting to use are:

- when selecting a row in a non-DM table, CTRL-D will select and show the corresponding DM record. Using CTRL-D one can than "toggle" between both the datasets
- when selecting a row in a SUPPxx dataset, CTRL-S will select and show the corresponding source record(s). Also here, one can the toggle between the two datasets by the use of CTRL-B.

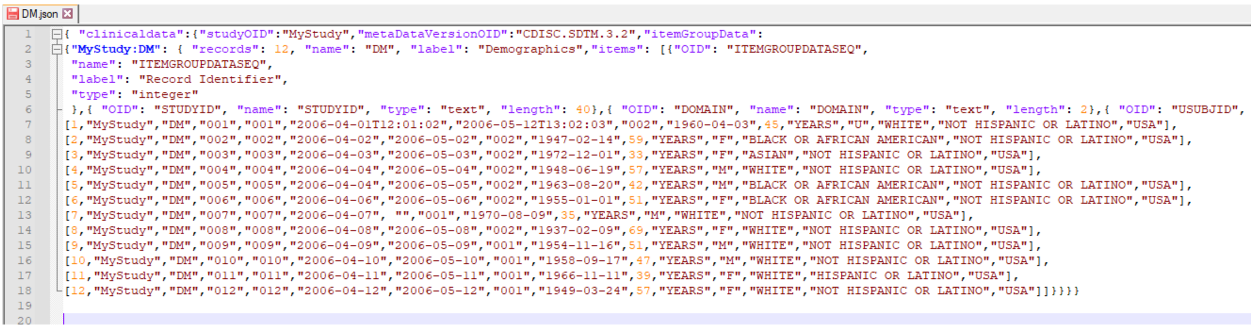
If we had checked the checkbox "Bring SUPPQUAL data back to original dataset", the view for the CM dataset would become:



CMENDY	CMSTRF	AELINENO
75	AFTER	1
	AFTER	
74	AFTER	
70	AFTER	
	AFTER	
72	AFTER	
-21	BEFORE	
-2578	BEFORE	2
-2486	BEFORE	3
-2478	BEFORE	1
92	AFTER	1
-2298	BEFORE	1
-1761	BEFORE	1
	BEFORE	1
-2423	BEFORE	1

Define-XML metadata:
Name: AELINENO
Label: AE Line Number
Mandatory: No
Datatype: integer
Length: 2
Non-Standard Variable (NSV)

The Dataset-JSON files themselves are very compact (file sizes considerably smaller than for SAS-XPT). For example, for DM.json:



```
{ "clinicaldata":{"studyOID":"MyStudy","metaDataVersionOID":"CDISC.SDTM.3.2","itemGroupData":
{"MyStudy:DM": { "records": 12, "name": "DM", "label": "Demographics","items": [{"OID": "ITEMGROUPDATASEQ",
 "name": "ITEMGROUPDATASEQ",
 "label": "Record Identifier",
 "type": "integer"
 },{ "OID": "STUDYID", "name": "STUDYID", "type": "text", "length": 40},{ "OID": "DOMAIN", "name": "DOMAIN", "type": "text", "length": 2},{ "OID": "USUBJID",
[1,"MyStudy","DM","001","001","2006-04-01T12:01:02","2006-05-12T13:02:03","002","1960-04-03",45,"YEARS","U","WHITE","NOT HISPANIC OR LATINO","USA"],
[2,"MyStudy","DM","002","002","2006-04-02","2006-05-02","002","1947-02-14",59,"YEARS","F","BLACK OR AFRICAN AMERICAN","NOT HISPANIC OR LATINO","USA"],
[3,"MyStudy","DM","003","003","2006-04-03","2006-05-03","002","1972-12-01",33,"YEARS","F","ASIAN","NOT HISPANIC OR LATINO","USA"],
[4,"MyStudy","DM","004","004","2006-04-04","2006-05-04","002","1948-06-19",57,"YEARS","M","WHITE","NOT HISPANIC OR LATINO","USA"],
[5,"MyStudy","DM","005","005","2006-04-04","2006-05-05","002","1963-08-20",42,"YEARS","M","BLACK OR AFRICAN AMERICAN","NOT HISPANIC OR LATINO","USA"],
[6,"MyStudy","DM","006","006","2006-04-06","2006-05-06","002","1955-01-01",51,"YEARS","F","BLACK OR AFRICAN AMERICAN","NOT HISPANIC OR LATINO","USA"],
[7,"MyStudy","DM","007","007","2006-04-07", "","001","1970-08-09",35,"YEARS","M","WHITE","NOT HISPANIC OR LATINO","USA"],
[8,"MyStudy","DM","008","008","2006-04-08","2006-05-08","002","1937-02-09",69,"YEARS","F","WHITE","NOT HISPANIC OR LATINO","USA"],
[9,"MyStudy","DM","009","009","2006-04-09","2006-05-09","001","1954-11-16",51,"YEARS","M","WHITE","NOT HISPANIC OR LATINO","USA"],
[10,"MyStudy","DM","010","010","2006-04-10","2006-05-10","001","1958-09-17",47,"YEARS","M","WHITE","NOT HISPANIC OR LATINO","USA"],
[11,"MyStudy","DM","011","011","2006-04-11","2006-05-11","001","1966-11-11",39,"YEARS","F","WHITE","HISPANIC OR LATINO","USA"],
[12,"MyStudy","DM","012","012","2006-04-12","2006-05-12","001","1949-03-24",57,"YEARS","F","WHITE","NOT HISPANIC OR LATINO","USA"]]}}}}
```

In future, once a "neutral zone" between sponsors and regulatory authorities has been established, such JSON structures (but not as "file") can be used, using RESTful web services, to populate the regulatory submission, even when the study is still in its starting phase and only a minor amount of the data has been collected.

It is of course a feature that we would LOVE to implement in future ...