

Taking Advantage of CDISC Standards and Planning the Study Specifications

Siddharth Kumar Lokineni, Syneos Health, Cary, USA

Vara Prasad Reddy Sakampally, Syneos Health, Cary, USA

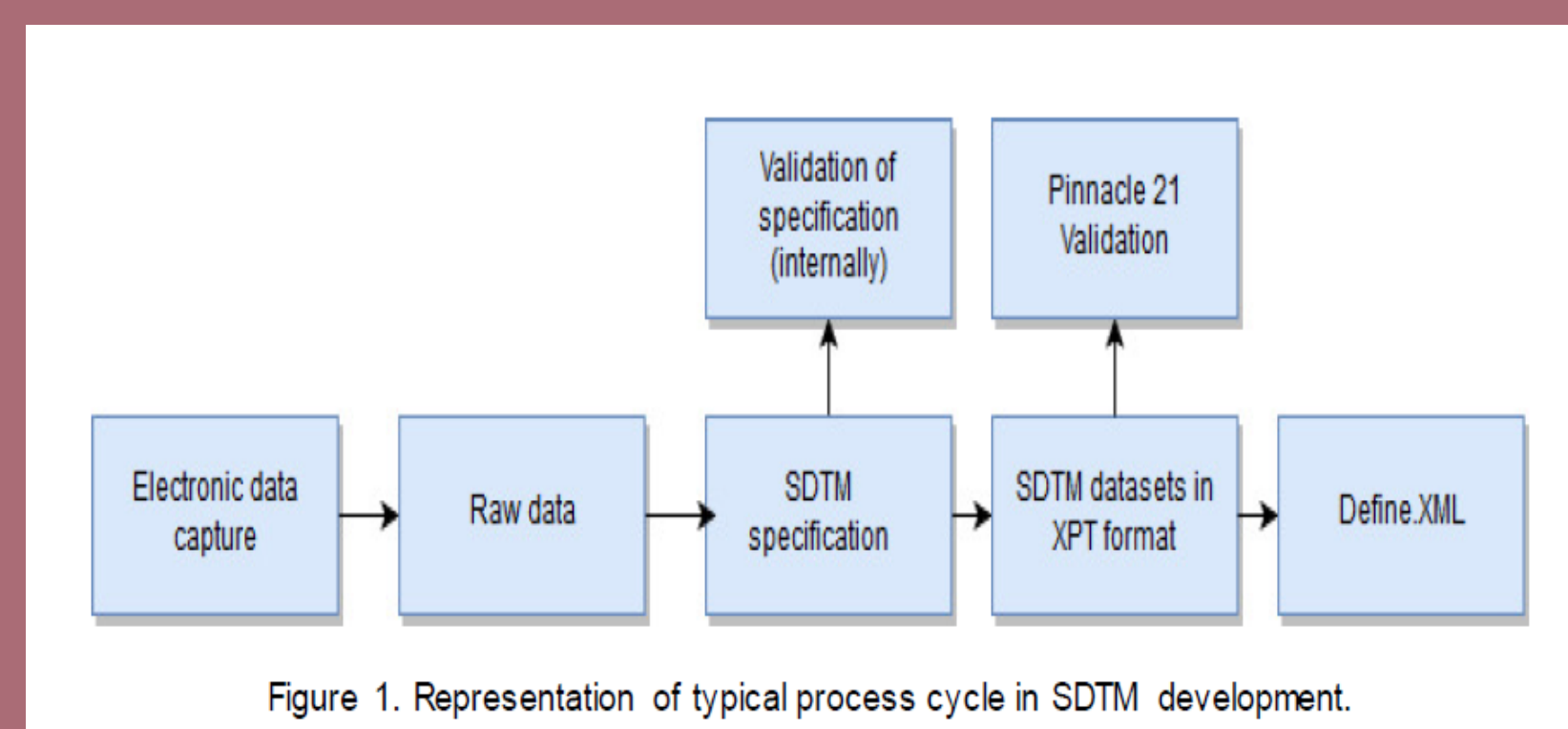
PhUSE US Connect 2019

Abstract

Implementing the CDISC® standards is a challenging task and will impact study time lines, budget, and quality if they are not considered before setting up a new study for programming. Addressing the key compliance issues and considering the necessary files needed for submission before stepping into the programming phase will help in a smooth flow of the study and avoid rework of many items. This paper will describe a method to design the specifications template to aid in annotating the CRF, checking mapping specifications against CDISC data standards, in generating the supplemental data definition document files and generating the define.xml that go along with the submission packet. Also, it will provide additional information on performing the gap analysis and additional checks on mapping all the data components captured in a study.

