PhUSE US Connect 2018

Paper SI12

How to Prepare High-quality Metadata for Submission

Varun Debbeti, Cytel Inc., Waltham, MA, USA

ABSTRACT

One of key factors in getting approval of any submission is preparing a high quality metadata that is devoid of issues. The issues may be due to inadequate preparation of metadata, and/or, the lack of knowledge of the latest requirements of agency, and/or traceability issues, and/or inadequate validation of metadata. This paper presents the utilization of metadata submission guidelines, FDA validator rules, FDA rejection criteria, FDA Business rules, FDA Technical Conformance Guide, eCTD validation criteria, and define.xml specifications guide in combination with tools like Pinnacle 21 ® community, Pinnacle 21 Enterprise and Pinnacle 21 validator tools to overcome some common issues and help in creating high quality metadata. This paper also discusses tips and tricks to avoid some specific issues from Pinnacle 21 validator report run on define.xml and xpt data.

INTRODUCTION

This paper primarily focuses on New Drug Applications (NDAs) and Biologics License Application (BLAs) of clinical data only to CDER and CBER evaluation centers in FDA.

WHAT IS METADATA?

Definition: "*Metadata is descriptive information about an object or resource, whether it is physical or electronic*". The simplest definition of metadata is "structured data about data". The description could be about its content, format, intended use, origin, derivation, etc. E.g., the define.xml is the metadata describing the structure and content of the submitted datasets. In Figure 2 below, a dataset has actual data about subject information, whereas metadata in a define.xml has information like dataset name, dataset description, class as per standard implementation guide, structure of dataset, its purpose, key variables to identify unique records, and its location etc. Metadata Constituents: It includes aCRF, Define.xml, Define.pdf, Reviewer's guides, Stylesheets, validation reports (P21 reports), and other reference documents etc. Considering the broader scope of metadata constituents, and available limit of this paper's length, discussion is focused on Define.xml issues, Reviewer's guides and clean validation reports.

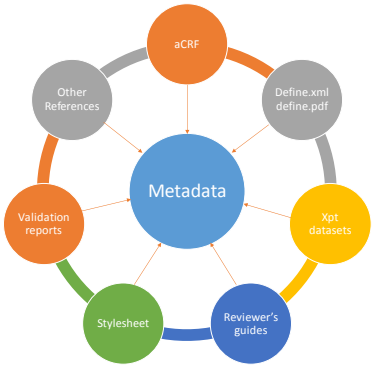
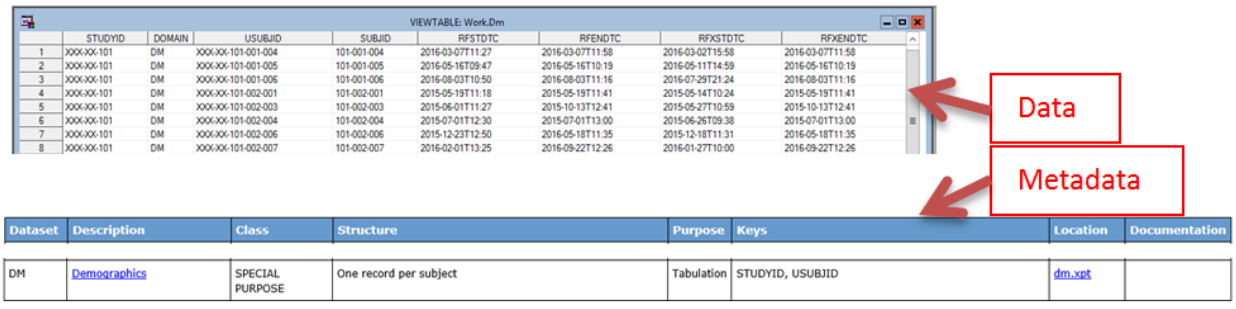


Figure 1 Constituents of Metadata in a typical Clinical data submission

Importance of metadata to understand the data: During the data interpretation process in a drug development environment, sometimes, we get overwhelming job of effectively managing different kinds of data. Figure 2 shows the flow of data during a clinical trial originated at investigative site using Electronic-Data-Capture system. Example: Patient data collected on CRFs from the investigative sites – for example, AE (Adverse Events), EG (Electrocardiogram), LB (Laboratory Results), and VS (Vital Signs) data. After raw data is collected into the clinical trial database, the next step is data standardization and manipulation. Derived/Analysis SAS ® datasets are generated based on specific rules and criteria that are defined by metadata. The repository of this metadata is called as Data Definition Table (DDT) or define.xml. The define.xml describes the metadata of the submitted electronic datasets, and is considered arguably the most important part of the electronic submissions for regulatory review. An insufficiently documented define.xml is a common deficiency that reviewers have noted.



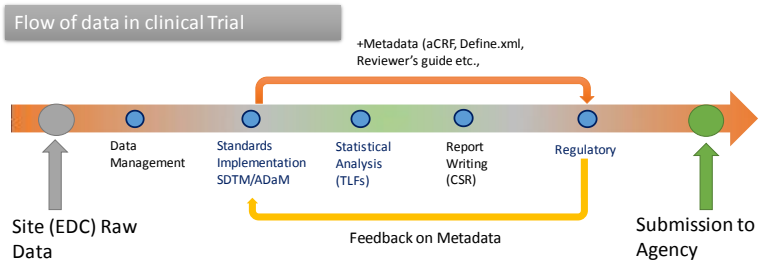


Figure 2 Data vs Metadata (on top); Flow of data in clinical trial (on bottom);

HOW TO PREPARE HIGH QUALITY METADATA

Reasons for low quality Metadata are the metadata not following the current requirements of review division, inadequate preparation and validation of metadata, traceability issues etc. This paper focuses on explaining how to overcome these issues by following six steps below with the help of standard guidance documents from FDA and SDOs like CDISC together with automatic tools like Pinnacle-21.

