Automating SDTM File Creation: Metadata Files Speeding the Process

Daphne Ewing, Auxilium Pharmaceuticals, Inc, Malvern, PA

ABSTRACT

For small to mid-size companies that cannot afford to purchase expensive tools to "automate" the SDTM file creation process, tools can be prepared to streamline this process. SDTM conversion can be complicated as clinical study nuances have to be built into the process. Metadata files that are built first by knowing what data are coming from the Clinical Data Management (CDM) system along with describing each CRF field as an SDTM variable begin the mapping process. Adding to these metadata files key information on how to transform the data from its source files into the resulting SDTM target variables round out the metadata files. Clinical Databases do not lend themselves to creation of 100% SDTM compliant files, thus requiring some processing/mapping prior to having SDTM submission ready files. The use of metadata files makes it efficient to create utility programs that will ensure that there are no holes in the process of getting from the source data into the SDTM structure. There are many moving parts to this conversion process, let the metadata self document the process.

INTRODUCTION

The Food and Drug Administration (FDA) and the Clinical Data Interchange Standards Consortium (CDISC) have begun working together to ensure that electronic submissions adhere to standards to make data review more effective and efficient. The Standard Data Tabulation Model (SDTM) created by CDISC serves this purpose. This SDTM standard has been modified several times. In its early inception, there was flexibility at many levels. With additional versions of the model approved, some of the flexibility was removed to get to a more standard structure in the common domains (e.g. DM – Demography, AE – Adverse Events, etc.). By providing the FDA with an electronic submission following CDISC Standards eliminates the need for Case Report Tabulations (CRT) to be submitted. This time consuming step, which is usually a rate limiting step at the end, no longer needs to be done. Planning the submission process and the data contained from the onset of a project lessens the burden at the end of a study/program. This paper is based on the CDISC SDTM Version 1.2 and the SDTM Implementation Guide Version 3.1.2.

UNDERSTANDING SDTM – AN OVERVIEW

The CDISC SDTM Guidelines groups data into files or domains. Each domain is named with two characters. There are several different types of domains described below:

1. Special Purpose (Demography – DM, Comments – CM, Subject Elements – SE and Subject Visits – SV)
2. Interventions (Concomitant Medications – CM, Exposure – EX and Substance Use – SU)
3. Events (Adverse Events – AE, Clinical Events – CE, Disposition – DS, Protocol Deviations – DV, Medical History – MH)
4. Findings (Drug Accountability – DA, ECG Test Results – EG, Inclusion/Exclusion Criterion Not Met – IE, Laboratory Test Results – LB, Microbiology Specimen – MB, Microbiology Susceptibility Test – MS, Pharmacokinetic Concentrations – PC, Pharmacokinetic Parameters – PP, Physical Exam – PE, Questionnaires – QS, Subject Characteristics – SC and Vital Signs – VS)
5. Trial Design Model – TDM (Trial Elements – TE, Trial Arms – TA, Trial Visits – TV, Trial Inclusion Criteria – TI, Trial Summary – TS)

Because clinical study data collected does not always fit into these standards, each domain can also have a relationship domain, called a Supplemental Qualifier, called a SUPPQUAL, to store any additional information related to the original data. Each domain can have its own related file named **SUPP- -** where the **- -** is the two letter domain abbreviation where the "related" data resides.

Each SDTM domain is comprised of variables with specific attributes. Each variable may be: REQUIRED, EXPECTED or PERMISSIBLE. The only key difference between REQUIRED and EXPECTED variables is that EXPECTED variable values are allowed to be missing, while REQUIRED must have a value. PERMISSIBLE variables should ONLY be included if the information contained within them was collected. Some variables in a domain may have values governed by rules called Controlled Terminology. SDTM Version 3.1.2 is designed with less missing values allowed within each domain compared to previous versions of the standard.