Introduction

The CDISC data standards facilitate uniformity across the industry to represent data collected during a clinical study. With the fast paced deliverables and challenging timelines to deliver quality analysis to clients, it is very important to get dataset programming correct the first time. It is vital to devise an efficient way of designing the specifications to get accuracy and consistency in programming and corresponding output data for CDISC SDTM mapping and ADaM data derivation. This in turn will help in speeding up programming cycle, avoid budgeting issues and rework. A specification is traditionally a multi-tabbed EXCEL file, displaying one domain per sheet, one row for each variable, and spreadsheet columns mimicking the CDISC implementation guide. But here the idea is to generate a specification file in a single tab instead of multiple tabs. This way we can reuse it for CRF annotation, Define.XML easily. A well-defined and organized specification document minimizes the time needed to acclimatize with the metadata required when working on Define documentation during the later stages, which significantly reduces the overall time for review. Note that the specs are reviewed before the dataset programming is started. This ensures quality when metadata information is reused in a submission for regulatory review.



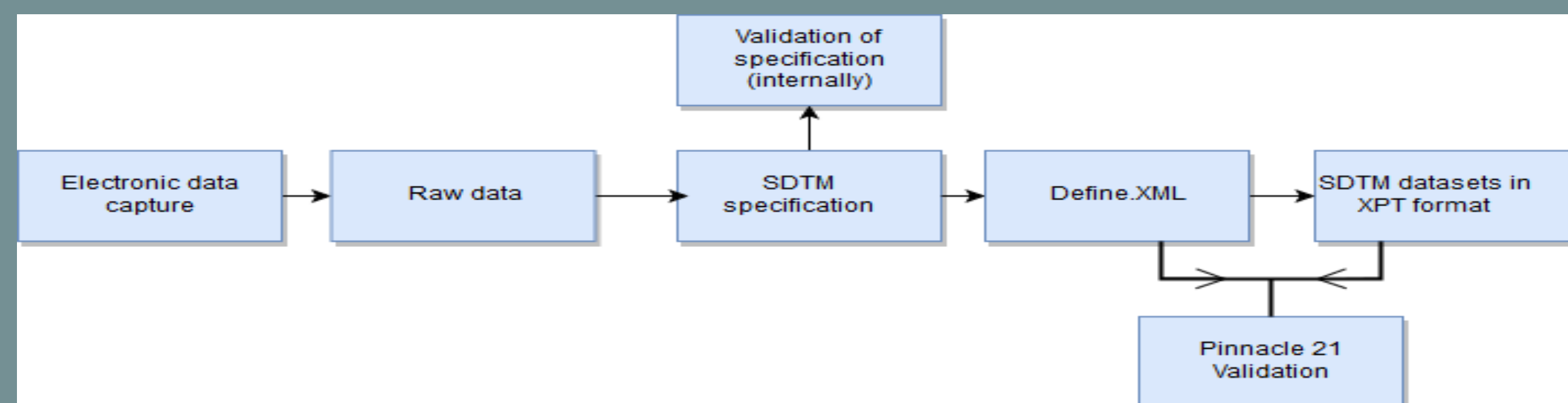
CDISC guidelines will be used to create these specifications. Although the paper begins with a brief description of the specifications structure, some prior knowledge of specification document and excel will be helpful.

SINGLE-TAB Specification uses

❖ **Variable Metadata update** :Metadata has many defining properties about the variable. The information about the variable label, type, format/CT, origin, role, length, and format are these properties. Usually there are certain variables that are repeated across domains. When there is a need to update the variable attributes across domains, a single tab spec will be really useful. The user will be filtering the EXCEL spreadsheet by the variable to update the attributes. In the below figure, the variable EPOCH is present in multiple domains. If an update is warranted to any of the attributes, having a single tab makes it much easier. In the below example, we can filter the third column to the variable EPOCH and make the change across domains.

Domain Prefix	Variable Name (minus domain prefix)	Variable Name	Variable Label	Type	Controlled Terms or Format	Origin	Length	Format
AE	EPOCH	EPOCH	EPOCH	Char	Protocol	Protocol	20	\$20.
CM	EPOCH	EPOCH	EPOCH	Char	Protocol	Protocol	20	\$20.
DA	EPOCH	EPOCH	EPOCH	Char	Protocol	Protocol	20	\$20.

