Benefits of Using Excel File in CDISC SDTM Data Mapping

Hong Wang, Clinforce Consulting Inc, King of Prussia, PA

ABSTRACT

This paper discusses the benefits of excel file utilization in mapping clinical trial study data sets to Clinical Data interchange Standards Consortium(CDISC) Study Data Tabulation Model(SDTM) format. The SAS® code provided in this paper was developed with BASE SAS version 9.1 on PC and 8.2 on UNIX.

INTRODUCTION

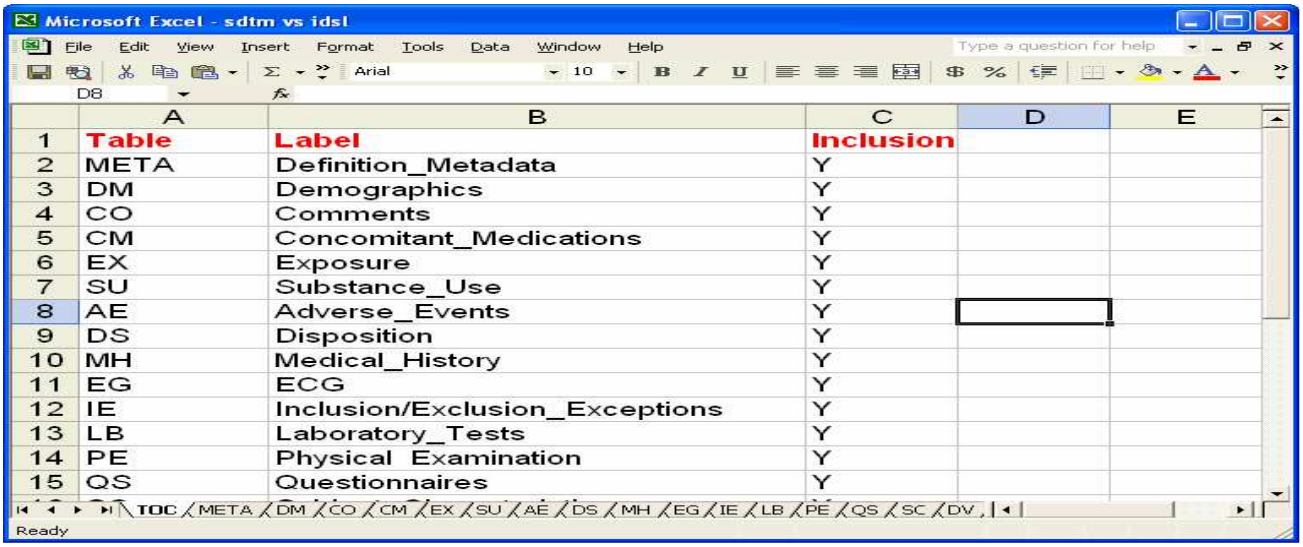
As some clinical trials already started and now are in the procedure of collecting clinical trial data, we couldn't redesign the standard CRF according to CDISC SDTM requirements. It is very important to develop a general tool for referencing SDTM metadata contents and attributes for all the projects to foster uniformity, evade repetitive hard-coding of attributes, and ensure the high quality in mapping clinical data to SDTM.

The master SDTM excel file can carry all the attributes of SDTM metadata. It can be generated according to the document in SDTM version 3.1.1 at www.cdisc.org. For each specific study, you can create your own study metadata selecting 'Y' in inclusion column and either export or Dynamic data exchange into Excel file by using SAS programming. From PC SAS version 8.2 or 9.1, read in the study SDTM multiple spreadsheets and pull all the spreadsheet metadata information into one whole SAS data set, then upload into the UNIX study working area.

From the UNIX SAS metadata, we can create the global macro variable for keeping variable (Req, Exp and some Perm variables) which should be kept in the SDTM domain data set and labeling variable attributes for each Variable.

