PharmaSUG 2022 - Paper DS-064

# Upgrading from Define-XML 2.0 to 2.1

Trevor Mankus, Pinnacle 21

## ABSTRACT

The Define-XML standard has developed significantly since its original inception in 2005 when version 1.0 was released. Fast forward to this year and versions 2.0 and 2.1 are the industry standard now. However, about 6 years separate the publication of the two versions and much has changed. This presentation will focus on highlighting the new elements and attributes introduced in version 2.1 as well as discuss some best practices that creators should follow in order to upgrade their existing define 2.0 to the latest published version of the standard. In addition, the presentation will show how Pinnacle 21 Enterprise can be used to aid in the up-versioning process using our Excel-like editor and also cover the new validation rules that were implemented so that you can feel confident your upgraded define.xml file complies with the latest guidance.

## INTRODUCTION

In 2019, the CDISC Define-XML Team released a new version of the Define-XML standard and as a result, added many new details that caused the task of creating a define.xml to be more difficult and time consuming. Major changes between Define-XML versions 2.0 and 2.1 are summarized in **Section 1.1.3 Relationship to Prior Define-XML Specifications** of the Define-XML v2.1 PDF document. One major change is the ability for producers of the define.xml document to reference more than just one CDISC standard and controlled terminology. The introduction of this change alone can be disruptive to many existing processes as companies across our industry begin to think about adopting the new version of the standard.

## ASSOCIATING STANDARDS AND CONTROLLED TERMINOLOGIES

In Define-XML v2.0, the define.xml produced could only reference a single standard. Programmers would point to the version of the CDISC standard they followed when they created their data. For example, a define.xml for Tabulation datasets could reference SDTM-IG 3.1.2 and a define.xml for Analysis datasets could reference ADaM-IG 1.1.

This was represented very simply within the MetaDataVersion element at the top of the define.xml:

```
<MetaDataVersion OID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2"
                 Name="Study CDISC01, Data Definitions"
                 Description="Study CDISC01, Data Definitions"
                 def:DefineVersion="2.0.0"
                 def:StandardName="SDTM-IG"
                 def:StandardVersion="3.1.2">
```

Starting with Define-XML v2.1, the **def:StandardName** and **def:StandardVersion** attributes were deprecated and replaced with a more comprehensive solution—the **def:Standards** element. This element contains children **def:Standard** elements; one for each referenced standard used in the study (including controlled terminology).

```
<def:Standards>
    <def:Standard OID="STD.1" Name="SDTMIG" Type="IG" Version="3.1.2"
                  Status="Final" def:CommentOID="COM.STD1"/>
    <def:Standard OID="STD.2" Name="SDTMIG" Type="IG" Version="3.2"
                  Status="Final" def:CommentOID="COM.STD2"/>
    <def:Standard OID="STD.2_1" Name="SDTMIG-MD" Type="IG" Version="1.0"
                  Status="Final" def:CommentOID="COM.STD3"/>
    <def:Standard OID="STD.3" Name="CDISC/NCI" Type="CT"
```

```
                    PublishingSet="SDTM" Version="2011-12-09"
                    Status="Final" def:CommentOID="COM.CT1"/>
      <def:Standard OID="STD.4" Name="CDISC/NCI" Type="CT"
                    PublishingSet="SDTM" Version="2015-12-18"
                    Status="Final" def:CommentOID="COM.CT2"/>
  </def:Standards>
```

The define.xml stylesheet parses this information into a nice table at the top of the document when viewed in a browser.

**Standards for Study CDISC01_1**

| Standard | Type | Status | Documentation |
|---|---|---|---|
| SDTMIG 3.1.2 | IG | Final | The CDISC01 study was modeled on a very old SDTMIG and no attempt was done yet to upversion it to a newer SDTMIG |
| SDTMIG 3.2 | IG | Final | As an example, the CDISC01 study was adjusted to include a new Domain available in SDTM IG 3.2 |
| SDTMIG-MD 1.0 | IG | Final | As an example, the CDISC01 study was adjusted to include a new Domain available in SDTMIG-MD 1.0. The XS Domain is expected to reference the device used with variable SPDEVID. |
| CDISC/NCI SDTM 2011-12-09 | CT | Final | Assuming the CT was not upversioned for this study |
| CDISC/NCI SDTM 2015-12-18 | CT | Final | The CT version applicable for the new Domain is the 2015-12-18 version |