❖ **Preparing Define.XML during specification and dataset**: It has often been observed that when a metadata discrepancy has been noticed in the define.xml/define.pdf, the entire cycle of updating and rerunning the datasets process has to be performed. To avoid this, we can use the single tabbed spec file which will be very similar to the specification file that Pinnacle21® uses as input to generate the define-xml/pdf. In other words the proposal here is that the define documentation should be done during the SDTM/ADaM development, particularly while dataset mapping specification are created as shown in the figure below. Generating define initially will subject it to multiple review cycles and also speeds up the entire process.



• Typically in the industry, define.xml is generated towards the final submission stage and assumes that it has to be done after the SDTM/ADaM development lifecycle. The figure above demonstrates SDTM development and implementation life cycle. Here, the single tabbed specification is made use of and the define.xml is created right after the specifications are worked on unlike the traditional way. Once the SDTM datasets are developed and validated, the datasets will also be passed through the Pinnacle21 to check for the CDISC compliance.

❖ **Ease of review for internal reviewers and clients**: Having a single tabbed specification file will make the job of validators and reviewers simpler. It is much convenient to check the logical consistency among similar variables. This also eases the communication with the clients. If clients find any inconsistency – it is easy to point out all the variables and cross check the logic implemented in other datasets. Figure illustrates the details mentioned.

Observation Class	Domain Prefix	Variable Name (minus domain prefix)	Variable Name	Variable Label	Type	Origin	Role	CDISC Notes (for domains) Description (for General Classes)	Core	Mapping Specifications
Events	AE	STDY	AESTDY	Study Day of Start of Adverse Event	Num	Derived	Timing	Study day of start of adverse event relative to the sponsor-defined RFSTDTC	Perm	AE start date - reference start date (+ 1 day if AE start date == reference start date)
Interventions	CM	STDY	CMSTDY	Study Day of Start of Medication	Num	Derived	Timing	Study day of start of medication relative to the sponsor-defined RFSTDTC	Perm	CMSTDTCT - RFSTDTC (+ 1 day if CMSTDTCT == RFSTDTC)
Events	DS	STDY	DSSTDY	Study Day of Start of Disposition Event	Num	Derived	Timing	Study day of start of event relative to the sponsor-defined RFSTDTC	Perm	DSSTDTCT - DM RFSTDTC (+ 1 day if DSSTDTCT == DM RFSTDTC)
Events	MH	STDY	MHSTDY	Study Day of Start of Observation	Num	Derived	Timing	Actual study day of start of observation expressed in integer days relative to the sponsor-defined RFSTDTC	Perm	Medical history start date - reference start date (+ 1 day if medical history start date == reference start date)
Special-Purpose	SV	STDY	SVSTDY	Study Day of Start of Visit	Num	Derived	Timing	Study day of start of visit relative to the sponsor-defined RFSTDTC	Perm	SVSTDTCT - RFSTDTC (+ 1 day if SVSTDTCT == RFSTDTC)

• Having a single tabbed specification file will make the job of validators and reviewers simpler. It is much convenient to check the logical consistency among similar variables. This also eases the communication with the clients. If clients find any inconsistency – it is easy to point out all the variables and cross check the logic implemented in other datasets.

Design of specification

The spec will contain the following sheets, each serving a purpose.

❖ **Change log sheet**:

Below screenshot of the change log sheet has the following columns: Version, Date, Changes, Changed by, Reviewed by and Review Date. “Version” column documents the version of the spread sheet. “Date”, “Changes” and “Changed by” columns signify the date on which the changes were made, what were the changes and who made the changes respectively. The last two columns “Reviewed by” and “Review date” indicate who was the reviewer and when was the review done.

❖ **Change log sheet**:

This sheet tells the datasets that are needed or the study. It has the following details presented– Dataset, Dataset label, Dataset class per CDISC guidelines, Dataset structure per CDISC guidelines and Key variables used in final sort. CRF pages will give information of the pertinent CRF page for the related information. Dependencies column will help to sort the dependent and independent datasets which can be worked on in appropriate hierarchy.