SDTM MASTER METADATA EXCEL FILES

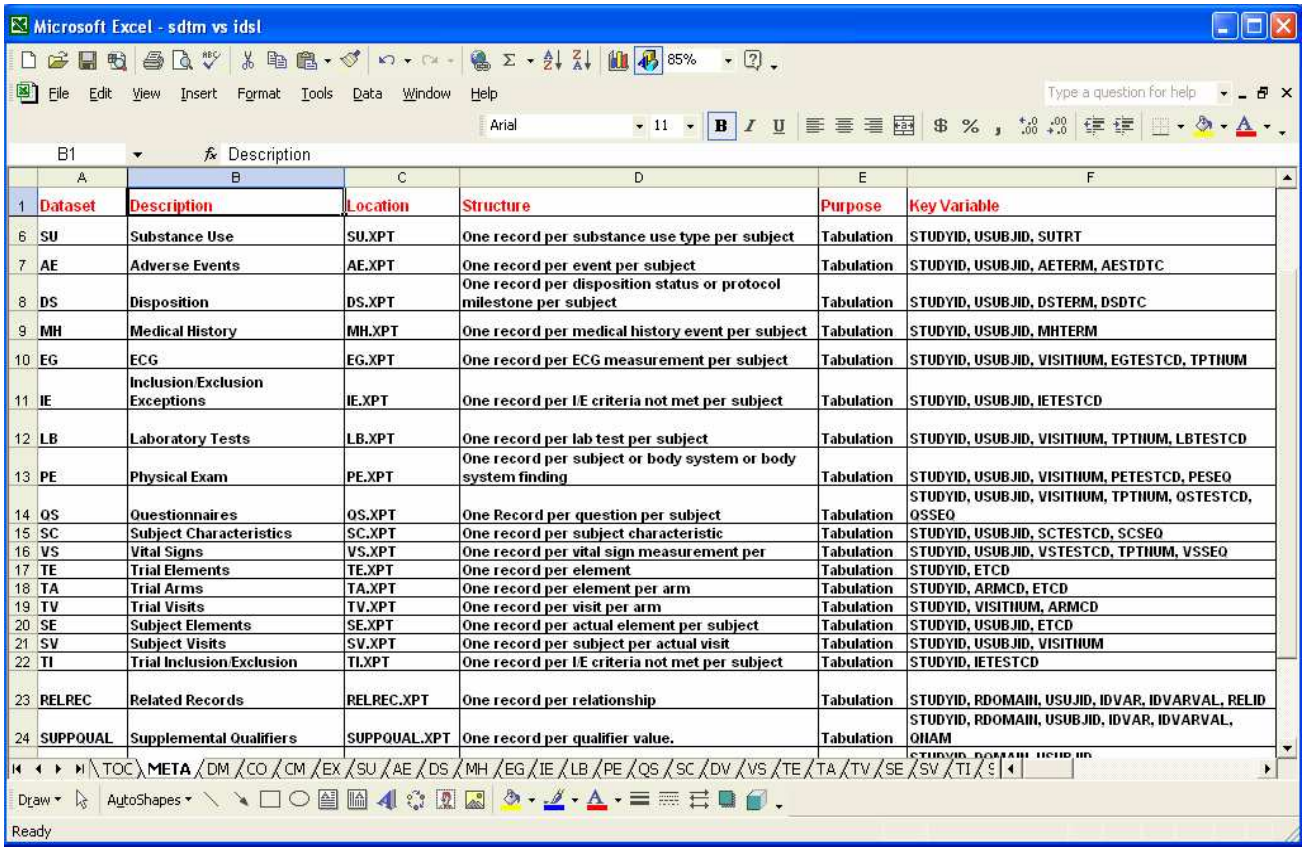
The master SDTM metadata was first entered into Excel. Each sheet of the master excel file has the additional column~ Inclusion(Y/N) when compared to the CDISC document. The reason for this added column is that variable Include(Y/N) can be reused by other users by just selecting 'Y' for their studies.
The sheet named TOC displayed three columns: Table, label and Inclusion(Y/N). See table 1



	A	B	C	D	E
1	Table	Label	Inclusion		
2	META	Definition_Metadata	Y		
3	DM	Demographics	Y		
4	CO	Comments	Y		
5	CM	Concomitant_Medications	Y		
6	EX	Exposure	Y		
7	SU	Substance_Use	Y		
8	AE	Adverse_Events	Y		
9	DS	Disposition	Y		
10	MH	Medical_History	Y		
11	EG	ECG	Y		
12	IE	Inclusion/Exclusion_Exceptions	Y		
13	LB	Laboratory_Tests	Y		
14	PE	Physical Examination	Y		
15	QS	Questionnaires	Y		