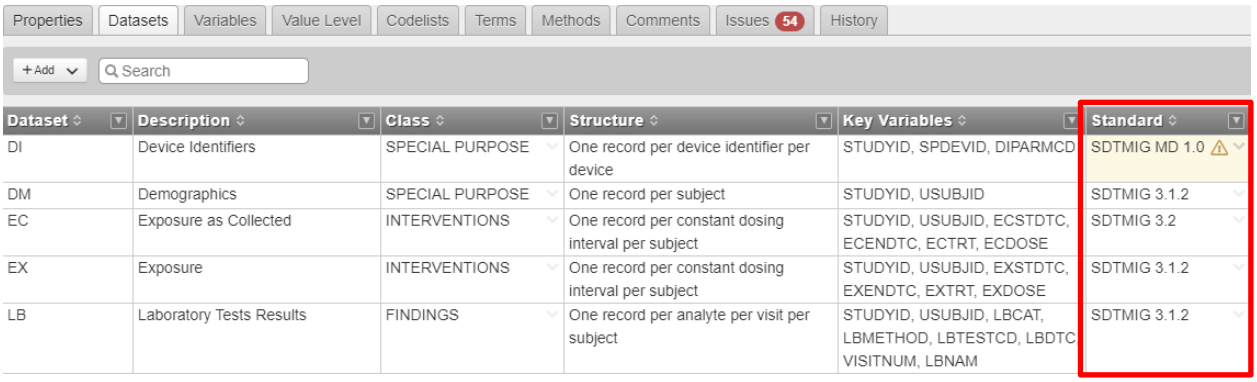
**Figure 1. The "Standards" table as seen when viewing define.xml in a browser**

Unfortunately, this can cause problems for many programmers as they now need to specify this additional metadata somehow in their process which can be both time consuming and also prone to data entry errors. But that's not all—these standards need to be linked to datasets and codelists. These links are made using the **def:StandardOID** attribute within the **ItemGroupDef** element (for datasets) or the **CodeList** element (for codelists):

```
<ItemGroupDef OID="IG.TS" Domain="TS" Name="TS" Repeating="No"
              IsReferenceData="Yes" SASDatasetName="TS"
              def:Structure="One record per trial summary parameter value"
              Purpose="Tabulation" def:StandardOID="STD.1"
              def:ArchiveLocationID="LF.TS">
  <Description>
    <TranslatedText xml:lang="en">Trial Summary</TranslatedText>
  </Description>
  <ItemRef…/>
  <def:Class Name="TRIAL DESIGN"/>
  <def:leaf ID="LF.TS" xlink:href="ts.xpt">
    <def:title>ts.xpt</def:title>
  </def:leaf>
</ItemGroupDef>
```