1. Knowledge on standards and latest requirements
2. Efficient preparation and Extensive Validation
3. Interpretation of issues
4. Evaluation of Impact of Issues
5. Fixing Critical Issues
6. Explaining non-fixable issues:

KNOWLEDGE ON STANDARDS AND LATEST REQUIREMENTS

FDA Data Standards Catalog [1]: The Catalog provides a listing of currently supported and/or required standards, their uses, the date FDA will begin (or has begun) to support a particular standard, and the date support ends (or will end), the date the requirement to use a particular standard will begin (or has begun), the date such requirement ends (or will end), and other pertinent information. The Agency may Refuse-To-File (RTF) for NDAs and BLAs in an electronic submission that does not have study data in conformance to the required standards specified in the Catalog.

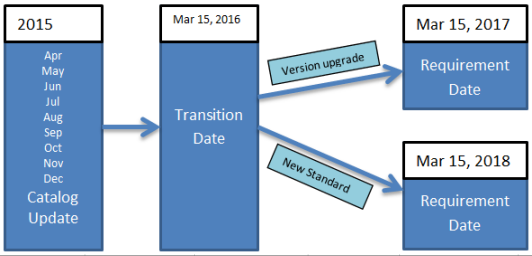
FDA Data Standards Catalog v4.10 (10-24-2017) - Supported and Required Standards

This table contains a listing of the data exchange, file formats and terminology standards supported at FDA. These standards have gone through all the steps necessary to make this part of the regulatory review process, including posting of regulatory guidance documents and associated implementation guidelines and technical specifications. The submission of standardized data using any standard not listed, or to an FDA Center not listed, should be discussed with the Agency in advance. This catalog is incorporated by reference in the guidance to industry, *Providing Regulatory Submissions in Electronic format-Standardized Study Data (http://www.fda.gov/downloads/Drugs/Guidances/UCM292334.pdf).*

Use	Data Exchange Standard	Exchange Format	Standards Development Organization (SDO)	Supported Version	Implementation Guide Version	FDA Center(s)	Date Support Begins (MM/DD/YYYY)	Date Support Ends (MM/DD/YYYY)	Date Requirement Begins (MM/DD/YYYY)	Date Requirement Ends	Regulatory Reference and Information Sources
Regulatory Applications (IND, NDA, ANDA, BLA, master files)	Electronic Common Technical Document (eCTD)	Extensible Markup Language (XML)	International Council for Harmonisation (ICH)	3.2.2	M2 eCTD: Electronic Common Technical Document Specifications	CDER, CBER	06/01/2008		05/05/2017 [5] 05/05/2018 [6]		Electronic Submissions-Electronic Common Technical Document (eCTD)
Clinical and Non-Clinical study data sets - Transport	SAS Transport (XPORT)	XPT	SAS	5	SAS Technical Support TS-140	CDER, CBER	Ongoing		12/17/2016 [1] 12/17/2017 [2]		For CDER and CBER only: Technical Conformance Guide
Clinical study datasets	Study Data Tabulation Model (SDTM)	XPT	Clinical Data Interchange Standards Consortium (CDISC)	1.1	3.1.1	CDER, CBER	Ongoing	01/28/2015		01/28/2015	CDISC.org - SDTM; See Technical Conformance Guide
Clinical study datasets	SDTM	XPT	CDISC	1.2	Version 3.1.2 Amendment 1	CDER, CBER	08/07/2013	03/15/2019 [1] 03/15/2020 [2]	12/17/2016 [1] 12/17/2017 [2]	03/15/2019 [1] 03/15/2020 [2]	CDISC.org - SDTM
Clinical study datasets	SDTM	XPT	CDISC	1.2	3.1.2	CDER, CBER	10/30/2009	03/15/2019 [1] 03/15/2020 [2]	12/17/2016 [1] 12/17/2017 [2]	03/15/2019 [1] 03/15/2020 [2]	CDISC.org - SDTM
Clinical study datasets	SDTM	XPT	CDISC	1.3	3.1.3	CDER, CBER	12/01/2012		12/17/2016 [1] 12/17/2017 [2]		CDISC.org - SDTM
Clinical study datasets	(SDTM)	XPT	CDISC	1.4	3.2	CDER, CBER	08/17/2015		03/15/2018 [1] 03/15/2019 [2]		CDISC.org - SDTM
Clinical study datasets	Analysis Data Model (ADaM)	XPT	CDISC	2.1	1.0	CDER, CBER	Ongoing	03/15/2019 [1] 03/15/2020 [2]	12/17/2016 [1] 12/17/2017 [2]	03/15/2019 [1] 03/15/2020 [2]	CDISC.org - ADaM
Clinical study datasets	Analysis Data Model (ADaM)	XPT	CDISC	2.1	1.1	CDER, CBER	03/15/2018		03/15/2019 [1] 03/15/2020 [2]		CDISC.org - ADaM
Study data definition	Define	XML	CDISC	1.0	N/A	CDER, CBER, CDRH	Ongoing	03/15/2018 [1] 03/15/2019 [2]	12/17/2016 [1] 12/17/2017 [2]	03/15/2018 [1] 03/15/2019 [2]	CDISC.org - Define-XML
Study data definition	Define	XML	CDISC	2.0	N/A	CDER, CBER, CDRH	08/07/2013		12/17/2016 [1] 12/17/2017 [2]		CDISC.org - Define-XML

FDA released the initial data standards catalog in December 2014, and had given the industry 2 years of time to meet these requirements. Any study starting after 17 Dec 2016 for NDA, BLA and ANDAs should follow the set standards in order to avoid refuse to file. FDA maintains this document using a transition-date approach, by making updates regularly to meet with the version updates to the FDA-supported standards, and any new standards that were not listed previously by publishing in Federal Register (https://www.federalregister.gov/documents/current#food-and-drug-administration) via a notice of transition date (Fixed Month/date: Next Calendar Mar 15). See example below where FDA allows 2 years period of time after a set transition date to support new standard before making it a requirement to use new standard. The time between transition date

and requirement start date is 1 year for any version upgrade to existing standards. However, sponsors are encouraged to use the update/new standard as soon as it is identified for use by FDA and the standards catalog is updated [3].