Table 1 Table of Content for master SDTM

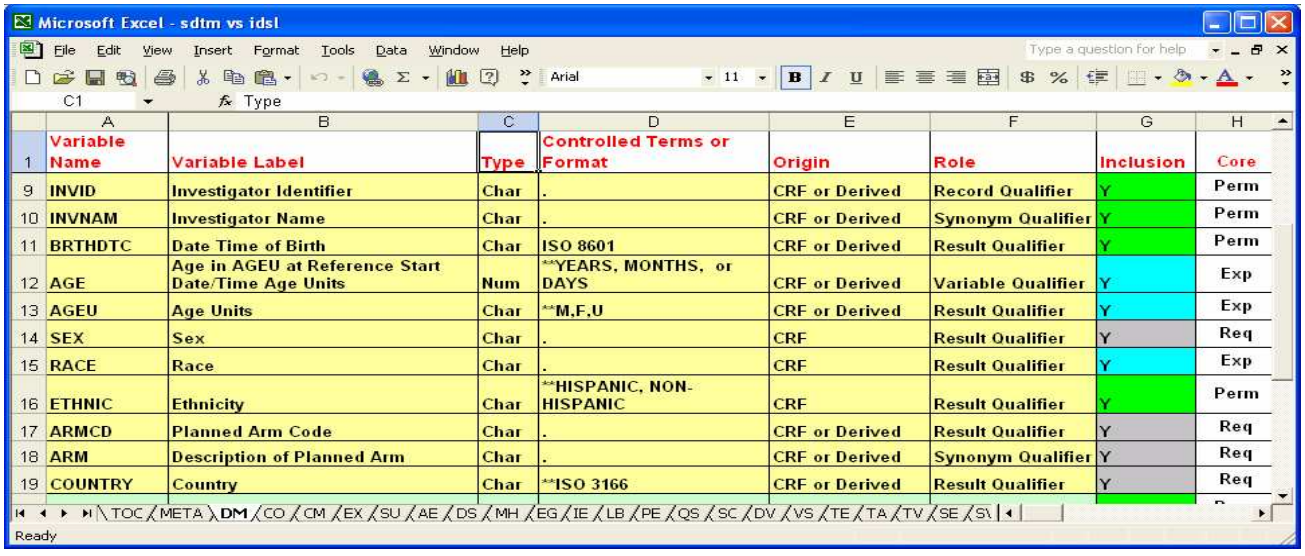
The second sheet named META display dataset name, Description, location, Structure, Purpose, key Variable and Inclusion See Table 2