ELECTRONIC SUBMISSION PLANNING

When companies submit New Drug Applications (NDAs) to the FDA, they are required to submit the data in SDTM format. For legacy studies that were analyzed prior to this requirement, this step is done retrospectively and a validation plan has to be in place to ensure that the data converted into SDTM can support the submission. However, by planning for the submission up front and getting the data into SDTM compliant files from the start, there are quite a few savings, not only financial but time saving:

1. No retrospective validation needs to be done
2. No patient profile programs/output need to be provided (FDA will prepare what they need from tools they have in place that work with the SDTM files)
3. No Listings need to be provided

Another important step to save time and to self document the traceability of data from collection to analysis is to generate the Analysis data sets directly from these SDTM files. CDISC has also compiled the **A**nalysis **Da**ta **M**odel (ADaM) standards to standardize these derived files that are provided to the FDA in an electronic submission. Figure 1, below shows the relationship between the different types of data and their relationship during submission planning:

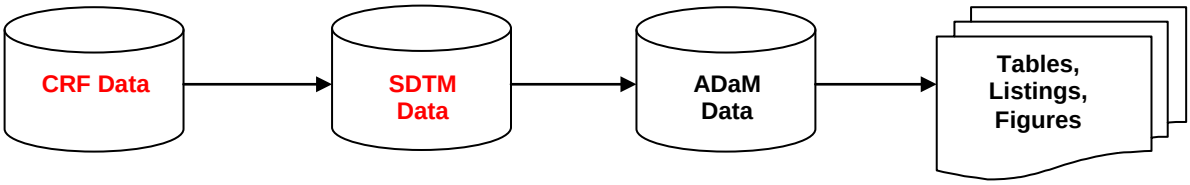

Figure 1. Submission Data/Output Relationships

With the above relationships, a processing flow is established. The data collected on the Case Report Forms (CRFs) has a structure and all the information collected there fits somewhere within the SDTM domain structure. All of the ADaM files can be created from the SDTM files and finally all Tables, Listings and Figures (TLFs) are generated from the ADaM data.

Understanding how to get from one of the above buckets to the next (source to target) is something that has to be documented and validated. Using metadata files that are both created by programs and have information entered by team members will make this process self documenting and provides the ability to validate the information. The rest of the discussion will be regarding getting the data into SDTM format from CRF format and preparing for the submission of the SDTM data.

Planning needs to occur to identify where each data point collected in a study needs to be stored in an SDTM domain. The SDTMIG provides the standard structure for the different variables that are allowed in each of the domains. Care must be taken to ensure that all necessary fields in the SDTM structure are created and that they are being gathered from the appropriate field on the CRF. One of the easiest ways to start this process is to annotate a CRF with SDTM domains and variable names (SDTM aCRF). By doing this, there is a relationship for each field collected and where it will eventually end up in the SDTM files.

DEFINE.XML

The FDA requires a Data Definition Document be provided that describes the data being submitted. The FDA has expanded to accepting Extensible Markup Language (.XML) files for this submission. CDISC has published Case Report Tabulation Data Definition Specification (CRT-DDS) and standard style sheets to help with this process. These can be found on the CDISC website. The full details of the XML file will not be presented in this paper, although some of the required sections are important to note as they can be produced as part of the design process when creating the SDTM files thus supporting the submission during this process. The following are the sections of the Define.XML:

1. SDTM Study Level metadata (Details about the study)
2. SDTM Domain Level metadata (Details about each SDTM domain including the name, a description, structure, a list of key variables and a link to the actual data set, which may be in transport format)
3. SDTM Variable Level metadata (Details about each variable found in each of the above domains including the name, a description, the type, the length and other pertinent information the help a reviewer)

4. SDTM Value Level metadata (Because some of the data are normalized, a Value Level metadata file will support the details required for these normalized domains. This file describes in more detail the different values that exist for TEST/TESTCD combinations within the SDTM Domains. This file can be prepared programmatically from the actual SDTM data.)

