

Automating SDTM – Get it prepared with your voice

Akhil Vijayan, Genpro Life Sciences, Thiruvananthapuram, India
Anish George, Genpro Life Sciences, Thiruvananthapuram, India
Anoop Ambika, Genpro Life Sciences, Thiruvananthapuram, India

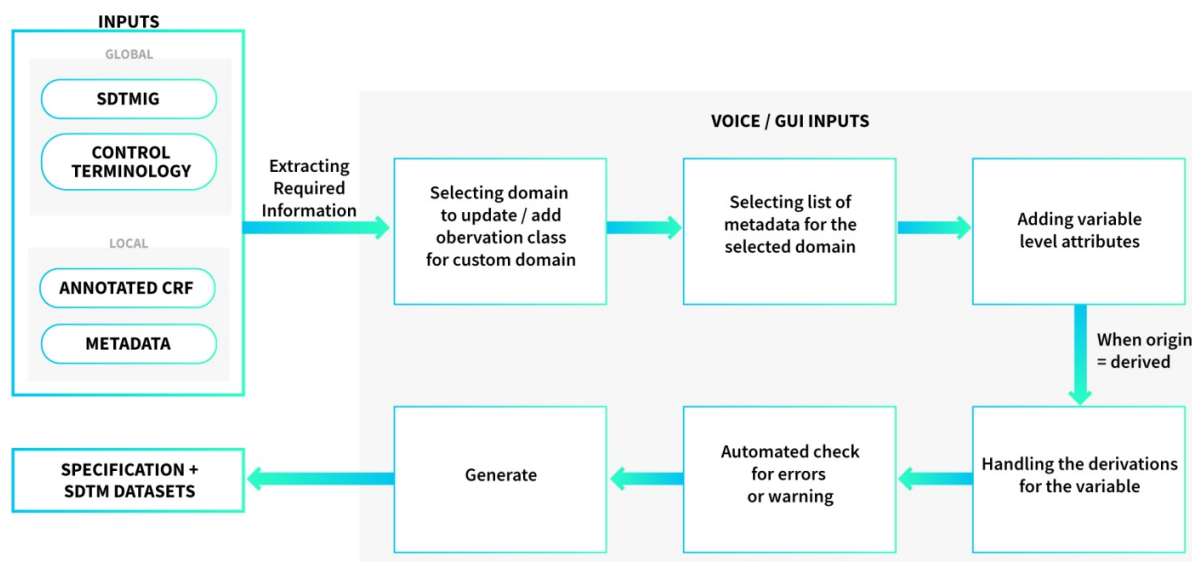
ABSTRACT

FDA mandates that data submissions for all studies conducted after 2016 must be in SDTM format. Majority of the sponsors are now looking at ways to generate datasets in a cost-effective manner. The process starting CRF annotations to the generation of the SDTM dataset is a long cycle comprising human effort and quality issues that follow. This paper introduces “**YESDTM**” a tool which generates the SDTM specification and datasets in a shorter time span using NLP and AI algorithms. It will take the annotated CRF and the raw datasets as input and then generate the SDTM Specification and datasets using guided voice/typed commands.

Once the user submits the study-specific information, the system performs a two-stage process. It captures and stores the domains and variable names from annotated CRF. The system will then load all the data and variables from the raw data and identifies the direct possible mappings automatically. The advantages of the system includes, less SDTM cycle time, improved efficiency, voice and text interfaces and the automatic mapping of all possible data points.

INTRODUCTION

Quality and Time are the two major factors that define efficient clinical research. Development of high quality SDTM datasets can be time consuming when implementing the QC checks at different phases of the development. Automation of the entire process can reduce the cycle time in developing high quality SDTM datasets. This paper portrays the working and methodology of the tool ‘YESDTM’ that aids in SDTM automation. The paper assumes the readers to have a basic understanding about SDTM.