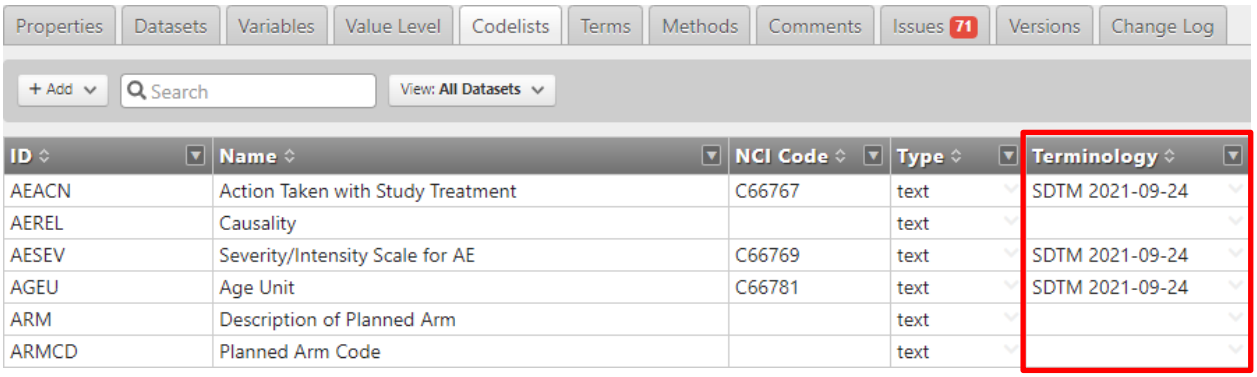
## HOW P21 ENTERPRISE HELPS

Pinnacle 21 Enterprise solved this problem by adding a new column on the Datasets tab called *Standard*. This column contains a drop-down with a controlled list of values that can be selected. The additional properties of these standards (Type, Status, etc.) are managed by Pinnacle 21—removing the burden from the programmer. Invalid combinations and values are shown with in-cell tooltips to help the programmer during define.xml creation. Likewise, a similar column was added to the Codelists tab called *Terminology* which works in the same manner. The absence of a configured value in these columns indicates that the dataset or codelist is non-standard and results in the generated define.xml containing the **def:IsNonStandard** attribute.




| Dataset | Description | Class | Structure | Key Variables | Standard |
|---|---|---|---|---|---|
| DI | Device Identifiers | SPECIAL PURPOSE | One record per device identifier per device | STUDYID, SPDEVID, DIPARMCD | SDTMIG MD 1.0 |
| DM | Demographics | SPECIAL PURPOSE | One record per subject | STUDYID, USUBJID | SDTMIG 3.1.2 |
| EC | Exposure as Collected | INTERVENTIONS | One record per constant dosing interval per subject | STUDYID, USUBJID, ECSTDTC, ECENDTC, ECTRT, ECDOSE | SDTMIG 3.2 |
| EX | Exposure | INTERVENTIONS | One record per constant dosing interval per subject | STUDYID, USUBJID, EXSTDTC, EXENDTC, EXTRT, EXDOSE | SDTMIG 3.1.2 |
| LB | Laboratory Tests Results | FINDINGS | One record per analyte per visit per subject | STUDYID, USUBJID, LBCAT, LBMETHOD, LBTESTCD, LBDTC VISITNUM, LBNAM | SDTMIG 3.1.2 |

**Figure 2. Pinnacle 21 Enterprise's Datasets tab, containing the *Standard* column**




| ID | Name | NCI Code | Type | Terminology |
|---|---|---|---|---|
| AEACN | Action Taken with Study Treatment | C66767 | text | SDTM 2021-09-24 |
| AEREL | Causality | | text | |
| AESEV | Severity/Intensity Scale for AE | C66769 | text | SDTM 2021-09-24 |
| AGEU | Age Unit | C66781 | text | SDTM 2021-09-24 |
| ARM | Description of Planned Arm | | text | |
| ARMCD | Planned Arm Code | | text | |

**Figure 3. Pinnacle 21 Enterprise's Codelists tab, containing the *Terminology* column**

## CLASS AND SUBCLASS

Another new concept introduced in Define-XML v2.1 is SubClass—allowing programmers to provide a more descriptive classification of their datasets. As of today, the only defined use cases for this concept exist for ADaM submissions. A class of BASIC DATA STRUCTURE can have subclass values of NON-COMPARTMENTAL ANALYSIS and TIME-TO-EVENT; a class of MEDICAL DEVICE BASIC DATA STRUCTURE can have subclass values of MEDICAL DEVICE TIME-TO-EVENT; and a class of OCCURRENCE DATA STRUCTURE can have a subclass of ADVERSE EVENT. Ultimately, the list of valid Class and SubClass combinations is driven by the Define-CT that is published quarterly which means programmers need to constantly be aware of new combinations with each publication.

Datasets

| Dataset | Description | Class - SubClass | Structure | Purpose | Keys | Documentation | Location |
|---|---|---|---|---|---|---|---|
| ADSL [ADaMIG 1.1] | Subject-Level Analysis | SUBJECT LEVEL ANALYSIS DATASET | one record per subject | Analysis | STUDYID, USUBJID | Screen Failures are excluded since they are not needed for this study analysis. See referenced dataset creation program and ADRG<br>adsl.sas<br>Analysis Data Reviewer's Guide [6] | adsl.xpt |
| ADQSADAS [ADaMIG 1.1] | ADAS-Cog Analysis | BASIC DATA STRUCTURE | One record per subject per parameter per analysis visit per analysis date | Analysis | STUDYID, USUBJID, PARAMCD, AVISIT, ADT | See referenced dataset creation program and ADRG<br>adqsadas.sas<br>Analysis Data Reviewer's Guide [Section2.1] | adqsadas.xpt |
| ADAE [ADaMIG 1.1] | Adverse Events Analysis Dataset | OCCURRENCE DATA STRUCTURE<br>- ADVERSE EVENT | one record per subject per adverse event | Analysis | STUDYID, USUBJID, AETERM, ASTDT, AESEQ | See SAS program<br>adae.sas | adae.xpt |