1. SDTM Controlled Terminology metadata (Many SDTM variables are defined allowing only certain values be presented…this is what is called Controlled Terminology. This step is a combination of what is contained within the SDTM files, but also has to be extended to include all possible options within the study so a reviewer can better understand the data.)

Having utilities and team members prepare/update the above metadata files as part of the conversion process from CRF data to SDTM data makes the process repeatable and electronically testable easing the manual burden of the process. Having these metadata files as input to the Define.XML process is an efficient/effective way to work as there is no need for retrospective work at submission time.

SOURCE DATA

The data collected on a study CRF is designed to support the protocol and provide information to analyze and prove the hypothesis in the clinical study protocol. Preparing the CRFs in a fully SDTM compliant manner is not only very difficult, but does not lend itself to data clean-up/review activities. There are, however, plenty of data collected during a clinical study that can be databased in SDTM compliant fields (variable names, attributes and even values).

PREPROCESSING CDM DATA

As noted above, the closer the CDM data (structure and values) are to SDTM format, the easier the transformation. However, there are certain fields on the CRF or required in SDTM that are not in a format that ease the conversion or mapping. Some examples of this are the RFSTDTC and RFENDTC in DM, dictionary coding values if they are not part of the CDM files, etc. Having a location to store these preprocessed files and a SAS libname pointing to this location is essential. The default preprocessing is simply making a copy of the CDM file (X_aa) in a preprocessing directory and renaming it slightly (P_aa) to become the source files. A SAS utility program, S_PREP.SAS, identifies all the CDM files and creates one SAS program (P_aa) per X_aa CDM file that simply makes a copy of the file and renames it in the new, SOURCE directory. Any additional pre-processing that is necessary for a given source file can then be placed in this program.

The term SOURCE DATA in this paper refers to the CRF data as it comes from the Clinical Data Management (CDM) tool (e.g. EDC system, ClinTrial, etc.) and then copied into a SOURCE location as noted above. Variables that are SDTM compliant in the CDM database can simply be copied into the appropriate SDTM domain. This is the simplest form of "mapping" data from the SOURCE to the TARGET (SDTM) files. It is an important step to know what data was collected and needs to be converted to SDTM. This is the starting point. Using a SAS utility program to prepare an excel file from the contents of the CRF data with all the necessary information like data set name, variable name, label, type, length, format is a starting point. The excel file created from this process becomes the SOURCE METADATA. This metadata file becomes an interactive tool by adding columns that will facilitate the mapping process. An example of the columns in this metadata file are noted in Figures 2a and 2b below:

MEMNAME	ORDER	NAME	LABEL	TYPE	LENGTH	FORMAT	Action	Con Term
P_AE	1	STUDYID	Study Identifier	CHAR	20		used	N
P_AE	2	S_SITEID	Site Number	CHAR	4		used	N
P_AE	3	S_SUBJID	Subject Number	CHAR	4		used	N
P_AE	4	S_PAGE	Page Number	NUM	8		not used	N
P_AE	5	S_AETERM	AE Verbatim	CHAR	100		used	N

Figure 2a. SOURCE Metadata Example

The SOURCE data set names, in the ***MEMNAME*** column above, are named with "P_" in front of them so as to distinguish them from SDTM files that are the two letter designations. The ***ACTION*** column indicates whether the field is used in SDTM or not and the ***CONTERM*** column identifies if the variable has controlled terminology associated with it. ConTerm values of "Y" are used in another step of the process to check/approve all unique values in the SOURCE files. Programmatically, all unique values from fields with controlled terminology can be determined from the source data and maintained in another metadata file described below. So, although the original metadata file can be produced, and even updated, programmatically, the review of the last two columns is done by a human, but once it is done, need not be done again. The utility program that updates this metadata file can also note in the ***ACTION*** column if a field is new or has been removed from the source data.