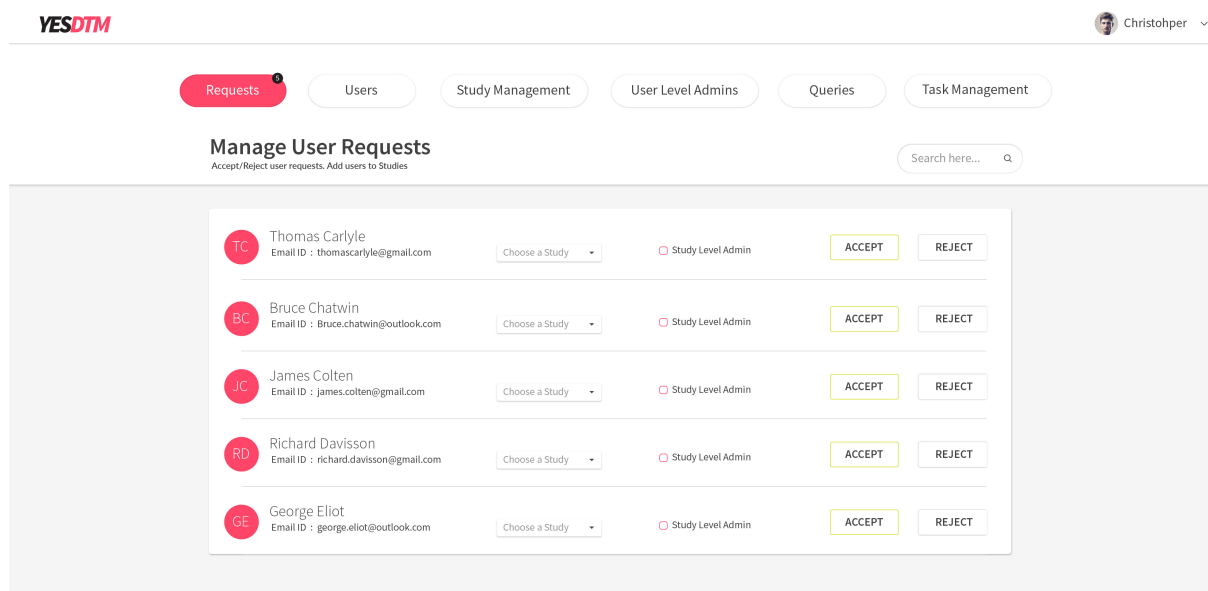


YESDTM requires certain inputs to generate the SDTM dataset for each study. The inputs are the annotated CRF, raw datasets, SDTM IG and the Control Terminology. Based on the annotated CRF and the SDTM IG, YESDTM will finalize the list of domains and the variables for the study. Users can add/remove variables based on the requirement. When the domains identified from the CRF are not in the SDTM IG, YESDTM will identify it as a custom domain.

PhUSE EU Connect 2018

DESIGN OF YESDTM

YESDTM functionalities are focused on the data accessibility, trial/user management and specification generation. Security of the information is an important factor in clinical trial and the responsible persons will not allow all the users of the study to access the entire things. There are some user level restrictions at every stage. As part of the security, access to the system is restricted using login credentials and three levels of user roles are allowed in YESDTM; the Super administrator, Study Level administrator and the trial level user. Super administrator has complete access to the application who will then define the privileges for the rest of the users. There are multiple tasks that the super administrator can perform which is depicted in the following figure.



It is difficult to populate all the information for generating the Define XML and datasets in a single module. Separate modules are designed in YESDTM to capture the inputs for Define XML and datasets such as the module for populate the domain level information, variable level information etc.

INPUTS FOR YESDTM

The inputs required for the working of YESDTM are categorized into two, local and global. Global inputs consist of the latest Study Data Tabulation Model Implementation Guide (SDTMIG) which is the backbone for the system and the CDISC SDTM Controlled Terminology. User can reuse the global inputs in other studies. PDF and the excel version of CT is available in the web, so extracting the information from CT is not a big deal. For extracting the information from SDTM IG the system seeks the help of artificial intelligence, with the application of Machine Comprehension and NLP modules YESDTM can read the unstructured PDFs and convert them into a structured format. After the completion of the extraction YESDTM will save the domains information in the central repository and the user needs to modify the input files based on the available versions.