	A	B	C	D	E	F
1	Dataset	Description	Location	Structure	Purpose	Key Variable
6	SU	Substance Use	SU.XPT	One record per substance use type per subject	Tabulation	STUDYID, USUBJID, SUTRT
7	AE	Adverse Events	AE.XPT	One record per event per subject	Tabulation	STUDYID, USUBJID, AETERM, AESTDTC
8	DS	Disposition	DS.XPT	One record per disposition status or protocol milestone per subject	Tabulation	STUDYID, USUBJID, DSTERM, DSDTC
9	MH	Medical History	MH.XPT	One record per medical history event per subject	Tabulation	STUDYID, USUBJID, MHTERM
10	EG	ECG	EG.XPT	One record per ECG measurement per subject	Tabulation	STUDYID, USUBJID, VISITNUM, EGTESTCD, TPTNUM
11	IE	Inclusion/Exclusion Exceptions	IE.XPT	One record per I/E criteria not met per subject	Tabulation	STUDYID, USUBJID, IETESTCD
12	LB	Laboratory Tests	LB.XPT	One record per lab test per subject	Tabulation	STUDYID, USUBJID, VISITNUM, TPTNUM, LBTESTCD
13	PE	Physical Exam	PE.XPT	One record per subject or body system or body system finding	Tabulation	STUDYID, USUBJID, VISITNUM, PETESTCD, PESEQ
14	QS	Questionnaires	QS.XPT	One Record per question per subject	Tabulation	STUDYID, USUBJID, VISITNUM, TPTNUM, QSTESTCD, QSSEQ
15	SC	Subject Characteristics	SC.XPT	One record per subject characteristic	Tabulation	STUDYID, USUBJID, SCTESTCD, SCSEQ
16	VS	Vital Signs	VS.XPT	One record per vital sign measurement per	Tabulation	STUDYID, USUBJID, VSTESTCD, TPTNUM, VSSEQ
17	TE	Trial Elements	TE.XPT	One record per element	Tabulation	STUDYID, ETCD
18	TA	Trial Arms	TA.XPT	One record per element per arm	Tabulation	STUDYID, ARMCD, ETCD
19	TV	Trial Visits	TV.XPT	One record per visit per arm	Tabulation	STUDYID, VISITNUM, ARMCD
20	SE	Subject Elements	SE.XPT	One record per actual element per subject	Tabulation	STUDYID, USUBJID, ETCD
21	SV	Subject Visits	SV.XPT	One record per subject per actual visit	Tabulation	STUDYID, USUBJID, VISITNUM
22	TI	Trial Inclusion/Exclusion	TI.XPT	One record per I/E criteria not met per subject	Tabulation	STUDYID, IETESTCD
23	RELREC	Related Records	RELREC.XPT	One record per relationship	Tabulation	STUDYID, RDOMAIN, USUJID, IDVAR, IDVARVAL, RELID
24	SUPPQUAL	Supplemental Qualifiers	SUPPQUAL.XPT	One record per qualifier value.	Tabulation	STUDYID, RDOMAIN, USUBJID, IDVAR, IDVARVAL, QNAM

Table 2 META sheet

The other sheets named as their domain names display the columns, the same as the CDISC SDTM version 3.1. See Table 3.



	A	B	C	D	E	F	G	H
1	Variable Name	Variable Label	Type	Controlled Terms or Format	Origin	Role	Inclusion	Core
9	INVID	Investigator Identifier	Char	.	CRF or Derived	Record Qualifier	Y	Perm
10	INVNAM	Investigator Name	Char	.	CRF or Derived	Synonym Qualifier	Y	Perm
11	BRTHDTC	Date Time of Birth	Char	ISO 8601	CRF or Derived	Result Qualifier	Y	Perm
12	AGE	Age in AGEU at Reference Start Date/Time Age Units	Num	**YEARS, MONTHS, or DAYS	CRF or Derived	Variable Qualifier	Y	Exp
13	AGEU	Age Units	Char	**M,F,U	CRF or Derived	Result Qualifier	Y	Exp
14	SEX	Sex	Char	.	CRF	Result Qualifier	Y	Req
15	RACE	Race	Char	.	CRF	Result Qualifier	Y	Exp
16	ETHNIC	Ethnicity	Char	**HISPANIC, NON-HISPANIC	CRF	Result Qualifier	Y	Perm
17	ARMCD	Planned Arm Code	Char	.	CRF or Derived	Result Qualifier	Y	Req
18	ARM	Description of Planned Arm	Char	.	CRF or Derived	Synonym Qualifier	Y	Req
19	COUNTRY	Country	Char	**ISO 3166	CRF or Derived	Result Qualifier	Y	Req