Type of requirement	Example document	Date Released by SDO (yyyy-mm-dd)	Federal Notice of FDA Support (yyyy-mm-dd)	Publication Date (yyyy-mm-dd)	Updated Data Standards Catalog (yyyy-mm-dd)	Transition Date	Date Requirement Begins (yyyy-mm-dd)
Initial Requirement	Providing Regulatory Submissions in Electronic Format — Standardized Study Data	-NA-	-NA-	2014-12-01	2014-12-01	-NA-	2016-12-01
Version Updates	SDTM 4.1	2016-02-15	2016-05-06	-NA-	2016-05-06	2017-03-15	2018-03-15
New Standards	DS Exchange 2.0	2017-02-20	2018-04-22		2018-04-22	2019-03-15	2021-03-15

Figure 3 Example showing data standard catalog update and transition date process

Other Standard documents and regulations: There are many other standard documents released by FDA every year for easy understanding of requirements. Reference to standard available documents https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm .

METADATA PREPARATION/VALIDATION

The submission package should consist of aCRF, XPT files of SDTM and ADaM domains, define.xml, define.pdf (for printing purpose), reviewer's guidance, Complex algorithms and any additional external files. There are many possible ways of validation of aCRF, Define.xml, and datasets. It is important to set right validators to get high quality metadata. Examples are provided below for some standard guidance documents used for validation.

Define.xml schema validation [4]: Validating a define.xml instance means executing a process to ensure that the XML document follows the structure and content rules specified in the define schemas. Processing a define.xml file is difficult or impossible for the receiving party when the file does not conform to the expected structure. For this reason, define.xml documents submitted to a regulatory agency or exchanged with an external partner are typically validated against the define.xml schema prior to processing their contents. CDISC provides define.xml validation schema files that can be downloaded and used to setup the external xml validators available. (The schema files are retrieved from http://www.cdisc.org/define-xml.). Websites for some XML tools evaluated are provided in the document. E.g., DefineValidator https://www.phaseforward.com/products/cdisc/, Oxygen XML Editor http://www.oxygenxml.com , Eclipse http://www.eclipse.org/, Cladonia XMLExchanger http://www.exchangerxml.com, Liquid XML http://www.liquid-technologies.com/, Validome http://www.validome.org

4. Data Conformance Summary

4.1 Conformance Inputs

Was OpenCDISC used to evaluate conformance? Yes/No

If yes, specify the versions of OpenCDISC and the OpenCDISC validation rules:

[Text Here]

Were sponsor-defined validation rules used to evaluate conformance? Yes/No

If yes, describe any significant sponsor-defined validation rules:

[Text Here]

Were the SDTM datasets evaluated in relation to define.xml? Yes/No

Was define.xml evaluated? Yes/No

Provide any additional compliance evaluation information:

Figure 4 A screenshot from Study-data-reviewers-guide

Define.xml Pinnacle 21 validation: As shown in figure above, as per Study-data-reviewers-guide template from PhUSE (released as SDRG Package v1.2) [5], it is desirable to validate define separately, and datasets in relation to define.xml. Pinnacle 21 community (previously OpenCDISC) toolset can do three types of validation. They are data only, data + define, and define only validation. All examples shown in this paper uses P21 report generated for data + define. Enterprise version of Pinnacle-21 is also available with subscription, and is a very helpful tool in creation of high quality define.xml. It has tools that automatically populates codelists and value-level metadata using study data submitted to the tool which saves a lot of time when compared to manual method.

FDA Business run/Validator rules [6]: FDA released latest version of business rules in Dec 2017 for sponsors to evaluate their study data before submission. FDA also released version 1.2 of validator rules in Dec 2017. The validator rules are technical

version of business rules. See example below, the business rule (FDAB035) applies technically to 6 different situations that are drafted as validator rules.

FDA Business Rule ID	FDA Business Rule	FDA Validator Rule	Domains
FDAB035	The definition of datasets, variables, and codelists in define.xml must reflect the actual study data.	Datasets included in study data must be described in the data definition document (define.xml).	ALL
		Domains referenced in data definition document (define.xml) should be included in the submission.	ALL
		Variables listed in the data definition document (define.xml) should be included in the dataset.	ALL
		Variables included in the dataset must be described in the data definition document (define.xml).	ALL
		Variable Data Types in the dataset must match the variable data types described in the data definition document (define.xml).	ALL
		Variable values should be populated with terms found in the user-defined codelist associated with the variable in define.xml.	ALL

Pinnacle 21 ID	Publisher ID	Message	Description	Category	Severity
SD1063	FDAC023	Dataset is not present in define.xml	Datasets included in study data must be described in the data definition document (define.xml).	Metadata	Error
SD0061	FDAC024	Domain referenced in define.xml but dataset is missing	Domains referenced in data definition document (define.xml) should be included in the submission.	Metadata	Warning
SD0054	FDAC025	Variable in define.xml is not present in the dataset	Variables listed in the data definition document (define.xml) should be included in the dataset.	Metadata	Warning
SD0060	FDAC026	Variable in dataset is not present in define.xml	Variables included in the dataset must be described in the data definition document (define.xml).	Metadata	Error
SD0059	FDAC034	Define.xml/dataset variable type mismatch	Variable Data Types in the dataset must match the variable data types described in the data definition document (define.xml).	Metadata	Error
SD0037	FDAC037	Value for variable not found in user-defined codelist	Variable values should be populated with terms found in the user-defined codelist associated with the variable in define.xml.	Terminology	Error

All the above validator rules are covered by P21 rules in Pinnacle 21 community toolset. However, some mismatches between define and datasets like below are not covered by P21 community rules and should be manually checked.