Additional fields are also included in this file to define the derivations. Since one SDTM domain may come from multiple source files, identifying the SDTM Domain and SDTM variable name in this file provides the relationship. For SDTM fields that are assigned and do not come from the source data, a row is added here with an Action of Assigned as noted below for the EPOCH variable.

MEMNAME	NAME	SDTM Action	SDTM Domain	SDTM Name	SDTM Rule
P_AE	STUDYID	Copied	AE		
P_AE	S_SITEID	Modified	AE	USUBJID	USUBJID = TRIM(S_SITEID)||'-'||TRIM(S_SUBJID);
P_AE	S_SUBJID	Dependent	AE		
P_AE	EPOCH	Assigned	AE		EPOCH = “FOLLOW-UP”;
P_AE	S_AETERM	Modified	AE		AETERM = STRIP(S_AETERM);

Figure 2b. SOURCE Metadata additional fields Example

TARGET DATA (SDTM STRUCTURE)

The term TARGET DATA in this paper refers to the SDTM data for a particular study. The first step to this process is identifying which SDTM domains need to be prepared for the study. Identifying the variables needed for each SDTM domain is the next step. Referencing the CDISC standards and having a list of all the possible variables for each domain, each CRF field can then be assigned to the appropriate SDTM domain variable. Two metadata files (one for domains, one for variables) can be prepared that serve two purposes:

1. To describe the variables to be created in SDTM structure
2. To support the Define.XML process

So, although the list of all possible domains and all possible variables within each domain can be created from the SDTM documentation, the review and identification of what must be included for a particular study has to be done by a project member familiar with the study and with the SDTM requirements, but once it is completed can be reused for additional studies as well. The key take away from this is that for each study, there has to be review to ensure that all data collected for the study has an appropriate “home” within the SDTM files and this can not requires human review/approval.

As noted above, the CDISC SDTM Version 1.2 document describes data structure for Interventions, Events and Findings classifications of data. The CDISC SDTMIG Version 3.1.2 describes the data structure for the standard domains, which are also classified as Interventions, Events, Findings and a couple of Special Purpose domains (DM and CO). The information contained in these documents can be copied into standard metadata files to be used as pick lists for study specific metadata files. More detail on the two different types of files is described below.

DOMAIN LEVEL TARGET METADATA

A metadata file to house the list of domains (or data sets) needed for a particular study should have enough information to support the needs identified above. Below is an example of the fields necessary in the domain metadata file.

Data set	Title	Class	Structure	Keys	Loca-tion	Data set Order
DM	Demographics	SPECIAL PURPOSE	One record per subject	STUDYID, USUBJID	dm.xpt	1
AE	Adverse Events	EVENTS	One record per adverse event per subject	STUDYID, USUBJID, AEBODSYS, AEDECOD, AETERM, AESTDTC	ae.xpt	10
CM	Concomitant Medications	INTERVENTIONS	One record per recorded medication occurrence per subject	STUDYID, USUBJID, CMDECOD, CMTRT, CMSTDTC	cm.xpt	20
ZE	Primary Efficacy Assessment	FINDINGS	One record per Efficacy Item per visit per subject	STUDYID, USUBJID, ZECAT, ZESCAT, ZELOC, ZETESTCD, ZEDTC	ze.xpt	30

Figure 3. DOMAIN Metadata Example

The above example is abbreviated and contains the DM domain (a special purpose data set), AE (an events data set), CM (an interventions data set) and ZJ (a custom findings domain). The ***Class*** column comes directly from the CDISC standards. The ***Structure*** column describes the layout of the file which supports the Define.XML to help a reviewer understand how the data in the file is laid out. The ***Keys*** column identifies the list of variables to use to order the data set prior to creating the --SEQ value. The ***Data Set Order*** column is not needed for the Define.XML, but has been added to be used by utility programs to ensure that programs are run in a certain order.