Go to the top of the Define-XML document

**Figure 4. The "Datasets" table as seen when viewing define.xml in a browser**

Initial need for this new concept came from a request out of the CDISC ADaM team because in the ADaM standard, there is no dataset naming conventions other than the subject-level analysis dataset (ADSL). However, certain rules exist for structures that meet specific needs—such as the adverse event dataset. Therefore, the define.xml subclass attribute was added to help validation software know which set of rules to execute against specific datasets.

```
<def:Class Name="OCCURRENCE DATA STRUCTURE">
  <def:SubClass Name="ADVERSE EVENT"/>
</def:Class>
```

## HOW P21 ENTERPRISE HELPS

Pinnacle 21 Enterprise solved this problem by adding a new column on the Datasets tab called *SubClass*. This new column has built-in validation that checks for valid combinations of Class and SubClass values. The built-in validation should help programmers catch issues earlier on in the process and more importantly, help ensure data compliance against CDISC by running a comprehensive suite of validation checks.

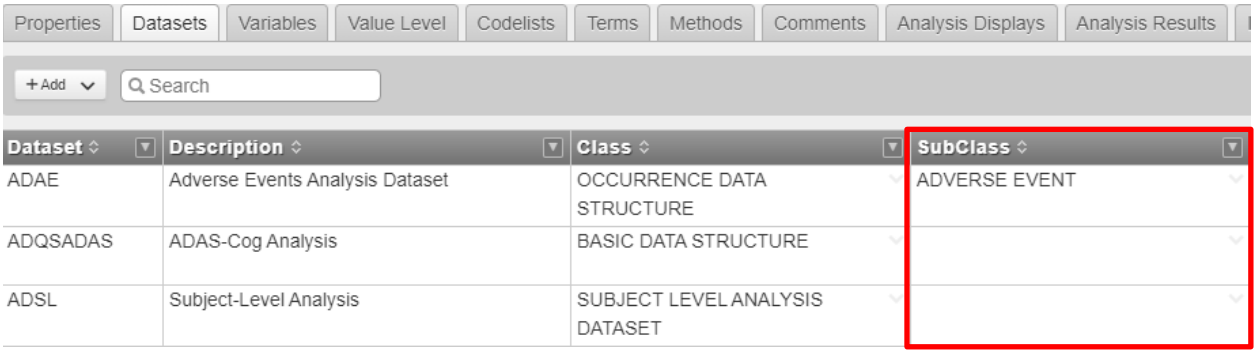



**Figure 5. Pinnacle 21 Enterprise's Datasets tab, containing the *SubClass* column**

## DATASETS AND VARIABLES WHICH HAVE NO DATA

One of the more trivial additions that came with Define-XML v2.1 has to do with identifying empty datasets and variables with no values. This is flagged with a new attribute called **def:HasNoData** on both the ItemGroupDef element (for datasets) and the ItemRef element (for variables). This optional attribute is only needed when the condition is met (i.e., the object contains no data) and will therefore have a value of "Yes". In addition, whenever a dataset or variable is flagged as being empty, a comment becomes conditionally required to explain why no data is present. When this flag is set the define.xml stylesheet will show **[No Data]** next to the dataset or variable to indicate that it is empty.