1. Attributes like length not matching between data and define
2. Sequence of model permissible variables in data which are not matching in define.xml are not found by P21 community e.g., When QSMETHOD is added to QS domain, and the order of QSMETHOD in define is different in define and dataset, it is not identified by P21 community tool.

SDTM-Metadata Submission Guidelines [7]: The purpose of the Study Data Tabulation Model - Metadata Submission Guidelines (SDTM-MSG) is to provide guidance for preparing the eCTD module 5 “sdtm” folder efficiently. Following are some cases where the guidance could be useful to solve issues.

Case 1: SDTM-MSG says “In the event that no records are present in a dataset (e.g., a small study where no subjects took concomitant medications), the empty dataset should not be submitted or described in the define.xml.” It is a simple rule to follow by checking number of records in datasets and removing empty datasets. Sponsor should verify if 1. All non-zero sas data has a corresponding XPT dataset generated 2. If any zero-record sas data has an XPT dataset generated, and should be removed so it is not submitted in define generation.

P21 community also has a rule to check this issue.

SD0001	FDAC014	No records in data source	Domain table should have at least one record.	Presence	Error

The case below is an extrapolated situation of this issue.

Comments collected on Inclusion/Exclusion Criteria CRF page are mapped to CO domain. So, CO and IE domains are related by USUBJID, RDOMAIN, IDVAR and IDVARVAL. At the first snapshot of raw data, there were comments entered on the CRF page, and subsequently mapped to CO domain. In the next snapshot of data, the comments are removed from the page. So, the CO xpt/dataset are expected to have zero records, and by the regular check, one should be able to find and remove this CO xpt/dataset with zero records. However, if a programmer forgot to remove old xpt files before re-running for the second snapshot of raw data, and the CO program is set in a way that it will not create zero record xpt/datasets, it leads to the old xpt not being replaced with zero record dataset. The tool that checks and removes zero record datasets will overlook the CO, and causes invalid reference of the new IE domain to old CO Xpt file/dataset. So, it is always recommended to remove old xpt files in the folder before re-running for a different snapshot of raw data even if the programs won’t create empty xpt/datasets.

CO					
	SD0077	FDAC074	Invalid referenced record	Error	1

Case 2: Split domains

Domain Name vs Dataset name: A dataset is just a collection of data. Most commonly a data set corresponds to the contents of a single database table, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set. Domain is defined as per the SDTM model. As per SDTM IG 3.2 [8] “A domain may be comprised of more than one physical dataset, for example the main domain dataset and its associated Supplemental Qualifiers dataset.” For e.g., QS domain may include QS and SUPPQS datasets. There is a common misconception that the split domain datasets (e.g., QSCG, QSCS, or QSMM) should have domain variable values with 4 letters same as dataset name, and/or the define.xml controlled terminology-code list for Domains to include 4 letter split dataset names. SDTM-MSG

says “The domain variable value for all split domains is QS; however, the dataset names are unique and prefixed with QS. The annotated CRF refers to the domain name (QS) as opposed to the dataset name (QSCG, QSCS, or QSMM)”. It should be noted that the domain name of split domains does not change due to the fact that they are split in to multiple datasets and it is advised that the sponsor check with their review agency regarding exactly what needs to be included in the submission, i.e. the split datasets or both the split datasets and the un-split datasets.

Define-XML-2-0-Specification [9]: This specification describes latest Define_XML model i.e., 2.0.0 which is used to describe SDTM and ADaM datasets for FDA submissions. It supports SDTM v3.1.2 or higher, and ADaM v1.0 or higher. A define.xml document will have 7 sections; Dataset-level metadata, Variable-level metadata, Value-level metadata, Controlled terminology, External Dictionaries, Computation algorithms, and Comments. The specification document helps in developing and validation of a standard define.xml file.

Dataset-level metadata: Image below shows how a data-set level metadata appears in browser mode

Tabulation Datasets for Study [redacted] (SDTM-IG 3.2)

Dataset	Description	Class	Structure	Purpose	Keys	Location	Documentation
TA	Trial Arms	TRIAL DESIGN	One record per planned element per arm	Tabulation	STUDYID, ARMCD, TAETORD	ta.xpt	
TE	Trial Elements	TRIAL DESIGN	One record per planned element	Tabulation	STUDYID, ETCD	te.xpt	
TI	Trial Inclusion/Exclusion Criteria	TRIAL DESIGN	One record per I/E criterion	Tabulation	STUDYID, IETESTCD, TIVERS	ti.xpt	
TS	Trial Summary	TRIAL DESIGN	One record per trial summary parameter value	Tabulation	STUDYID, TSPARMCD, TSSEQ	ts.xpt	
TV	Trial Visits	TRIAL DESIGN	One record per planned visit per arm	Tabulation	STUDYID, VISITNUM, ARMCD	tv.xpt	

The define.xml that is included in a submission should always describe each dataset that is included in the submission and provide the natural key structure of each dataset. A natural key is a piece of data that uniquely identify that entity, and distinguish it from any other observation in the dataset. The advantage of natural keys is that they exist already, and one does not need to introduce a new “unnatural” value to the data schema. One of difficulties with natural keys is that it always has potential to change. Because they have business meaning, natural keys are often coupled to the business, and they may need to be reworked when business requirements change. An example of such a change is shown below where addition of a position or location that becomes a key in a new study, but were not in the standard available natural keys.