Table 3 DM Domain of master SDTM

SDTM STUDY METADATA EXCEL FILES

Procedures for creating your own study metadata file:
First reading the multiple data sets into multiple SAS data sets;
Delete the inclusion equal N records and drop the inclusion variable;
Then outputting the data sets into another excel which stands for study metadata file.

```
libname sds "C:\Documents and Settings\hw69253\Desktop";

/*SAS code for reading sheets TOC and META and DDE to Excel file*/

%macro getexcel(shname=,getn=, out=);
PROC IMPORT OUT= &out(where=(inclusion='Y'))
            DATAFILE= "C:\Documents and Settings\hw69253\Desktop\sdtm vs idsl.xls"
            DBMS=EXCEL REPLACE;
            SHEET="&shname";
            GETNAMES=&getn;
            MIXED=YES;
            SCANTEXT=YES;
            USEDATE=YES;
            SCANTIME=YES;
RUN;

%mend;

%getexcel(shname='TOC$',getn=YES, out=TOC);
%getexcel(shname='META$',getn=YES, out=META);

data toc0;
Table='Table';
Label='Label';
run;

data Toc1;
length table $10. label $40.;
format table $6.;
set toc0 toc(drop=inclusion);
run;

filename extoc DDE 'excel|TOC!r1c1: r30c2' notab;
data _null_;
set toc1;
file extoc;
put Table "09"X Label "09"X;
run;

data meta0;
length dataset $8. Description_ $25. Location $8. Structure $40. Purpose $10.
Key_Variable $50.;
dataset='Dataset';
Description_='Description';
location='Location';
Structure='Structure';
Purpose='Purpose';
Key_Variable ='Key Variable';
run;
```

```
data meta1;
format dataset $8. location $8.;
set meta0 meta;
run;

filename exmeta DDE 'excel|META!r1c1: r30c6' notab;
data _null_;
set meta1;
file exmeta;
put Dataset "09"X  Description_ "09"X  Location "09"X  Structure "09"X
Purpose "09"X Key_Variable "09"X ;
run;

/*SAS Code for reading each individual sheet and DDE to Excel file*/
%macro set;

data tot; set toc;
  if _n_=1 then delete;
run;

data _null_;
   set tot end=eof;
   call symput("sheet"||left(put(_n_,8.)), trim(left(table)));
   if eof then call symput('totn', _n_);
run;

%do i=1 %to &totn;
  %getexcel(shname='&&sheet&i$',getn=YES, out=&&sheet&i);

  data dat0;
   length variable_name $14. variable_label $40. Type $8.
Controlled_Terms_or_Format $25. Origin $20. Role $20. Core $4.;
  variable_name='Variable Name';
  Variable_label='Variable Label';
  Type='Type';
  Controlled_Terms_or_Format='Controlled Terms or Format';
  Origin='Origin';
  Role='Role';
  Core='Core';
run;

data &&sheet&i..1;
 format  Controlled_Terms_or_Format $25. Origin $20. Role $20. Core $4.;
 set dat0 &&sheet&i;
 run;

 %if &&sheet&i =SUPPQUAL %then %do;

  filename ex DDE "excel|&&sheet&i!r1c1: r100c7" notab;
  data _null_;
  set &&sheet&i..1 ;
  file ex;
  put variable_name "09"X  variable_label "09"X  Type "09"X
Controlled_Terms_or_Format "09"X  Origin "09"X Role "09"X  Core "09"X ;
run;
```

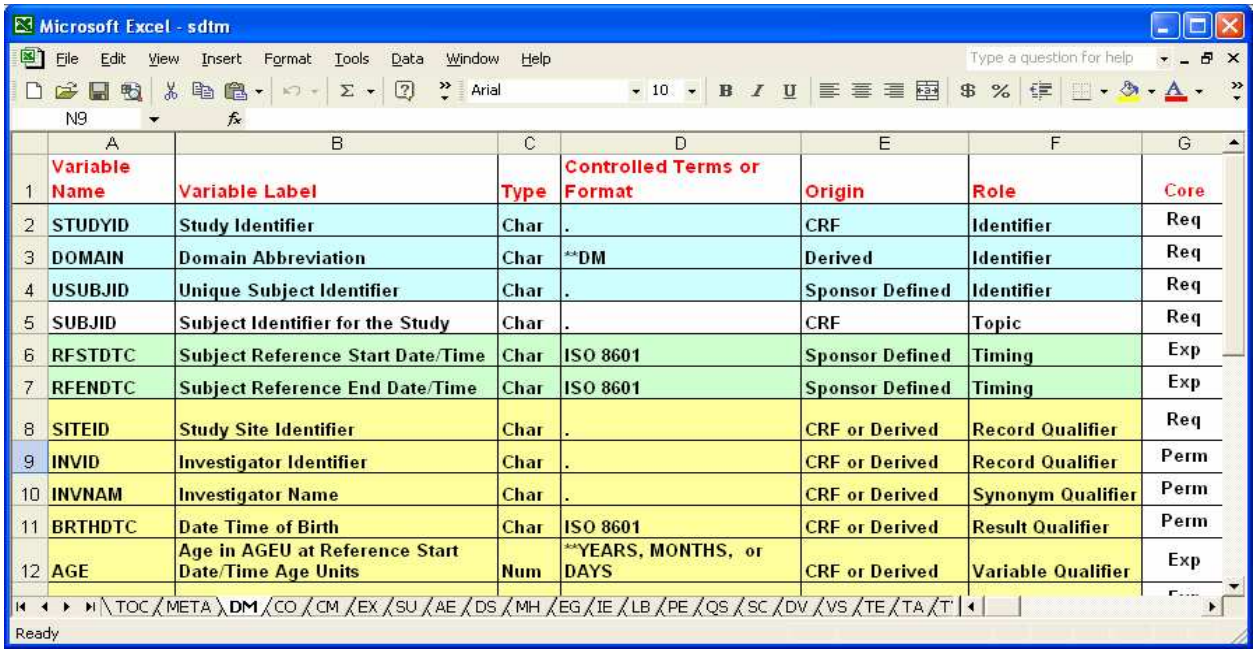
```
  %end;
  %else %do;

  filename ex&&sheet&i DDE "excel|&&sheet&i!r1c1: r100c7" notab;
  data _null_;
  set &&sheet&i..1 ;
  file ex&&sheet&i;
  put variable_name "09"X  variable_label "09"X  Type "09"X
Controlled_Terms_or_Format "09"X  Origin "09"X Role "09"X  Core "09"X ;
  run;

 %end;
%end;

%mend;
%set;
```