SDTM annotated CRF and the raw datasets are considered as the local inputs in YESDTM. Once user submits the study specific information, the system requires two step processes to complete the initial phase. First system reads the annotated CRF to capture the annotated domains and variables names and stores it in a separate memory. Similar to the SDTM IG annotations in the aCRF is in a tabular format and extracting the information from the CRF is also achieve with the use of artificial intelligence. When the category of the variable is "required" or expected" SDTM will present those variables in the domains regardless of the variable annotation in the CRF.

Next system will start processing with the loaded raw data. YESDTM will load all the data and identifies the variables name similar to the SDTM IG variables. The system will use this information to give suggestion to the user at the time of spec development. After the fulfillment of the two-stage process, the framework will recognize the list of domains identified for the study. User can add or remove domains from the identified list.

The picture below shows an example of the annotation in the CRF.

PhUSE EU Connect 2018

DM=Demographics

Form: Demographics

Date of Birth	BRTHDTC		①
Age	AGE/AGEU		Fixed Unit: years ②
Sex	SEX	Female ☐ ③ Male ☐	
Ethnicity	ETHNIC	Hispanic or Latino ☐ ④ Not Hispanic or Latino ☐ Not Reported ☐ Unknown ☐	

Next system will start processing with the loaded raw data. The system will load all the data and identifies the variables name similar to the SDTM IG variables. The system will use this information to give suggestion to the user at the time of spec development. After the fulfillment of the two-stage process, the framework will recognize the list of domains identified for the study. User can add or remove domains from the identified list.

WORKING OF YESDTM

The development stage of YESDTM is spitted into three the extraction of data from pdf, voice interaction with YESDTM and generation of the specification. Python modules are available for extracting the tabular information from pdf. Extraction is performed based on vertical and horizontal lines and their intersecting points on the page as cell-separators. We can specify the pages we need to extract when calling the extraction method, which includes an option to extract tables from all pages. After the execution of the code it will return the tabular data and this can be converted to the required format- only thing is we have to specify the key / identification variables for identifying the domain. Key variables can be any value from the tabular data. For example if we specify data going to LB when it contains 'LBTESTCD'. Conversion of the result data to any format is possible.

Voice recognition is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT). The capacity to make reliable client encounters with voice connections requires a comprehension of how individuals normally speak with their voices and the basics of voice collaboration. Sometimes the system require 'training' which means an individual speaker reads text or isolated vocabulary into the system and is called as "speaker dependent" speech recognition systems.

For each study YESDTM expecting study level information to start processing. The administrators need to add such information like the title of the study, protocol ID and the inputs. YESDTM will have the list of domains once the user submit the aCRF into YESDTM .The system is organized to present the identified domain in the task management module. Administrators can update the task management module. Updates includes add or remove identified domain, assign task for users etc. Domain generation is beginning once the user selects the assigned domains from the task list. After selecting a domain user will lead to next module where YESDTM expecting the domain level information from the user. YESDTM will save the domain information in the repository and it will populate automatically these information for the next time . In the CRF annotation if the variable name is begin with SUPP YESDTM identifies it as a supplementary domain and this will also presented in the current module. For the rest of the cases user can add supplementary domains.

PhUSE EU Connect 2018

YESDTM

Christopher

Requests

Users

Study Management

User Level Admins

Queries

Task Management

Assign Task Here..

Domain Type	Domain Name	Title	Specs Author	Specification Target Completion Date	Specs Actual Completion Date	Status	Specs Reviewer
Events	AE	Adverse Events	Select User	Set Target Date	--/--	---	Select User
Interventions	EX	Exposure	Select User	Set Target Date	--/--	---	Select User
Special Domains	DM	Demographics	Select User	Set Target Date	--/--	---	Select User
Interventions	EC	Exposure as Collected	Select User	Set Target Date	--/--	---	Select User
Special Purpose	SV	Subject Visits	Select User	Set Target Date	--/--	---	Select User

ADD DOMAIN

Once user submit the domain level information YESDTM lead the user to the main module where the user need to add the variable level information for the selected domain.