Physical Examination (PE) domain example: Sponsor A chooses the following natural key for the PE domain: Sponsor B collects data in such a way that the location (PELOC) and method (PEMETHOD) variables need to be included in the natural key to identify a unique row, but they do not collect a visit variable; instead they use the visit date (PEDTC) to sequence the data. Sponsor B then defines the following natural key for the PE domain.

Sponsor-A natural keys: STUDYID, USUBJID, VISTNUM, PETESTCD

Sponsor-B natural keys: STUDYID, USUBJID, PEDTC, PETESTCD, PELOC, PEMETHOD

If the key variables are not able to identify unique records, there could be four types of duplicate records based on the reason.

1. True double entry in the data per USUBJID, VISTNUM, PETESTCD and only PEORRES is different among the duplicate records identified. These should be queried to CDM and only record with appropriate result should be kept in the data, and resolved before locking study
2. Duplicates identified due to insufficient standard Key variables – The above example of PE falls under this category, and should be added in define.xml for easy identification by reviewers.
3. Duplicates identified due to incorrect standard Key variables. Example: Sponsor mapped FA domain with Key variables USUBJID FACAT FASCAT FATESTCD VISITNUM whereas the standard keys described in specification are USUBJID FATESTCD FAOBJ – these should be fixed before submission
4. Duplicates identified due to missing values in key variables. Example: Standard keys are STUDYID USUBJID DATESTCD and DADTC in DA domain. If the data has missing values of DADTC, multiple records will have same DADTC (i.e., missing) as per Pinnacle tool and are identified as duplicates.

Variable-level metadata: Invalid data type errors: CDISC SDTM and ADaM data types are specified as character (“Char”) or numeric (“Num”), whereas define.xml supports wide range of data types.

Demographics (DM) [Location: dm.xpt]

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin	Derivation/Comment
STUDYID	Study Identifier	1	text	11		Assigned	[redacted]
DOMAIN	Domain Abbreviation		text	2	DOMAIN	Assigned	Assigned as "DM"
USUBJID	Unique Subject Identifier	2	text	19		Derived	Concatenate [redacted]
SUBJID	Subject Identifier for the Study		text	11		Assigned	Raw dataset DEMOG.SUBJECT

Following table illustrates mapping of some commonly used data types in define.xml to the SDTM/ADaM datatypes:

Define-XML Data Type	Submission Data Type	Length	Considerations
text	Char	Maximum allowable length.	SAS Version 5 transport files restrict variable lengths to 200 characters.
integer	Num	The largest allowable integer width.	Use for numeric or equivalent variables that have discrete whole values (non-fractional). Can be positive, negative, or zero. ADaM date variables, are provided as integers.
float	Num	The largest allowable whole number width plus the maximum number of decimal digits.	Use for numeric variables that may contain a fractional component. It represents the set of all the decimal numbers with arbitrary lengths.
datetime	Char	N/A	Use if values for SDTM or Send variable represent Date Times (YYYY-MM-DDTHH:MM:SS.SS).

If the data types like “Char” and “Num” are used in define.xml, it leads to errors during validation of define.xml. Error from a P21 validation report is shown below.

OD0075	Invalid Data Type value for variable			Error	112
OD0075	Invalid Data Type value for variable	The Data Type attribute for Variable must have a value of 'text', 'integer', 'float', 'datetime', 'date', or 'time' for Define-XML v1.0, and 'text', 'integer', 'float', 'datetime', 'date', 'time', 'partialDate', 'partialTime', 'partialDatetime', 'incompleteDatetime', or 'durationDatetime' for Define-XML v2.0.	Terminology	Error	

Miscellaneous uses: With availability of modern tools that creates and validates define.xml, there is very less need for people to learn the XML coding, but when certain errors show up in the report, one should have reference for the terms used in the XML code, the define specification document is most helpful during those times. To understand the error below one should know what def:ExtendedValue means. As per define_xml specification document, it is an indicator in xml code to check if the codelist is mentioned as Extensible or not in the standard controlled terminology. In the figure below, the top one is issue description from a P21 report, middle one is from controlled terminology document, and the bottom one is from a sample define.xml code using EGTEST codelist.

DD0029	Required attribute def:ExtendedValue is missing or empty	Error	36

Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)	CDISC Definition	NCI Preferred Term
C71152		Yes	ECG Test Name	EGTEST	ECG Test Name	Terminology codelist used with ECG Test Names within CDISC.	CDISC SDTM ECG Test Name Terminology
C116140	C71152		ECG Test Name	Acute Myocardial Ischemia ECG Change	Acute Myocardial Ischemia ECG Change	An electrocardiographic finding assessment of new or presumed new significant ST-segment-T wave (ST-T) changes or new left bundle branch block consistent with acute myocardial ischemia. (Thygesen K, Alpert JS, Jaffe AS, et al.: the Writing Group on behalf of the Joint ESC/ACCF/AHA/WHF Task Force for the Universal Definition of Myocardial Infarction. J Am Coll Cardiol 60(16):1-18, 2012)	Acute Myocardial Ischemia by ECG Assessment

define.xml

```
</CodeList>
<CodeList OID="CL.(EGTEST)" Name="(EGTEST)" DataType="text">
   <EnumeratedItem CodedValue="Overall Interpretation" OrderNumber="51" def:ExtendedValue="Yes"/>
   <EnumeratedItem CodedValue="PR Interval" OrderNumber="52" def:ExtendedValue="Yes"/>
   <EnumeratedItem CodedValue="QT Interval" OrderNumber="53" def:ExtendedValue="Yes"/>
   <EnumeratedItem CodedValue="Sinus Rhythm Normal Y/N" OrderNumber="54" def:ExtendedValue="Yes"/>
   <EnumeratedItem CodedValue="Summary (Mean) QRS Duration" OrderNumber="55"
                   def:ExtendedValue="Yes"/>
   <EnumeratedItem CodedValue="Ventricular Rate" OrderNumber="56" def:ExtendedValue="Yes"/>
   <Alias Name="C71152" Context="nci:ExtCodeID"/>
```

Missing Alias – Define.xml specification document is helpful to understand an error in P21 validation report of a define.xml when Split domains are included in submission. When split domains are included in the package, the P21 tool expects a common label that ties all split datasets together. As per Define_xml specification document, an alias context should be added after last item ref under itemgroupdef code of the split dataset like QSCS below. The below example shows Domain description as alias context.