The last row in Figure 3 above is an example of a custom domain. Some recommendations on custom domains:

1. Group like datasets with a naming convention, for example use Xa for Interventions, where "a" can be any letter, Ya for Events and Za for Findings.
2. Where possible, use the same custom domain name for the same purpose across multiple studies, for example, if ZE is used for primary efficacy in one study, use it for the same thing across studies.

VARIABLE LEVEL TARGET METADATA

As mentioned above, the SDTMIG document provides the variables and their attributes for each type of SDTM domain (interventions, events and findings) along with specific detail for all possible variables for the primary domains. Although this information is helpful, it is also more information than is necessary for a particular study, and does not provide variable lengths in most cases. Therefore, having a study specific variable level metadata file is helpful for preparing the SDTM domains, checking to be sure they are accurate and then providing input to the Define.XML process including the relationship to codelists/controlled terminology. This can all be done with the same file which again is a time saver. Below is a partial example of a study specific variable level metadata file:

Data set	Variable	Label	Data Type	Len-gth	Origin	Role	Manda-tory
DM	USUBJID	Unique Subject Identifier	text	20	Assigned	Identifier	Yes
DM	SUBJID	Subject Identifier for the Study	text	20	CRF Page	Topic	Yes
DM	RFSTDTC	Subject Reference Start Date/Time	text	19	Assigned	Record Qualifier	No
DM	RFENDTC	Subject Reference End Date/Time	text	19	Assigned	Record Qualifier	No
DM	SITEID	Study Site Identifier	text	4	Assigned	Record Qualifier	Yes
DM	INVNAM	Investigator Name	text	30	Assigned	Synonym Qualifier	No
DM	BRTHDTC	Date/Time of Birth	text	19	CRF Page	Record Qualifier	No

Figure 4. Variable Level Metadata Example

This file appears to be much like the Source metadata file describing the data set name and each variable contained within each domain. In order to fully support the Define.XML process, a few more columns are included in this file (but not shown here). They include:

1. Computation Method/Comment – description of traceability or how the variable was modified/derived from its source – used for Define.XML
2. Code List Name – This is the name associated with the list of possible values for this variable. Most of these come from CDISCs referenced Controlled Terminology, but some of these are related to the CRF values.
3. Display Format – The format (SAS only) needed to display the variable. Typically this is date formats and sometimes used for numeric values to ensure only a certain number decimal points are included.
4. Significant Digits – Related to the Display Format, identifies how many decimal places are needed for the variable.
5. Variable Order – This field is not needed for the Define.XML process, but supports the left to right order of the variables within the file created.

CONTROLLED TERMINOLOGY/CODE LISTS

CDISC has incorporated controlled terminology with many of the fields. In the SDTMIG definition for each of the domains defined is a column indicating controlled terminology. The values for each controlled term can be found in an external excel file also provided by CDISC. This file can be downloaded and used for reference. Some of the data collected in a study may not be part of controlled terminology lists, but the possible values need to be presented as part of the Define.XML to help a reviewer understand the data.

In order to ensure that all data for a study conforms to the controlled terminology and to support the Define.XML process, it is important to know what the source values are for fields that will need Code Lists. Using SAS to first identify each source variable that will need Controlled Terminology (ConTerm=Y in the Source file above), the program can then load all the unique CRF values for these variables into a spreadsheet like the example below:

MEMNAME	NAME	VALUE	CRFORDER	APPROVED
P_DM	ETHNIC	HISPANIC OR LATINO	1	<inits>
P_DM	ETHNIC	NOT HISPANIC OR LATINO	2	<inits>
P_DM	RACE	AMERICAN INDIAN OR ALASKA NATIVE	1	<inits>
P_DM	RACE	ASIAN	2	<inits>
P_DM	RACE	BLACK OR AFRICAN AMERICAN	3	<inits>
P_DM	RACE	NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	4	<inits>
P_DM	RACE	OTHER	6	<inits>
P_DM	SEX	M	1	<inits>
P_DM	SEX	F	2	<inits>