Table 4 shows the each domain attributes of your study.



	A	B	C	D	E	F	G
1	Variable Name	Variable Label	Type	Controlled Terms or Format	Origin	Role	Core
2	STUDYID	Study Identifier	Char	.	CRF	Identifier	Req
3	DOMAIN	Domain Abbreviation	Char	**DM	Derived	Identifier	Req
4	USUBJID	Unique Subject Identifier	Char	.	Sponsor Defined	Identifier	Req
5	SUBJID	Subject Identifier for the Study	Char	.	CRF	Topic	Req
6	RFSTDTC	Subject Reference Start Date/Time	Char	ISO 8601	Sponsor Defined	Timing	Exp
7	RFENDTC	Subject Reference End Date/Time	Char	ISO 8601	Sponsor Defined	Timing	Exp
8	SITEID	Study Site Identifier	Char	.	CRF or Derived	Record Qualifier	Req
9	INVID	Investigator Identifier	Char	.	CRF or Derived	Record Qualifier	Perm
10	INVNAM	Investigator Name	Char	.	CRF or Derived	Synonym Qualifier	Perm
11	BRTHDTC	Date Time of Birth	Char	ISO 8601	CRF or Derived	Result Qualifier	Perm
12	AGE	Age in AGEU at Reference Start Date/Time Age Units	Num	**YEARS, MONTHS, or DAYS	CRF or Derived	Variable Qualifier	Exp

Table 4 DM Domain of Study SDTM

Transform Excel file into SAS data.

When we get the study data set, read the multiple excel sheet metadata into SAS data in PC SAS 9.1 and upload to UNIX. See the data view as Table 5.

```
/*upload the attrib data to unix through remote connect*/
options comamid=tcp;
%let unixnode=xxxxx;
options remote=unixnode;
filename rlink 'C:\Program Files\SAS\SAS 9.1\connect\SASlink\tcpunix.scr';
signon;
rsubmit;
libname unixlib v8 'YYYYYYYYY'
proc upload data=sdtm out=unixlib.sdtm;
run;
endrsubmit;
signoff;
```

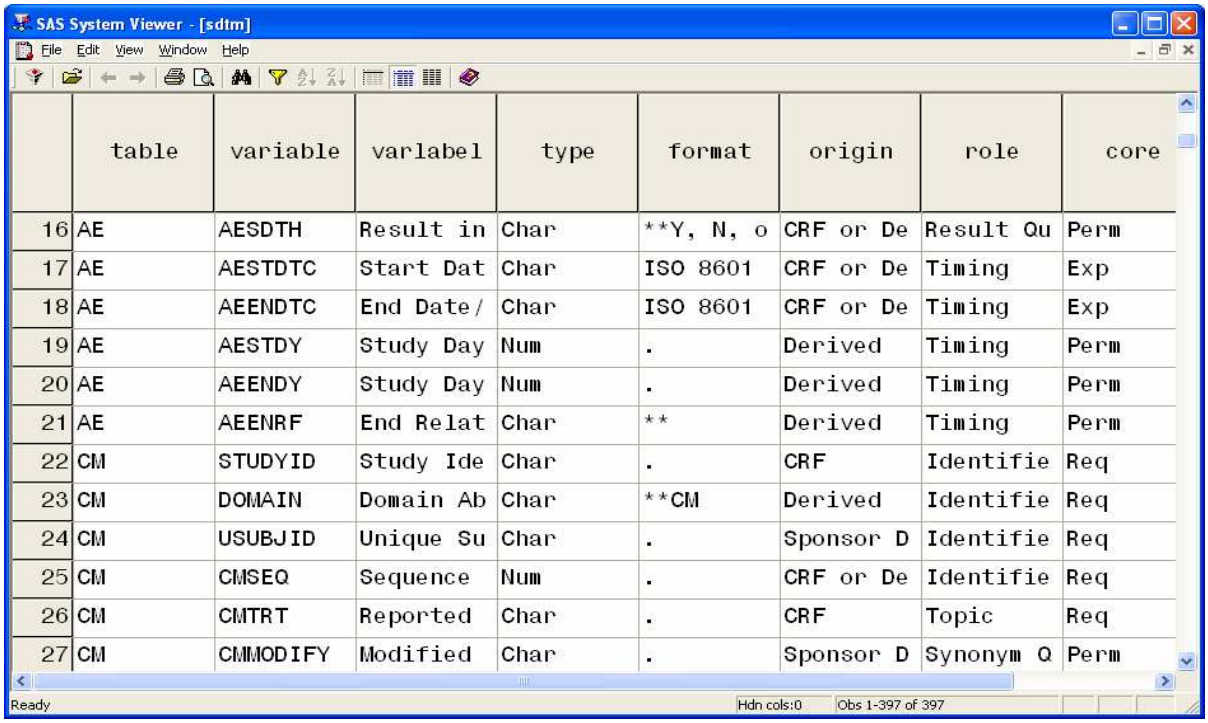


	table	variable	varlabel	type	format	origin	role	core
16	AE	AESDTH	Result in	Char	**Y, N, o	CRF or De	Result Qu	Perm
17	AE	AESTDTC	Start Dat	Char	ISO 8601	CRF or De	Timing	Exp
18	AE	AEENDTC	End Date/	Char	ISO 8601	CRF or De	Timing	Exp
19	AE	AESTDY	Study Day	Num	.	Derived	Timing	Perm
20	AE	AEENDY	Study Day	Num	.	Derived	Timing	Perm
21	AE	AEENRF	End Relat	Char	**	Derived	Timing	Perm
22	CM	STUDYID	Study Ide	Char	.	CRF	Identifie	Req
23	CM	DOMAIN	Domain Ab	Char	**CM	Derived	Identifie	Req
24	CM	USUBJID	Unique Su	Char	.	Sponsor D	Identifie	Req
25	CM	CMSEQ	Sequence	Num	.	CRF or De	Identifie	Req
26	CM	CMTRT	Reported	Char	.	CRF	Topic	Req
27	CM	CMMODIFY	Modified	Char	.	Sponsor D	Synonym Q	Perm

Table 5 One SAS data set pulled all the domain datasets

Based on this SAS data set, we can generate the two global variables, &keepvar and &labelvar.

GLOBAL VARIABLE &KEEPVAR

Macro variable &keepvar was generated from the above SAS data set by calling b_keepvar.sas The purpose is to make sure that Req, Exp and some Perm variables are in final mapped to SDTM data sets.

```
%macro b_keepvar;

/*get the domain data you worked on */
  data sdtm;set sdtm.sdtm;
   if table="&domain";
 run;

/*get the count of variables in the specified domain*/
proc sql;
 select count(distinct variable) into: n
from sdtm;

/*transpose the variables in horizontal*/
proc transpose data=sdtm out=sdtm_t;
 var variable;
run;

/*put all the column variable into macro variable*/
  %do i=1 %to &n ;
  proc sql;
  select Col&i into: c&i
  from sdtm_t;
 %end


 /*create the global variable keepvar and assign the initial value as blank*/
%global keepvar;
```

```
/*using the do loop to get the variables listed in the global variable*/
%do i=1 %to &n ;
%let Keepvar=&keepvar &&c&i;
%end;
%put keepvar=&keepvar;

%mend;
```

call %b_keepvar in your mapping code b_ex.sas

```
%let domain=EX;
%b_keepvar;
data sdtm.ex(keep=&keepvar);
  set ex;
run;
```

keepvar =OMAIN USUBJID EXSEQ EXTRT EXDOSE EXDOSU EXDOSFRM EXDOSFRQ EXDOSTOT EXROUTE EXLOT EXSTDTC EXENDTC EXSTDY EXENDY EXTPT EXTPTNUM

From log window, you can see the keepvar macro variable carrying all the needed variable from SDTM data.

GLOBAL VARIABLE &LABELVAR

Retrieve the variable labels from the above SAS metadata and generated the global macro variable &labelvar by calling b_labelvar.sas.

```
%macro b_labelvar;
  data sdtm;set sdtm.sdtm;
   if table="&domain";
    l=variable || '=' ||"'" || trim(varlabel) || "'";
 run;
 proc sql;
  select count(distinct variable) into: n
from sdtm;
data _null_;
 set sdtm;
  %do i=1 %to &n;
    if _n_=&i then do;
    call symput("t"||trim(left(_n_)),trim(left(l)));
  end;
 %end;
run;
%global labelvar;
 %do i=1 %to &n;
    %let labelvar=&labelvar &&t&i;
%end;
%put   labelvar=&labelvar;
%mend;
```

In your mapping code, call b_labelvar.sas and then assign the variable labels in final SDTM mapped data.

```
%let domain=EX;
%b_labelvar;
data sdtm.ex;
  set ex;
  label &labelvar;
run;
```

&labelvar equals to
STUDYID ='Study Identifier' DOMAIN ='Domain Abbreviation' USUBJID ='Unique Subject Identifier' EXSEQ ='Sequence Number' EXTRT ='Reported Name of Drug, Med, or Therapy' EXDOSE ='Dose per Administration' EXDOSU ='Dose Units' EXDOSFRM='Dose Form' EXDOSFRQ='Dosing Frequency Per Interval' EXDOSTOT='Total Daily Dose using EXDOSU' EXROUTE ='Route of Administration' EXLOT ='Lot Number' EXSTDTC ='Start Date/Time of Treatment' EXENDTC ='End Date/Time of Treatment' EXSTDY ='Study Day of Start of Treatment' EXENDY ='Study Day of End of Treatment' EXTPT ='Planned Time Point Name' EXTPTNUM='Planned Time Point Number'

You will see it will label all the variables in the final SDTM data

CONCLUSION

SDTM master file metadata excel files can be easily changed into study SDTM files by using SAS. This prevents hard coding or manually entering data which causes man-made errors and thus provides quality. The use of these two macros b_keepvar.sas and b_labelvar.sas helps to decrease the boring typing work and errors occur which often happen in SAS 'keep' and 'label' statements. These two macros are not only simple, but also provide accurate information from SDTM study metadata. It will guarantee that required and expected variables of SDTM meta are in the final mapped SDTM data as you put all these variables in your metadata Excel file. Users don't need to QC each study excel metadata file as it comes from the master files which can be used by many programmers. One only needs one time to QC master file.

The master SDTM excel file can be applied across many projects as it is the standard metadata. The only thing the users need to do is to select their study metadata. It is easy for users to track the variables and save QC time. The excel file application provides a smart and helpful tool in catching the SDTM metadata information and makes the mapping SDTM data work easily and correctly.

TRADEMARK

ACKNOWLEDGEMENTS

The author has special thanks to Dr. Shi-Tao Yeh for review and comments and thanks to Dr. Scott Patterson for the project to work on.

CONTACT INFORMATION

Hong Wang, M.S.
Clinforce, Inc
2202 Renaissance,
King of Prussia, PA
(484)322-0604 x 237 / (610)787-3692
hong.4.wang@gsk.com