DD0063	Missing Alias	Error	3

```
<!-- Dataset Definition (QSCS) -->
<ItemGroupDef OID="IG.QSCS"
  Domain="QS"
  Name="QSCS"
  Repeating="Yes"
  IsReferenceData="No"
  SASDatasetName="QSCS"
  Purpose="Tabulation"
  def:Structure="One record per questionnaire per question per visit per subject"
  def:Class="FINDINGS"
  def:CommentOID="COM.DOMAIN.QS"
  def:ArchiveLocationID="LF.QSCS">
  <Description>
    <TranslatedText xml:lang="en">Questionnaire-QSCS</TranslatedText>
  </Description>
  <ItemRef ItemOID="IT.STUDYID" OrderNumber="1" Mandatory="Yes" KeySequence="1"/>
  …
  <ItemRef ItemOID="IT.QS.QSDY" OrderNumber="17" Mandatory="No"
    MethodOID="MT.QSDY"/>
  <ItemRef ItemOID="IT.QS.QSEVLINT" OrderNumber="18" Mandatory="No"/>
  <Alias Context="DomainDescription" Name="Questionnaires"/>
  <def:leaf ID="LF.QSCS" xlink:href="qscs.xpt">
  <def:title>qscs.xpt</def:title>
  </def:leaf>
</ItemGroupDef>
<!-- Dataset Definition (QSMM) -->
<ItemGroupDef OID="IG.QSMM" Domain="QS"
```

eCTD Technical Conformance Guide [10]: The eCTD Technical Conformance Guide (Guide) provides specifications, recommendations, and general considerations on how to submit electronic Common Technical Document (eCTD)-based electronic submissions to the Center for Drug Evaluation and Research (CDER) or the Center for Biologics Evaluation and Research (CBER). Some examples are shown below to demonstrate how the document is helpful in preparation of high quality metadata.

- eCTD only accepts file names in lower case, without special character or blank space
- File names of external file should match with the file names referred in the xml code of the define.xml
- The link to an external file name must be less than 200 characters long.
- eCTD older structure [10] allows for automatic ADaM data validation that includes SDTM AE, DM and EX, but in sponsor's environment, it is a manual process. If sponsor miss this step, there may be additional issues (due to cross reference errors from AE, DM and EX with other ADaM data) that were probably missed out to be fixed or explained in the reviewers guide.

Technical Rejection Criteria [2]: Study data standard catalog must be followed for studies that start after December 17, 2016. Technical rejection criteria document is additional to the existing validation criteria to enforce deadlines. There are two rules so far if not followed will lead to rejection/refusal.

- Rule #1734 – Trial Summary (TS) dataset must be present for each study in Module 4 and 5
- Rule # 1736 – Demographic (DM) dataset, Subject level analysis dataset (ADSL) and define.xml must be submitted in Module 5 for clinical data

The above 2 rules are covered by P21 tool set as below.

Pinnacle 21 ID	Message	Description	Category	Severity
SD1020	Missing DM dataset	Demographics (DM) dataset must be included in every submission.	Presence	Error
AD0001	Missing ADSL dataset	ADaM Subject level (ADSL) dataset should be included in every submission.	Presence	Error
SD1115	Missing TS dataset	Trial Summary (TS) dataset must be included in every submission.	Presence	Error
DD0101	Missing Define.xml	FDA eCTD submissions must include a define.xml file for each study in Module 4 (nonclinical) and Module 5 (clinical)	Presence	Error

Study Data - Technical Conformance guide (TCG) [11]: This guide provides guidance on technical conformance like the exchange format for electronic submissions, study data submission format like the use of SDTM and ADaM models by SDOs, Use of standard terminologies, study data validation and traceability etc.,

As it is very important to use the controlled terminology, we have developed a small utility macro that checks the Controlled terminology in the findings data programmatically. It works by importing latest CT to a sas dataset. See Appendix I

Traceability Issues: It is an important part of a regulatory review to trace the analytical results back to the collected data on CRF. Traceability permits an understanding of the relationships between the analysis results (tables, listings and figures in the study report), analysis datasets, tabulation datasets, and source data, i.e., Report →Define.xml→ADaM→SDTM→raw data. Traceability has been a proven issue for Legacy data conversions. Legacy study data are study data in a format that is not standardized and not supported in FDA review process, and not ever listed in the FDA data standards catalog. Sponsors should use processes for legacy data conversion that account for traceability. As per TCG, there are issues like "Limited traceable path from SDTM to the ADaM datasets" and/or "Limited ability to replicate ADaM datasets (i.e., analysis variable imputation or derived variables) using SDTM datasets" when legacy study data and legacy analysis data are independently converted to SDTM and ADaM formats, respectively, rather than ADaM datasets being created directly from the SDTM datasets (converted from legacy study data). In these situations, it is very important to check if all the raw data is mapped to SDTM during the legacy data conversion process. We have developed a small utility for doing these kinds of checks. See Appendix II. The tool helps to read all existing SDTM logs to find input data library references, to know which all raw input

datasets are being used to generate all sdtm datasets. An excel file output will be generated with columns for 'list of raw data' and 'SDTM mapped'. A third column 'SDTM Suggested' can be manually added for easy comparison.