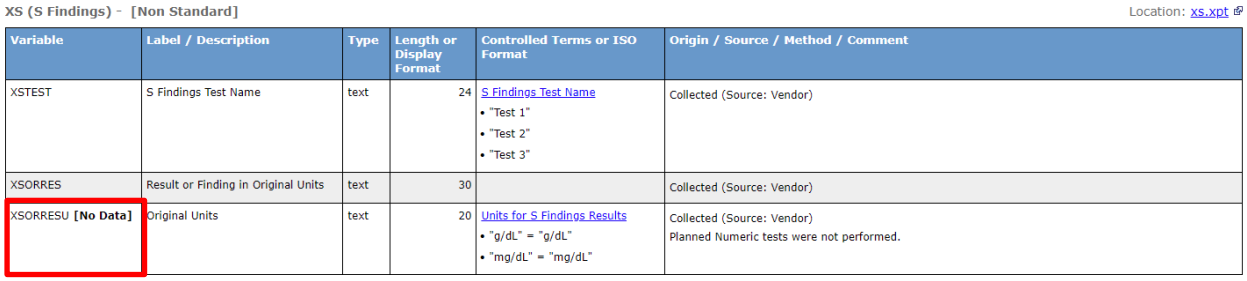



**Figure 6. The "Variables" table in the define.xml showing XSORRESU marked with No Data**

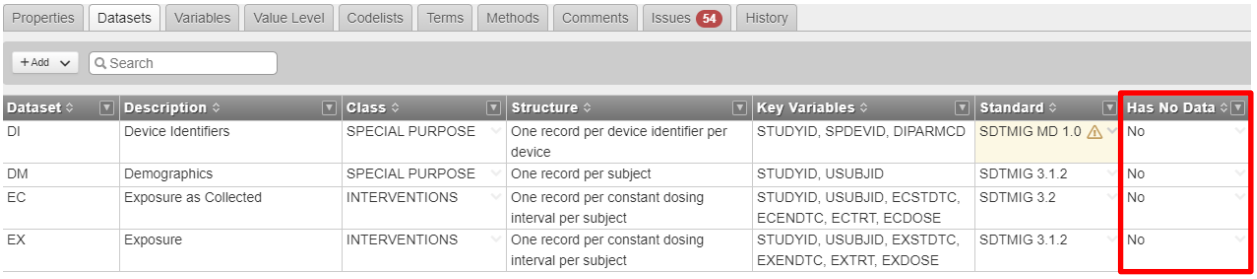
Populating this concept can be incredibly arduous for programmers, especially since variables may remain empty throughout the course of an ongoing study and only get values towards the end as data lock nears. It's also not good programming practice to include empty datasets in your submission package. In fact, versions of the SDTM Implementation Guide up to and including 3.3 had text stating:

> *"In the event that no records are present in a dataset […], the empty dataset should not be submitted and should not be described in the Define-XML document."*

This text was removed in SDTM-IG 3.4; however, still exists in the form of an FDA Validator Rule—*CG0408*.

## HOW P21 ENTERPRISE HELPS

Pinnacle 21 Enterprise solved this problem by adding a new column on the Datasets tab and Variables tab called *Has No Data*. This new column is a Yes/No field that is intended to flag empty datasets or empty variables. When the field is marked "Yes", P21E generates the **def:HasNoData="Yes"** attribute in the define.xml file. Additionally, the Comment field becomes conditionally required to indicate that an explanation is needed for why the dataset or variable is empty.




| Dataset | Description | Class | Structure | Key Variables | Standard | Has No Data |
|---|---|---|---|---|---|---|
| DI | Device Identifiers | SPECIAL PURPOSE | One record per device identifier per device | STUDYID, SPDEVID, DIPARMCD | SDTMIG MD 1.0 | No |
| DM | Demographics | SPECIAL PURPOSE | One record per subject | STUDYID, USUBJID | SDTMIG 3.1.2 | No |
| EC | Exposure as Collected | INTERVENTIONS | One record per constant dosing interval per subject | STUDYID, USUBJID, ECSTDTC, ECENDTC, ECTRT, ECDOSE | SDTMIG 3.2 | No |
| EX | Exposure | INTERVENTIONS | One record per constant dosing interval per subject | STUDYID, USUBJID, EXSTDTC, EXENDTC, EXTRT, EXDOSE | SDTMIG 3.1.2 | No |