YESDTM

Press MIC to add value in Origin

Christopher

< Study 1 / SV

SDTM IG : 3.2 | CT VERSION : 300CT2018

Task List

aCRF

Datasets

Queries

Dictionary

Order	Variable	Label	Data Type	Length	Significant Digits	Mandatory	Codelist	Origin	Pages	Method	Predecessor	Role	Comment
1	STUDYID	Study Identifier	text			Yes		Assigned				Identifier	ADD
2	DOMAIN	Domain Abbreviation	text			Yes	DOMAIN	Assigned				Identifier	ADD
3	USUBJID	Unique Subject Identifier	text			Yes		Derived		ADD		Identifier	ADD
4	VISITNUM	Visit Number	float	2		Yes	VISITNUM	Assigned				Topic	ADD
5	VISIT	Visit Name	text			No	VISIT	Assigned				Timing	ADD
6	VISITDY	Planned Study Day of Visit	integer			No		Derived		ADD		Timing	REMOVE
7	EPOCH	Epoch	text			No	EPOCH	Derived		ADD		Timing	REMOVE
8	SVSTDT	Start Date/Time of Visit	text			No	ISO8601	CRF	24			Timing	ADD
9	SVENDTC	End Date/Time of Visit	text			No	ISO8601	CRF	24			Timing	ADD
10	SVSTDY	Study Day of Start of Visit	integer			No						Timing	REMOVE

ADD VARIABLE

VALIDATE

GENERATE

Inputs from user are limited when the selected domain is from the SDTM IG otherwise user need to add the relevant details. 'Variable', 'label' like fields are ruled in its backend, user can enter up to 8 digits in variable column, 40 in length column, 3 possible entries are allowed in the data type column etc. YESDTM will suggest adding EPOCH in all the finding domains.

The role variable is directly mapped from the SDTM implementation guide. YESDTM will automatically populate values for 'variable', 'label', 'mandatory', 'date types' and 'roles' when the selected domain is standard. User can add/remove permissible variables from the domain.

How YESDTM populate variable 'Length': Significant digit variable is only used when the data type is 'float'. The length of the variable will be populated automatically at the time of generating the data. For all the data type is populated as 'text' YESDTM will use the maximum length from the variable and populate that as the length after execution.

How YESDTM populate the filed 'Mandatory': Three categories of variables are specified in the "Core" column of the domain models in the SDTM IG for measuring the compliance. When the core variable is "Required" or

PhUSE EU Connect 2018

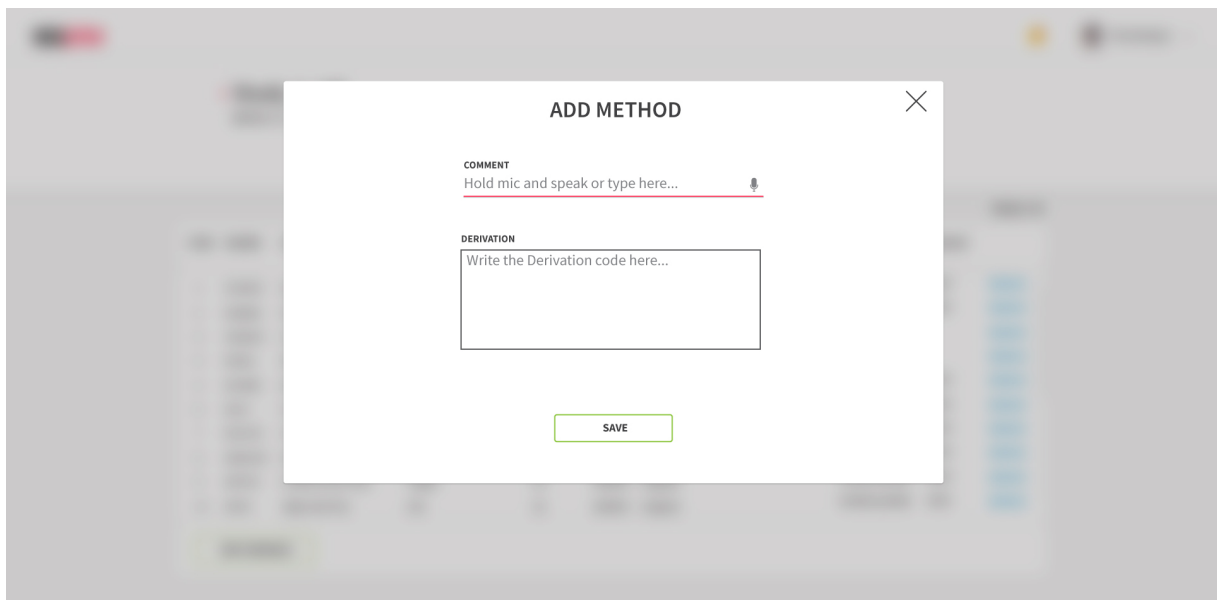
“Expected” YESDTM populate the value “Yes” in Mandatory and populate “No” when the core variable is “Permissible”.