Other manual checks: Despite the fact many automated tools and standard guides are available to perform validation, manual check of metadata is essential to avoid some common human errors. Example Issue 1: Missing decimal point in MedDRA version e.g., Define.xml showing "20" instead of "20.0" or "20.1". Tools generally don't understand non-standard version references. Example Issue 2: Define.xml references for SDTMIG 3.1.3, while submitted data utilized SDTMIG 3.2 version. Validation of define will result in many unexpected results and need to be fixed.

INTERPRET VALIDATION RESULTS

Validation results from tools like P21 community tool set will have issues reported along with severity (like 'Error', 'Warning', and 'Notice'). P21 interprets errors as issues reported with 100% confidence (e.g., AE start date is after the latest Disposition date), and warnings as potential issues requiring manual review (e.g., EXDOSU value not found in 'Unit' extensible codelist). Review agencies like FDA interprets Error/Warning as issues with severe impact on review process and Reject as critical issues that prevent review and automation processes. Example: See two cases below where FDA Business Rules uses words like "must" and "should" to include a treatment emergent flag and Epoch in SUPPAE and all domains respectively. On other hand, P21 reports these issues with a severity as 'Warning' against considering it as an 'Error' or 'Reject'. Despite having differences in interpretation, sponsors are encouraged to fix as many issues that are fixable and provide cleaner metadata for reviewers.

FDA Business Rule ID	FDA Business Rule	FDA Validator Rule	Domains	SDTM 3.1.2	SDTM 3.1.3	SDTM 3.2
FDAB001	A treatment-emergent flag must be submitted.	A treatment-emergent flag should be included in SUPPAE according to SDTM IG v3.1.2 #8.4.3.	SUPPAE	X	X	X
FDAB022	EPOCH should be included for clinical subject-level observations (e.g., adverse events, laboratory, concomitant medications, exposure, and vital signs).	Variables requested by FDA in policy documents should be included in the dataset. E.g., EPOCH and ELEMENT.	ALL	X	X	X

P21 has following rules that checks against these issues and reports them as warnings.

	SD1077	FDAC021	FDA Expected variable EPOCH not found	Warning	1
	SD1097	FDAC022	No Treatment Emergent info for Adverse Event	Warning	6794

EVALUATE IMPACT OF ISSUES ON DATA

Meetings with review divisions (e.g., CDER and CBER) play a crucial role in evaluating the impact of issues of data on the approval of INDs, NDA or BLAs. Meetings can be organized at different stages of drug development process and they include Pre-IND meetings, end-of-phase2 meetings, and pre-submission meetings and post-submission meetings etc. If sponsors are uncertain about how some issues are going to affect their approval, they are highly encouraged to work with review divisions on the specific issues to avoid delays in approval process. As per FDA's Providing Regulatory Submissions In Electronic Format — Standardized Study Data [3] "Sponsors and applicants may submit technical questions related to data standards at any time to the technical support team identified by each Center (see the Study Data Standards Resources Web page for specific contact information). Sponsors and applicants may also request a separate Type C meeting to discuss substantive data standardization issues for NDAs and BLAs. An example of such an issue might be a sponsor's desire to use a standard (e.g., therapeutic area standard in SDTM format) that is not currently supported by FDA."

FIX CRITICAL ISSUES

Critical issues are issues that will lead to refusal to file the approval. The perfect examples are two issues discussed under Technical Rejection Criteria section of this paper. It is certainly obvious that any known critical issues should be fixed by sponsor, and these are not supposed to be just Not fixed and explained in the reviewers' guide.

EXPLAIN NON-FIXABLE ISSUES

As a part of submission package, reviewers guide is optional, but is highly recommended to include for easy understanding of complex issues associated with data review and validation. If a CRF design or data collection issue is causing multiple validation issues, then this should be explained in detail either in dataset descriptions section (Section 3.3 in clinical SDRG or section 5.2 in ADRG). Some study specific issues could be mentioned in section 4.1. Example1: When compound adverse events (that are combination of simple adverse events) are mapped to a custom Events domain (e.g., XC), and it is used as input to ADAE along with AE, then P21 report on ADAE may throw errors due to the domain variable having values other than AE. Example 2: Study day 0 is not ideal as per SDTM model, but some sponsors may prefer to have study day zero for treatment onset day. This could lead to error in all submitted SDTMs and ADaM study day variables in the p21 validation report.

Generally, explanations for issues from P21 report generated on data are included in reviewers guide. But, sponsors are encouraged to include and explain issues generated on Define xml alone, and data + define.xml issues can also be added to reviewer's guides.

CONCLUSION

It is important to have a balance between use of standard guidance from FDA, and automatic tools like Pinnacle 21 for creation of high quality metadata. All sponsors should educate their Standards implementation teams about the available resources and how they could be used to have successful approvals for submissions for regulatory review.

REFERENCES

[1] Data standards catalog https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm
[2] Technical rejection criteria https://www.fda.gov/downloads/drugs/developmentapprovalprocess/formssubmissionrequirements/electronicsubmissions/ucm523539.pdf .
[3] Providing Regulatory submissions https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm
[4] Define.xml schema validation https://www.cdisc.org/system/files/all/standard_category/application/pdf/definereport_v1_0.pdf
[5] SDRG release package V1.2 http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide
[6] FDA Business rules/Validator rules https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm
[7] SDTM-Metadata submission guidelines https://www.cdisc.org/standards/foundational/study-data-tabulation-model-implementation-guide-sdtmig/metadata-submission
[8] SDTM IG 3.2 https://www.cdisc.org/standards/foundational/sdtmig
[9] Define-XML-2.0-Specification http://www.phusewiki.org/wiki/images/8/89/Define-XML-2-0-Specification.pdf
[10] eCTD Technical conformance guide https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm
[11] Study data – Technical conformance guide https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm .

CONTACT INFORMATION

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

Varun Debbeti
Cytel Inc.
460 Totten Pond Rd, STE 640
Waltham - MA / 02452
Work Phone: 6175287176
Email: varun.debbeti@cytel.com
Web: www.cytel.com

APPENDIX I: Sample code to check XXTEST and XXTESTCD controlled terminology in findings domain

```
*=============================================================================;
*Check Controlled Terminology for LBTESTCD and LBTEST values;
*=============================================================================;
**Read Controlled Terminology;
data ct;
      set inlib.sdtm_terminology_2015_09_25;
run;
**Check CT values for LBTESTCD, LBTEST**;
%macro ctchk (test = ,par= );
data ct_&test;
      set ct;
      Where CTNAME = upcase("&test");
      &par = CTVALUE;
run;
Proc sort data = ct_&test out = ct_&test; by &par;run;
Proc sort data = LB out = paramchk (keep = lbtestcd lbtest) nodupkey;
      by &par;
run;
data ctchk;
      length lbtestcd lbtest $200.;
      merge paramchk (in =a) ct_&test (in = b);
      by &par;
      if a then output ctchk;
      if &par ne ctvalue then
      put "NOTE: &test Controlled Terminology not Used ……  "
            &par = ctvalue = ;
run;
%mend ctchk;
%ctchk (test = lbtestcd, par = lbtestcd);
%ctchk (test = lbtest, par = lbtest);
```

APPENDIX II:

```
*=============================================================================;
*The macro searches in the log files for the word "RAW.XXXX". XXXX is raw dataset
name, RAW is raw library *based on the file name of the SDTM log in which the
corresponding "RAW.XXX" key word is found, the macro assigns the SDTM domain name to
a corresponding raw dataset. There could be multiple SDTM domains that a single RAW
data is mapped to.; *it then outputs an excel file with two columns, 1. Raw data
name 2. SDTM domains it was mapped to;
*=============================================================================;
%macro Read_ASCII(File=, Out=);
*file parameter is input of file names of SDTM log files;
*out parameter is for name of output dataset that has information about where the
raw dataset is mapped to;
    FILENAME ASCII "&path\log\%lowcase(&file)"; /*path where all log files of SDTM
programs are located*/
    DATA &Out;
        length filename Pgmname domain Statement KeyWord  $200 Row Keyword_Index
Line_Number 8;
        %let nKW=2;/*number or raw data available in the source folder*/
        array KW(&nKW) $200  (
                  'raw.ab',/*input the list of  raw datasets manually*/
                  'raw.ae');
        array TX(&nKW) $200  (
                  'raw.ab',
                  'raw.ae');
```

```
        INFILE ASCII LENGTH=LENLINE END=EOF;
        INPUT @1 LINE $VARYING255. LENLINE @;
        Pgmname="&File";
        Line_Number=_n_;
        do k=1 to dim(KW);
            if index(upcase(Line), upcase(strip(TX(k)))) then do;
                Statement=strip(Line);
                KeyWord=KW(k);
                Keyword_Index=k;
                Row=_n_;
            end;
        end;
            if row ne .;
            filename = strip(scan(keyword,2,"."))||".sas7bdat";
            domain = strip(upcase(scan(pgmname,"1",".")));
    RUN;
    proc print data=&out;
        title "&File";
        var Filename KeyWord  Row line;
        ***where not missing(KeyWord);
    run;
%mend;
filename pipedir pipe " dir ""&path\log\*.log"" /b"; /*path where all the SDTM log
files are located*/
data Files_found;
   infile pipedir;
   length fileNAME $ 150;
   input fileNAME $;
   filetype = scan(filename,2,".");
run;
proc sort data=files_found;
    by FileName;
    where upcase(FileType)='LOG'  and length(scan(filename,1,".")) <=4; /*to filter
out NON-SDTM domain logs*/
run;
*call the above macro using call execute statement;
data _null_;
    set files_found;
    call execute('%Read_ASCII(File=' || strip(FileName) || ', Out=Out' ||
strip(put(_n_, z2.)) || ')');
run;
*stack all the output datasets from macro calls;
data mapdomain;
 set out: ;
 N = _n_;
 keep filename domain N;
run;
proc sort data = mapdomain out = mapdomain2 nodupkey;
 by filename domain ;
run;
proc transpose data = mapdomain2 out = dom3 prefix = _;
 by filename;
 id domain;
 var  domain;
run;
data dom4;
 length domain $30.;
 set dom3 (drop = _name_);
 domain = catx("," ,of  _:); /*Concatenate all columns*/
```

```
 filename = upcase(scan(filename,1,"."));
 keep filename domain ;
run;
/*Get raw data list from source folder*/
libname raw "&path\raw";
 create table rawdatalist as
 select unique memname, name, type, label, format
 from sashelp.vcolumn where upcase(libname)='RAW';
quit;
data rawdatalist;
 set rawdatalist;
 rename memname = filename;
run;
proc sort data = rawdatalist nodupkey; by  filename;run;
proc sort data = dom4; by filename; run;
/*merge with the domain list to find which raw data is not mapped to any SDTM*/
/*Assign "NOT SUBMITTED" if the raw data is not mapped to any SDTM*/
data dom5;
 merge rawdatalist (in = a) dom4 (in = b);
 by filename;
 if a;
 if domain = "" then domain = "NOT SUBMITTED";
run;
/*export the data as an excel file*/
proc export
  data=dom5
  dbms=excel
  outfile="&path\domainmapping.xlsx"
  replace;
run;
```