Figure 5. CRF Value Metadata Example

Because the CRF may have possible data values that may not have been selected by any of the subjects in the study, these will have to be added manually to this CRF Value file to accurately represent all the options (supporting Define.XML). Careful review of each of the fields relative to the options on the CRF may identify values or rows in this file that need to be added to ensure that a complete picture is described in the Define.XML for the reviewer. The ***Approved*** column in this file is used to identify rows that are expected in the final file. If there were data entry issues, the row need not be approved, or conversely, if the data doesn't exist in the database, but the value is an option on the CRF, it can be approved. As this iterative process continues, all approved rows are always maintained and the file can be filtered on non-approved rows each time new data arrives to ensure that all possible values are reviewed and acceptable for the submission data.

PROCESSING SOURCE TO TARGET

Knowing what you have, the Source and where you need to get to, the Target, and having this information in machine readable format allows for utilities and processes to be automated and results to be consistent and accurate. The CDM data is typically stored in a consistent location with a SAS libname that will point to it. Similarly, the SDTM resulting files should be placed in a different, but consistent location with a SAS libname that will point to it. Another location is described below that will help the process as well.

TARGET VARIABLE ATTRIBUTES

Since the Variable Level Metadata file defines the files and variables in each of the SDTM files, another SAS utility program can be created. This program, S_ATTRIB.SAS, reads the Variable Level Metadata file and creates one SAS macro (S_aa) per SDTM file that contains SAS ATTRIB statements and a KEEP statement to be used in further processing to ensure that the SDTM file created contains ONLY those variables defined and in the format they were defined.

AUTOMATICALLY GENERATED PROGRAMS TO CREATE SDTM FILES

With these metadata files and macros in place, a SAS Utility program, SPEC2SDTM can be prepared that incorporates the metadata into SAS programs it creates that can then be run to generate the SDTM datasets. This utility program will create a program for each SDTM file that contains:

1. A SAS Program header
2. Call to setup programming environment
3. Create a data step from the Source file for each of the domains providing information for it and apply the Rules in that file within the data step

4. Combine all files into a working copy of the SDTM file
5. Sort the data based on the appropriate information in the SDTM domain Key field
6. Apply the sequence variable
7. Call the attribute macro to ensure the SDTM domain is created with all the appropriate variables/attributes
8. Create the permanent SDTM file with a label as identified also in the SDTM domain file.

After each SDTM program is automatically generated, another utility program is run to identify all the different SDTM programs and creates a batch file to run them all. This .BAT file is then run and the SDTM files are created.

The process of creating SDTM files begins after the CDM data structure is stable and when there is enough data in the CDM database to create all the different files needed but not necessarily when the database is clean and locked. So the above steps are iterated at agreed upon intervals and reviewed to ensure that all new data values are handled appropriately.

Supplemental qualifier data sets are created within this process as well, but the details have not been provided within the scope of this paper.

CONCLUSION

In summary, knowing the source data layout followed by identifying in metadata files the target SDTM structure provides the framework for conversion of data into submission ready files and having the metadata necessary for preparing the Define.XML files necessary for submission. Knowing the relationships between these metadata files allows for automated tools to be prepared to check that all related information paints a clear and accurate picture.

REFERENCES

www.cdisc.org

Study Data Tabulation Model (SDTM) Version 1.2

SDTM Implementation Guideline (SDTMIG) Version 3.1.2

ACKNOWLEDGMENTS

I would like to thank my colleagues at Auxilium Pharmaceuticals, Inc. for working through the details of implementing CDISC/SDTM standards. Additional gratitude goes to staff at PharmaStat, LLC for training and support of our SDTM implementation efforts.

CONTACT INFORMATION

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

Daphne Ewing, Auxilium Pharmaceuticals, Inc.
Voice: (484) 321-5969
FAX: (484) 321-2269
E-mail: dewing@auxilium.com