**Figure 7. Pinnacle 21 Enterprise's Datasets tab, containing the *Has No Data* column**

## ORIGIN ENHANCEMENTS

Lastly, representation of origin metadata was enhanced to identify the source in addition to the origin details. What we know as Origin in Define-XML v2.0 is now referred to as Type. The *Type* attribute indicates how the data originated while the *Source* attribute identifies the party responsible for the data's origin. Terminology used for both Type and Source is managed and published with Define-CT in the form of non-extensible codelists. Note that legacy value "CRF" was replaced with "Collected":

| Type | Definition |
|---|---|
| Collected | A value that is actually observed and recorded by a person or obtained by an instrument. Note that a collected entry translated to a synonymous controlled term still has a type Collected. |
| Derived | A value that is calculated by an algorithm or reproducible rule, and which is dependent upon other data values, including data values available within the dataset or externally provided data values. |
| Assigned | Data that is either:<br>• Determined by individual judgment as provided by an evaluator, or<br>• Coded terms supplied as part of a coding process, or<br>• Values set independently of any subject-related data value in order to complete a dataset. |
| Protocol | Data that is defined as part of the study protocol, investigator instructions, standard operating procedures or trial design preparation. |
| Predecessor | An entry that is copied from a variable in another dataset. |

**Table 1. Origin Type and Definitions in Define-XML v2.1**

| Type | Source | | | | Notes |
|---|---|---|---|---|---|
| | Subject | Investigator | Vendor | Sponsor | |
| Collected | ePro | CRF | Lab data, ECG | X | This term should be used for clinical data that were actually observed or recorded by a person or received from an instrument; it should not be used for data that have been interpreted, calculated, or derived from other information. |
| Derived | X | X | Lab data, ECG | SDTM | Derivation examples include calculations performed during data collection (e.g., --DY). Other derivation examples: calculations within ePRO (e.g., questionnaire section scores) and calculations within EDC (e.g., BMI, BSA). |
| Assigned | X | X | Adjudicator | SDTM | Examples of this include third-party attributions by an adjudicator, coded terms that are supplied as part of a coding process, and values that are set independently of any subject-related data values in order to complete SDTM fields such as DOMAIN and --TESTCD |
| Protocol | X | X | X | SDTM | An example would be VSPOS (Vital Signs Position), which could be specified in the protocol and be provided by other means (e.g. CRF, eDT). |
| Predecessor | X | X | X | X | Use when a value is an exact copy of another value in an SDTM dataset. |

**Figure 8. Type and Source values for SDTM datasets**

For SEND datasets, only the Type attribute is used. Source is not a value attribute. For ADaM datasets, the matrix is much more simple—Source must equal “Sponsor” when Type is “Derived” or “Assigned”.

### HOW P21 ENTERPRISE HELPS

Pinnacle 21 Enterprise solved this problem by adding a new column on the Variables and Value Level tabs called *Source*. This new column contains built-in validation to ensure the combination of values between Origin and Source is a valid one.

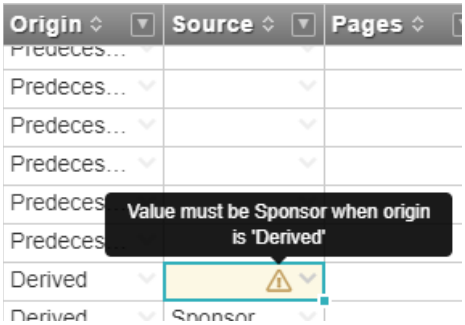



**Figure 9. The “Source” column in Pinnacle 21 Enterprise**

## CONCLUSION

Upgrading your study define.xml from 2.0 and 2.1 can seem like a very tedious and arduous task, but with the right tools and proper knowledge of the changes it can go smoothly. Some fields, such as the HasNoData flag, are best left to be populated towards the end of the study when data is finalized while others make sense to populate as early as possible (e.g., SubClass) to ensure you get proper validation results. The decision on *when* to upgrade is ultimately driven by the regulatory agency requirements. Starting with v5.1, Pinnacle 21 Enterprise helps answer the decision of *when* by allowing programmers to toggle the version of the Define-XML standard they want to follow and curates the user interface accordingly.

## REFERENCES

CDISC Define-XML Team. “CDISC Define-XML Specification Version 2.0.” Accessed March 29, 2022. https://www.cdisc.org/standards/data-exchange/define-xml/define-xml-v2-0.

CDISC Define-XML Team. “CDISC Define-XML Specification Version 2.1.” Accessed March 29, 2022. https://www.cdisc.org/standards/foundational/define-xml/define-xml-v2-1.

## CONTACT INFORMATION

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

Trevor Mankus
Pinnacle 21
tmankus@pinnacle21.com