For the variable names ending with “DTC”, YESDTM identify it as a date variable and populate the value as “IS8601” in codelist and the corresponding details are populated in the ‘Dictionaries’. Users can view/edit dictionaries within the dictionary module. For the rest of the variables, user need to assign only the codelist name.

How YESDTM populate ‘codelist’: YESDTM expecting the name of the “CDISC Submission Value” as the codelist name when codelist exist in the CT. After the generation of the data YESDTM will compare the data value with the selected control terminology value based on the codelist name. User can view/update this in the codelist module. For the variable name such as DOMAIN, VISIT, VISITNUM and EPOCH, variable names ending with TEST and TESTCD etc, YESDTM will suggest codelist names appropriately from the SDTM IG.

Since there only 3 possible values as the origin YESDTM provides dropdown for the origin with values “ASSIGNED”, “CRF” and “DERIVED”. YESDTM expecting the inputs for the columns ‘pages’, ‘method’ and ‘comment’ respect to the entry in the origin column. Users to enter the CRF page numbers only when the origin is “CRF”. User can enter multiple page number in the column separated by comma.

An **ADD** button will be present in the comment column when the origin is ASSIGNED or CRF. Similarly an **ADD** button will show up in the method column when the origin is DERIVED. User can add method/comment by clicking the add option within the respective column which will lead the user to a sub module where the user can add the derivation comment in the respective sections.



There are two fields in this module; comment and derivation. The comment value will be presented as the derivation comment for the specification/define XML and the derivation value will contain the SAS code for generating the variable.

User need to follow the YESDTM standards to add the SAS code. The dataset name should be the two digit abbreviation of the domain followed by the variable name. eg: if the user is deriving a variable named, RACE in the DM domain then the dataset name will be “DMRACE”. User need to keep the required variables for merging based on the domain. YESDTM will consider all the variables other than the derived variables as *by* variables. YESDTM will generate each variable separately and then combine those datasets to generate the domain.

Example for comment/Derivation:

Comment:

Set to ‘MULTIPLE’ when RAW.DM.XX=“Yes” Where XX refers to RACEAM, RACEAS and RACEBL;

Set to ‘AMERICAN INDIAN OR ALASKA NATIVE’ when RAW.DM.RACEAM=“Yes”;

Set to ‘ASIAN’ when RAW.DM.RACEAS=“Yes”;

Set to ‘BLACK OR AFRICAN AMERICAN’ when when RAW.DM.RACEBL=“Yes”;

Derivation:

```
Data dmrace;  
    Length race $200;
```

PhUSE EU Connect 2018

```
Set raw.dm;  
If strip(raceam)="yes" and strip(raceas)="yes"  
  and strip(racebl)="yes" then race="MULTIPLE";  
else if strip(raceam)="yes" then race="AMERICAN INDIAN OR ALASKA NATIVE";  
else if strip(raceas)="yes" then race="ASIAN";  
else if strip(racebl)="yes" then race="BLACK OR AFRICAN AMERICAN";  
keep subject race;  
run;
```

Note: Here YESDTM consider the subject as the by variable.

