PhUSE US Connect 2019

Paper PP06

Stay relaxed... An Utility to auto create specification!!

Kaveri Pachiappan, Symbiance Inc, New Jersey, USA

ABSTRACT

In the Pharmaceutical industry, specifications are core which mandates for all programming activities. Currently, SDTM and ADAM specifications are configured manually by mapping the raw data to the standard domains. As our conventional process is time consuming and it would require expertise inputs. This paper proposes an automatic and efficient approach to configure certain parts of SDTM specifications using SAS macros. This enables the auto-population of mapping rules into the specification, significantly reducing the manual work and make our deliverable on time with higher quality. The utility has the standardized metadata repository, which directs the creation of standard derivation and mapping in the specification. Rules or Derivations could be based on CDISC standards and Sponsor defined mappings. It will also facilitate Use and Re-Use of existing/previous Study based on rules in metadata. This will reduce a considerable amount of time and manpower involved which in turn reduces the manual errors.

INTRODUCTION

Specification generation process is to integrate/map data for different studies of a compound into a common integrated database. The source data can be any standard, also sponsor defined standards with different therapeutic area. By using SAS® macro templates and metadata, we can generate the SDTM spec without doing repetitive work for spec writing process.

METADATA CREATION

Metadata creation is a one-time process and this repository is an increasingly important aspect of resource discovery. Standardized metadata has the potential to increase discovery and it also supports reusability of standard derivations, automation of business processes and full data traceability. In this repository it's all about metadata, including transformational data, can be managed and governed centrally, which facilitates us for a faster and consistent specification generation. Metadata repository is created based on:

- Data collection based on prior study information
- Different versions of SDTM IG.
- Sponsor specific Data standards.
- Metadata can be updated for New information/scenarios.

SDTM SPECIFICATION:

SDTM specification configuration is the first and very important step in the SDTM development. The SDTM domains will be created based on the algorithms defined in the SDTM specifications. The attributes of the variables in SDTM domains are also defined in the SDTM specifications. All the variables present in the specification would comply with the SDTM IG.

TRADITIONAL APPROACH TO CREATE DATASET SPECIFICATION

The creation of clinical SDTM dataset specification is generally a manual mapping of raw datasets to the standard domains. This method requires a greater number of resources, expert inputs which will consume time as its needed to develop each specification from scratch through manual review and interpretation of standards and specification created earlier from other studies. Also, the mapping specification for each dataset would comply with the SDTM IG standards.

For example, to create SDTM mapping specification, statistical programmer and/or statistician starts with reviewing raw datasets with the annotated CRF and dataset standards. The author manually reconciles differences and drafts the study dataset specification. The spec authors must understand the layout of the data and each variable present in all the source dataset and with the SDTM standards, then only they can able to draft the specification. So, it requires expert inputs who have knowledge on standards and latest requirements to create the SDTM specification.

ENHANCED PROCESS

This enhanced process places a metadata driven flow as the driving force. Metadata repository has been created to map source informations into standard target data structures. There is always a certain percentage of source to target mapping that can be applied as from one study to another. Standards and metadata artifacts created and used throughout this lifecycle must then be managed in a centralized repository to facilitate their automated reuse. This paper specifically describes a metadata-driven data flow use for a dynamic and automated platform which integrates clinical data from multiple sources of raw datasets based on metadata managed and governed in a repository and generates SDTM specification.

For example, the standard domains within a study for a given therapeutic area, there is a limited percentage of variation to be customized from one study to another, whereas in custom domains, the level of customizing the derivation will be more than the standard domains, however there is still a fair amount of reuse expected to be leveraged for studies within the same therapeutic area. Once this metadata is created it can be reused for all the studies. The mapping logic and derivations would be similar for SDTM specifications for all the studies.

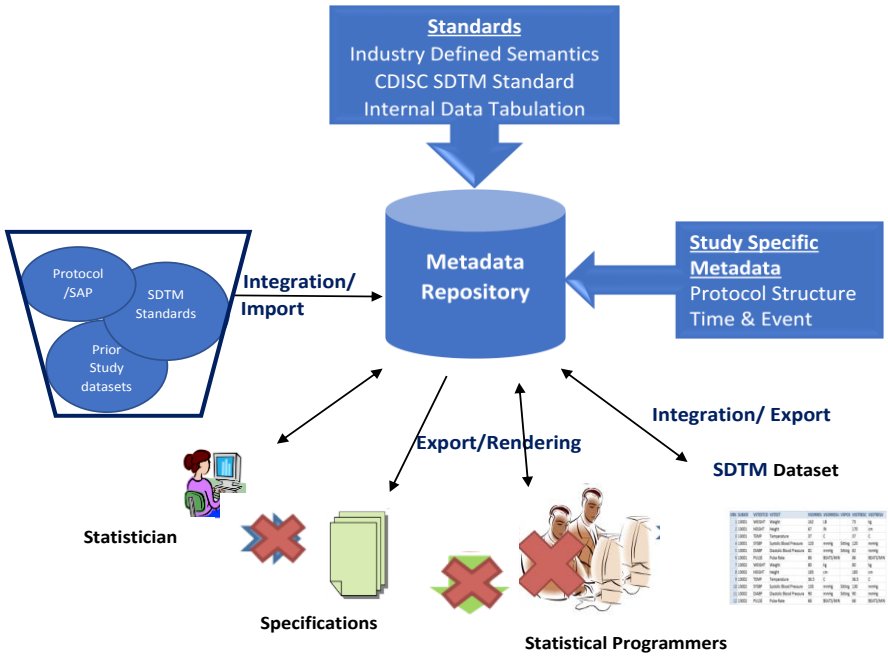

Figure. 1: Enhanced process with metadata repository

This reuse reduces considerable time consumption and effort which eventually becomes cost efficient. Mapping libraries can be maintained at various levels such as Global, Therapeutic Area and study specific. Thus, rather than the statistical programmer or statistician manually interpreting the CRF, raw datasets, SDTM IG standards and prior study specifications, one has all those definitions and standards already created in a centralized repository. So, re-use can be applicable for existing content to create a new set of metadata to represent the study specific datasets and specifications. Only adding items for structural and transformational metadata that are required to be customized for that specific study. By this approach any resource in an organization can be able to generate specification which in turn reduces manual error and deliverables could be delivered within time and metadata approach will maintain specification consistency across all over the studies.

'*%M_SPECGEN'* macro which crowds up the basic frame of SDTM mapping with the use of standard metadata, which can reduce the manual work for SAS programmers or in some case it can entirely handle some simple domains without any further modifications. This will significantly increase the efficiency and speed of the SDTM conversion process. The following sectors will present the procedural details about the development of this macro.

GENERAL SETUP

%M_SPECGEN macro reads all the raw datasets one after another whichever present in the raw library and it is used to compare with metadata which is created in Microsoft excel spreadsheet and imported as SAS® dataset. For each dataset, there will be a metadata which needs to be mapped under its respective domain. Each source dataset is mapped to its target domain and each source variables is mapped to its target variables by using the metadata repository. This repository can be updated only for new therapeutic area/new scenarios and later that can also be used for future purpose. This macro will check the information of study structure against the metadata repository.

SOURCE DATA

To read source datasets - The raw dataset variables from each dataset can be captured using PROC CONTENTS. It is only necessary to keep the variables MEMNAME and NAME which store the dataset name and variable names respectively.

Below code illustrates how raw dataset metadata is captured through CONTENTS procedure.

```
proc contents data =raw._All_ out=tdconts(keep=memname name) noprint;
run;

proc sort data =tdconts out=tdconts_s nodupkey;
      by memname name;
run;

proc sql noprint;
      select distinct count(memname) into: lvacnt from tdconts_s;
      select memname into:lvars1-:lvars%trim(%left(&lvacnt)) from tdconts_s;
run;
quit;

%put &lvacnt &lvars1;
```

DATASET	VARIABLE ID	VARIABLE	LABEL	DATA TYPE	FORMAT
AECDISC	3	SUBJID	Subject ID Number	num	
AECDISC	4	AESEQ	AE Sequence	num	
AECDISC	5	AE	Adverse Experience, Literal	char	
AECDISC	6	STRT_DD1	Onset Date Day (DD)	char	
AECDISC	7	STRT_DD2	Onset Date Month(MON)	char	
AECDISC	8	STRT_YY	Onset Date Year(YYYY)	char	
AECDISC	9	STRT_DT	Onset Date, Derived	num	DATE9.
AECDISC	10	STOP_DD1	End Date Day (DD)	char	
AECDISC	11	STOP_DD2	End Date Month (MON)	char	
AECDISC	12	STOP_YY	End Date Year(YYYY)	char	
AECDISC	13	STOP_DT	End Date, Derived	num	DATE9.

Figure. 2: Source dataset

STANDARD METADATA:

Once the standard metadata is created, it can be reused for generating the mapping spec across all studies. The metadata is prepared based on each SDTM IG version, sponsor data collection and based on rules or derivations for comment column. So, it will maintain the consistency across all the studies. This macro will

1. Check all raw informations against all standard domains
2. Create specification Excel
    a. Write domain description (domains/key variables / sort order).
    b. Write pool metadata and study metadata + derivations to Excel.

Class	Domain	Variable Name	Variable Name General	Variable Label	Type	Length	Format	Role	Core	Mandatory
Events	AE	USUBJID	USUBJID	Unique Subject Identifier	Char	24		Identifier	Req	Yes
Events	AE	AESEQ	SEQ	Sequence Number	Num	8		Identifier	Req	Yes
Events	AE	AESPID	SPID	Sponsor-Defined Identifier	Char	3		Sponsor Defined	Perm	No
Events	AE	AETERM	TERM	Reported Term for the Adverse Event	Char	70		Topic	Req	Yes
Events	AE	AESTDTC	STDTC	Start Date/Time of Adverse Event	Char	16	ISO 8601	Timing	Exp	No
Events	AE	AEENDTC	ENDTC	End Date/Time of Adverse Event	Char	16	ISO 8601	Timing	Exp	No
Events	AE	AESTDY	STDY	Study Day of Start of Adverse Event	Num	8		Timing	Perm	No
Events	AE	AEENDY	ENDY	Study Day of End of Adverse Event	Num	8		Timing	Perm	No
Events	AE	AESER	SER	Serious Event	Char	1	Y, N	Record Qualifier	Exp	No
Events	AE	AEACN	ACN	Action Taken with Study Treatment	Char	40		Record Qualifier	Exp	No
Events	AE	AEREL	REL	Causality	Char	20		Record Qualifier	Exp	No
Events	AE	AEOUT	OUT	Outcome of Adverse Event	Char	30		Record Qualifier	Perm	No

Figure. 3: Standard metadata repository

The metadata repository has several variables which includes,

CLASS: Events is one of the general observational class which incidents independent of planned study evaluations occurring during the trial.

DOMAIN: Each domain dataset is distinguished by a unique, two-character code that should be used consistently throughout the submission.

VARIABLENAME: The Variable Name (limited to 8 characters for compatibility with the SAS® Transport format)

VARIABLENAME GENERAL: General variable name without the prefix two-character domain code.

LABEL: A descriptive Variable Label, using up to 40 characters, which should be unique for each variable in the dataset

TYPE: The data Type (e.g., whether the variable value is a character or numeric)

ROLE: The Role of the variable, which determines how the variable is used in the dataset.

CORE VARIABLES:

REQUIRED – These variables are necessary for the proper functioning of standard software tools used by reviewers. They must be included in the data set structure and should not have a missing value for any observation.

EXPECTED – These variables form the fundamental core of information within a domain. They must be included in the data set structure; however, it is permissible to have missing values.

PERMISSIBLE – These variables are not a required part of the domain and they should not be included in the data set structure if the information they were designed to contain was not collected.

In addition to the variables shown in figure.3 SDTM IG version, type of CRF collection (sponsor defined), controlled terminology terms and formats also be presented in the metadata.

CALLING PROCEDURE FOR M_SPECGEN UTILTY

```
%m_specgen (ipdata    =          /*** Source data library ***/
            Metapath  =         /*** Path for metadata file ***/
            Whrcond   =         /*** Metadata, where to get type of collection ***/
            Lookupt   =         /*** Target look up table location ***/
            Outpath   =         /*** Path, where to generate mapping spec ***/
              Sdtmig  =        /*** Specifies the SDTM IG version which need to use ***/
);
```

MACRO PARAMETERS TO BE PASSED BY THE USER

IPDATA: Input data path, where the raw datasets are available to generate SDTM specification.
METAPATH: Path for the standard metadata repository which has been created
WHRCOND: To specify the sponsor defined type of CRF collection. we can also fetch the metadata based on sponsor type as well.
LOOKUPT: Path for the lookup table to be created for each domain.
OUTPATH: Output path where the final SDTM specification must be stored.
SDTMIG: To specify the SDTM IG version which needs to follow to generate dataset.

The '%M_SPECGEN' macro will

- Fetch all the datasets from the raw library and metadata file from metapath
- Compare the variables in Source Dataset and Metadata using '*%m_specgen*' macro.
- Where statement in macro filters the type of data collection which is based on each therapeutic area under company standards.
- Create target domain table with matched variables and its transformation.
- Check and update the generated specification for the values which is not available in the Metadata.

SOURCE TO TARGET TRANSFORMATION (SDTM STRUCTURE)

The term TARGET in this paper refers to the SDTM model that identifies which source information can fit to its corresponding SDTM structure. Then the author needs to Identify the variables for each SDTM domain. Referencing the standard metadata and having a list of all the possible mapping variables for each domain, for each sponsor can be verified in the look up table. For each SDTM domain, a separate lookup table will be created which has source to target domain information and the mapping derivations.

TABLE: Lookup.Ae

Source_Dataset	Source_Variable	Source_Label	Data_Type	Length	Format	Target_Domain	Target_Variable	SDTM_Label	Mapping_Logic
AECDISC	AE	Adverse Experience, Literal	Char	70		AE	AETERM	Reported Term for the Adverse Event	If AETERM missing then delete from the dataset
AECDISC	STRT_DD1	Onset Date Day (DD)	Char	16	ISO 8601	AE	AESTDTC	Start Date/Time of Adverse Event	If not missing(STRT_DD1, STRT_DD2, STRT_YY) then concatenate and convert into ISO 8601 format.
AECDISC	STRT_DD2	Onset Date Month (MON)	Char	16	ISO 8601	AE	AESTDTC	Start Date/Time of Adverse Event	as above
AECDISC	STRT_YY	Onset Date Year(YYYY)	Char	16	ISO 8601	AE	AESTDTC	Start Date/Time of Adverse Event	as above
AECDISC	STRT_DT	Onset Date Derived		.					
AECDISC	STOP_DD1	End Date Day (DD)	Char	16	ISO 8601	AE	AEENDTC	End Date/Time of Adverse Event	If not missing(STOP_DD1, STOP_DD2, STOP_YY) then concatenate and convert into ISO 8601 format.
AECDISC	STOP_DD2	End Date Month (MON)	Char	16	ISO 8601	AE	AEENDTC	End Date/Time of Adverse Event	as above
AECDISC	STOP_YY	End Date Year(YYYY)	Char	16	ISO 8601	AE	AEENDTC	End Date/Time of Adverse Event	as above

Figure. 4: AE look up table (source to target information)

Figure.4, identifies the source dataset AECDISC is mapped to AE domain and the source date and time variables STRT_DD1, STRT_DD2, STRT_YY, STRT_DT will be concatenated to get the standard variable AESTDTC under AE domain. The corresponding mapping logic and attributes will be fetched from the metadata repository. The list of all possible domains and all possible variables within each domain can be created from the SDTM documentation, the review and identification of what must be included for a study has to be done by a project member who is familiar with the study and with the SDTM requirements. Once completed it can be reused for additional studies as well. Even the mid-level programmer will be able to generate specification.

MAPPING SPECIFICATION DOCUMENT

In general, Mapping specification document is a multi-tabbed single EXCEL spreadsheet which is utilized to create a data set specification, displaying one domain (i.e.SAS data set) per sheet, one row for each distinct variable and columns for variable attributes and mapping logic or comments.

ODS TAGSETS. EXCELXP

To use ODS tagsets. excelxp to create multi-tabbed Excel files that contains specification, for each page one domain will be displayed and with the use of report procedure one can limit the variables which are needed to be display in specification. Report procedure has powerful formatting, summarizing, and analysis capabilities that made it easier to create anything from a simple listing to a very complex report. In the report itself one can keep the required variables in specification among the variables present in lookup table.

Below code illustrates spec generation.

```
ods _all_ close;
ods tagsets.ExcelXP path="&outpath" file="SDTM_mapping_spec.XML"
style= XLsansPrinter;

/***** report procedure *****/

proc report data=&lookupt..&&ds&i nowd headline headskip split='*' missing;
   column var1 var2 var3 var4;
   define var1     / display "Variable Name";
   define var2     / display "Variable Label";
   define var3     / display "Type";
   define var4     / display "Length";
run;
quit;

ods tagsets.ExcelXP close;
```

- The first ODS statement closes all destinations.
- The second ODS statement uses the ExcelXP tagset to generate the XML output and then store the output in a file. XML extension should be used instead of XLSX or XLSX because Excel 2007 and later it displays a warning if the XML extension is not used. The STYLE option controls the appearance of the output, such as the font and color scheme.
- The report procedure will print all the variables defined and will place it into the XML file created.
- The last ODS statement closes the ExcelXP destination and releases the XML file so that it can be opened with Excel.

FINAL AUTOMATED SPECIFICATION

Mapping specification will be created using Microsoft spreadsheet, for each SDTM domain, a separate mapping file will be created in the output location. The spec contains the informations like target variable name, variable label, length, type, its controlled terminology, origin, role of the variable, core information, from which dataset the mapping file created, source variable and its mapping logic or comments for programmers to generate SDTM dataset. With this overall approach we can able to complete 80% of the specification, remaining columns need to fill manually.

Variable Name	Variable Label	Type	Length	Controlled Terms or Format	Origin	Role	Core	From Dataset	Variable Name or Fixed Value	Mapping Logic	General Comments / Questions
USUBJID	Unique Subject Identifier	Char	24		Sponsor Defined	Identifier	Req	AECDISC	ID	Create Subject ID in a combination of Studyid and subjectid	
AESEQ	Sequence Number	Num	8		Derived	Identifier	Req			Unique identification of records per subject. Sort by USUSBJID, AESTDTC, AEBODSYS, and AETERM	
AETERM	Reported Term for the Adverse Event	Char	70			Topic	Req	AECDISC	AE	If AETERM missing then delete from the dataset	
AESTDTC	Start Date/Time of Adverse Event	Char	16	ISO 8601		Timing	Exp	AECDISC	STRT_DD1, STRT_DD2, STRT_YY		
AEENDTC	End Date/Time of Adverse Event	Char	16	ISO 8601		Timing	Exp	AECDISC	STOP_DD1, STOP_DD2, STOP_YY		
AESTDY	Study Day of Start of Adverse Event	Num	8		Derived	Timing	Perm	SDTM.DM, SDTM.AE	DM.RFSTDTC, AE.AESTDTC	For RFSTDTC and AESTDTC not missing: If <RFSTDTC then AESTDTC-RFSTDTC; If AESTDTC>=RFSTDTC then AESTDTC-RFSTDTC+1	
AEENDY	Study Day of End of Adverse Event	Num	8		Derived	Timing	Perm	SDTM.DM, SDTM.AE	DM.RFSTDTC, AE.AEENDTC	For AEENDTC and RFSTDTC not missing: If AEENDTC<RFSTDTC then AEENDTC-RFSTDTC; If AEENDTC>=RFSTDTC then AEENDTC-RFSTDTC+1	

Figure. 5: Generated mapping specification for AE domain

Figure. 5, is the automated mapping specification for AE using macro *%M_SPECGEN*. In this attrib and mapping logic will be fetched from the standard metadata file. The date variables AESTDTC & AEENDTC in SDTM should be in ISO8601 format as per SDTM IG. __SDTY and __ENDY is the study start and end day of adverse event and this logic will be common for all the variables which has __SDTY and __ENDY.

ADVANTAGES

- Provide high level overviews of domains or raw datasets and ensures that all raw variables are mapped accurately which eliminates most of the manual work.
- Implementation of spec generation with metadata repository will help to significantly reduce cost and condense development timelines.
- This method of spec generation promotes reusability, process automation and end-to-end data traceability.
- Allow the specification to adopt a uniform structure across all studies. It will also facilitate Use and Re-Use of existing/previous study information based on rules in metadata.
- Reduce a considerable amount of time and manpower involved which in turn reduces the manual errors.
- When raw data has same or similar to SDTM variable names, this process can save time ensuring that the raw variable data fits correctly into the SDTM variable with the same name.

FUTURE AMENDMENT

Currently this macro tool is used generate SDTM specification with the standard metadata. This can also be implemented to create specification for ADAM datasets when we get similar kind of studies or for most of the simple domain mapping might be done in future. The concept is same as to have one sheet per ADAM dataset that will contain the domain structure and derivation rules for the derived variables.

CONCLUSION

Creating auto spec generation for the development of dataset conversion programs is an essential part of the SDTM/ADAM submission. SDTM specifications using *'%M_SPECGEN'* macro can eliminate most of the manual and repetitive work from the spec writing process. This Overall automation process produces more efficient, higher quality, and less error prone specifications. SAS macros also provide a data overview (look-up table) to ensure that raw variables are mapped, variables conform to their expected format, and enforce a uniform structure throughout the specification.

REFERENCES

1. Judith Goud, Priya Shetty. (2016), "How a Metadata Repository enables dynamism and automation in SDTM-like dataset generation" https://www.lexjansen.com/phuse/2016/dh/DH05.pdf

2. Foundational standards of CDISC, from protocol to analysis reporting, along with data exchange and controlled terminology. To add a few acronyms (CDASH, SDTM, ADAM) to your vocabulary. https://www.cdisc.org/standards/foundational/sdtm

3. SAS Institute Inc. (2008). SAS Component Language 9.2: Reference. Cary, NC: SAS Institute Inc

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kaveri Pachaiappan
Symbiance Inc. 51 Everette Dr,
Suite B-20, Princeton Junction
New Jersey 08550, USA
Email: kpachiappan@symbiance.com
Web: www.symbiance.com