❖ **Variable Sheet**: All the variables in every dataset are present in this sheet and it contains the below columns–

- Seq. for Order – Gives the information about the order of the variables in the dataset.
- Observation Class – Per the CDISC SDTM IG the observations class is presented here.
- Domain Prefix - Presents the two character abbreviations for the domain.
- Variable Name(without domain prefix and with domain prefix) – Variable name without the domain prefix will help to filter and sort the similar type of the variables. Variable name with the domain prefix will be domain specific.
- Variable Label – Label of the domain.
- Type – identifies the type of the variable (Numeric or Character).
- Controlled Terms or Format – This will have allowed CT’s or the formats for the Age units (Years, Months or days), Domains, Ethnicity, Race, Sex, Country, Date format (ISO 1806) or Y/N format.
- Origin – Source of the variable is provided (Protocol, Assigned, Derived or CRF).
- Role - Identifies the type of the variable. Per CDISC there are different type of variables (Identifier, Topic, Record Qualifier, Timing etc..).
- CDISC Notes (for domains) and Description (for General Classes) – Provides detail information per the CDISC guide lines for the domains and other important information. For example in the below figure, RFSTDTC has information per the SDTM IG and useful information for the programmer.
- Core – Tells if the variable is required, expected or permissible. This categorization is based on the CDISC SDTM IG.
- Mapping Specification – Gives the data sources or derivation algorithm for variable.
- Length and Format – Length and Format of the variable can be found here.
- Submission Comments- Provision for submission comments that can be used in the define.xml

Below figure illustrates the variable sheet

Seq. For Order	Observation Class	Domain Prefix	Variable Name (minus domain prefix)	Variable Name	Variable Label	Type	Controlled Terms or Format	Origin	Role	CDISC Notes (for domains) Description (for General Classes)	Core	Internal Specifications	Length	Format	Submission Comments
1	All Classes	STUDYID	STUDYID	Study Identifier	Char				Identifier	Unique identifier for a study					
2	All Classes	DOMAIN	DOMAIN	Domain Abbreviation	Char				Identifier	Two-character abbreviation for the domain must be relevant to the observation. It is also used as a prefix for identifier used to uniquely identify a subject across all studies for all applications or submissions involving an identifier used to identify a result from a group of subjects that is					
3	All Classes	USUBID	USUBID	Unique Subject Identifier	Char				Identifier						
4	All Classes	POOLID	POOLID	Pool Identifier	Char				Identifier						

❖ **Value level Spreadsheet**: Value level metadata information is presented in this sheet. It is designed to resemble define specifications sheet closely to allow it to copy-paste from this document. This will usually be a live document and will be updated during the course of the study.

❖ **Codelist**: This sheet lists all possible values as they appear for the data. The data sheet will be populated as new data gets populated. This will help in the final documentation of the SDTM.

❖ **SDTM Reference sheet**: It is always useful when a SDTM reference document is available handy. It has details and explanation about a variable. This information is documented from the SDTM IG and will come handy to the programmers and CRF annotators. Below figure gives an example of SDTM reference sheet.

Seq. For Order	Observation Class	Domain Prefix	Variable Name (minus domain prefix)	Variable Name	Variable Label	Type	Controlled Terms or Format	Origin	Role	CDISC Notes (for domains) Description (for General Classes)	Core	Internal Specifications	Length	Format	Submission Comments
1	All Classes	STUDYID	STUDYID	Study Identifier	Char				Identifier	Unique identifier for a study					
2	All Classes	DOMAIN	DOMAIN	Domain Abbreviation	Char				Identifier	Two-character abbreviation for the domain must be relevant to the observation. It is also used as a prefix for identifier used to uniquely identify a subject across all studies for all applications or submissions involving an identifier used to identify a result from a group of subjects that is					
3	All Classes	USUBID	USUBID	Unique Subject Identifier	Char				Identifier						
4	All Classes	POOLID	POOLID	Pool Identifier	Char				Identifier						

Conclusion

Use of single tabbed specifications will ease the update and review process, which in turn will save time. By standardizing the specification file, proprietary tools can be developed at an organizational level to assist in define.xml creation, CRF annotation, define.pdf creation that saves a lot of time and money for the organizations, thus keeping the projects under budget. When this approach was implemented compared to traditional specifications, it was observed that programming and study are a lot more organized. This user-friendly interface makes it straightforward to apply changes as the study progresses for similar variables across domains.

References

[1] SDTMIG v3.3
<https://www.cdisc.org/standards/foundation/sdtm>.

[2] Yurong Dai, Jiangang Jameson Cai
“Conversion of CDISC specifications to CDISC data–specifications driven SAS programming for CDISC data mapping” Proceedings of PharmaSUG 2017 Conference.

[3] Vara Prasad Sakampally, Bhavin Busa
“Consider Define.xml Generation during Development of CDISC Dataset Mapping Specifications”

Acknowledgements

We would like to thank our manager Kelly Swartz, Programming director Nancy Fish and our colleagues Vishnu Alamuri, Prasad Marasa for reviewing and providing valuable inputs when ever asked for. Thanks to Syneos Health for supporting us to be at the PhUSE 2019.