The SAS code will be saved by YESDTM, ID for these variables will then be created using domain abbreviations and the variable name separated by a dot (DM.RACE). Similarly user needs to complete the derivation for all the variables. At the time of execution these for all the variable will be save within a code. Merging is done based on the variable name and domain name with the by variables kept in the code. The SAS code will be then run with the help of command prompt and receive the log and dataset as an output for each domain.

VALIDATION OF THE SPECIFICATION

The developer can validate the specification with the help of pinnacle. Pinnacle API is a RESTful service, seamlessly integrated with your data warehouse, statistical computing environment and/or metadata repository to enable a true end-to-end process. After submitting the spec for validation, the user will get the pinnacle report of the current domain. User need to update the specification based on the warning / error displayed in the report. The cycle will be continued until pinnacle shows no errors in the report. Once the specification is finalized user can generate the datasets.

The specification reviewer will receive a notification when the user updates the actual completion date in the Task. The entire cycle will be repeated based on the comment of the reviewer.

DATA VISUALIZATION

Similar to the Pinnacle YESDTM also provide the data visualization with help of MOSS, which is a dynamic software framework, intended to eliminate the dependency on the data team, and allow clinicians to analyze trial data independently. MOSS is a web based application which will combine the features of CDISC, Natural Language Processing and Machine learning. It allows users to create and save dashboards and provide data drill down capabilities. MOSS supports a variety of Table and Graph presentation formats for the users to choose from. Statistical capabilities are also built-in which will help the clinicians perform adhoc analysis.

CONCLUSION

YESDTM can be connected to any raw data sources which enable access to the real-time data and connectivity to EDCs. Applications like Pinnacle -21 help YESDTM to address the validation issues during the development phase itself. It provides the administrators to track the study from the beginning to end. It also help with better resource management. The most important feature of YESDTM is that, it provides quality output in a lesser cycle time. The users can interact with YESDTM using voice and text interfaces. Users can also post the queries in the tracker which helps to keep a log of all the assumptions made for the study. The users of the tool are expected to have a basic understanding of the SDTM methodology and SAS. Sponsors can monitor the Safety and Efficacy, Data Quality and Risk Indicators of a clinical trial by effectively visualizing and analyzing data with help of MOSS.

LIST OF ABBREVIATIONS

1	AI	Artificial Intelligence
2	CRF	Case Report Forms
3	CT	Controlled Terminology
4	EDC	Electronic Data Capture
5	NLP	Natural Language Processing
6	QC	Quality Check
7	SDTM	Study Data Tabulation Model

PhUSE EU Connect 2018

REFERENCE

- Lawrence, R. (2008). Fundamentals of speech recognition. Pearson Education India.
- Wood, F. (2008). The CDISC Study Data Tabulation Model (SDTM): History, Perspective, and Basics. PharmaSUG Proceedings, June.
- <https://moss.clapresearch.com/>

PhUSE EU Connect 2018

CONTACT INFORMATION

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

Author Name : Akhil Vijayan
Company : Genpro Life science India Private Ltd
Address : Technopark, India
City / Postcode : Thiruvananthapuram, 695581
Work Phone : 781-373-8455
Email : akhil.vijayan@genproindia.com
Web: : www.genproindia.com

Author Name : Anish George
Company : Genpro Life science India Private Ltd
Address : Technopark, India
City / Postcode : Thiruvananthapuram, 695581
Work Phone : 781-373-8455
Email : anish.george@genproindia.com
Web: : www.genproindia.com

Author Name : Anoop Ambika
Company : Genpro Lifescience India Private Ltd
Address : Technopark, India
City / Postcode : Thiruvananthapuram, 695581
Work Phone : 781-373-8455
Email : anoop.ambika@genproindia.com
Web : www.genproindia.com: