Paper SI04

Mapping Raw Data to SDTM Define using SAS Clinical Standards Toolkit

Andrew Bullivant, PHASTAR, London, United Kingdom

ABSTRACT

Mapping raw study data to SDTMs can be time consuming and at times inconsistently applied within a study or across studies. This paper describes a method for quickly defining the minimum set of metadata needed to describe the relationship between raw and SDTM. It goes on to describe how that metadata is used to create SAS Clinical Standards Toolkit SDTM metadata datasets which are in turn used to generate Define.xml early in the study reporting process at PHASTAR.

INTRODUCTION

The aim of this paper is to give an overview of the processes followed to generate the metadata required to produce an SDTM define in an automated way utilising the metadata structure and macros available in the SAS Clinical Standards Toolkit (CST). This process has been trialed at PHASTAR.

The first part of the paper describes a simple process for recording metadata that describes the link between raw data variables and SDTM variables. This process also captures the metadata needed to describe the link between raw data and the CRF. It uses an Excel spreadsheet to capture the metadata.

The paper goes on to describe the process used to programmatically create SAS Clinical Standards Toolkit (CST) SDTM metadata from this raw to SDTM mapping spreadsheet in combination with other metadata sources.

DESCRIPTION OF PROJECT FIRST USED IN

There was a desire to evaluate the potential for using the SAS CST within the company standard processes. A suitable project involved the mapping of several studies raw data to produce CDISC SDTM datasets and the Case Report Tabulation (CRT) documentation required for submission.

The studies had been reported previously so the main output from the project was the CRT package.

There were little or no similarities between the studies' raw data. The studies were for the same indications though so mapping decisions needed to be consistent from study to study even though the raw data was not collected consistently.

With these conditions it was important to produce the CRT as efficiently as possible and early in the overall process to allow the team to focus on them. The process needed to efficiently move from study to study and be adaptable to different underlying raw data.

OVERVIEW OF THE PROCESS

The process described in the rest of the paper falls into three main steps:

1. Generation of raw variable to SDTM variable mapping information
2. Creation of SAS CST metadata datasets
3. Generating SDTM Define.xml using the SAS CST define macro

GENERATION OF RAW VARIABLE TO SDTM VARIABLE MAPPING INFORMATION

The studies in the project had the following characteristics:

- raw data was different across studies
- there were unusual raw data types where the mapping to SDTM was not immediately obvious
- raw metadata (e.g. CRF page numbers) was limited or unavailable

A spreadsheet was designed to allow a single reviewer to:

- record raw to SDTM variable link
- record label for SDTM supplemental information.
- identify not submitted raw variables and reason why
- record other SDTM variable metadata (e.g. CRF page number)
- ensure SDTM variable information matches to CDISC SDTM Implementation Guide
- record other useful information for programmers

An example of the spreadsheet is shown on the next page.

The first six columns (MEMNAME, NAME, TYPE, LENGTH, LABEL, FORMAT) are the raw variable metadata available directly from a SAS proc contents of the all raw datasets present.

The columns “eDT or CRF” and “CRF Page” are populated by the reviewer based on a review of the raw data itself and the available raw annotated CRF information.

“SDTM Domain” and “SDTM Variable” is the key information input by the reviewer. Each of these fields have dropdown input lists which are based on the SDTM implementation guide information. The dropdown input list for “SDTM Variable” updates depending on the SDTM Domain chosen.

This forces only standard values to be input into those fields. The conditional input lists are created using Data Validation options in Excel combined with named ranges for the SDTM implementation guide information. With a named range for the domains and then named ranges for each of the Domains which contain the possible variables in each domain.

To aid the reviewer the “CDISC Variable Label” and “CDISC Notes” are auto populated into the spreadsheet based on the SDTM domain and variable chosen. This is done through a VLOOKUP function in Excel.

Information relating to assignment of supplemental variables is described in separate columns.

“Other Details” is used by the reviewer to record reasons for not submitting a raw variable. This column is also used to record important information for programmers about the raw data noted while reviewing the actual raw data (e.g. partial or missing dates noted in the data)

“Other SDTM Variable” records variable names when a single raw variable maps to more than one SDTM variable.

“Where Condition” records information for when a raw variable maps to a subset of records in the resulting SDTM. The remaining columns SUPP QNAM and SUPP QLABEL are used to record the metadata when a raw variable maps to a SDTM supplemental variable rather than a main SDTM variable.

EXAMPLE RAW TO SDTM SPREADSHEET

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	MEMNAME	NAME	TYPE	LENGTH	LABEL	FORMAT	eDT or CRF	CRF Page	SDTM Domain	SDTM Variable	Other details	Other SDTM Variable	Where Condition	Assigned	SUPP QNAM	SUPP QLABEL	SDTM Variable Label	CDISC Notes
1	DMOG	AGE	1	8	Age (Years)		CRF	6	DM	AGE							Age	Age expressed in AGEU. May be derived from RFSTDTC and BRTHDTC, but BRTHDTC may not be available in all cases (due to subject privacy
2	DMOG	ALC_AMT	1	8	Alcohol Amount		CRF	6	SU	SUDOSE			SUTRT="ALCOHOL"				Substance Use Consumption	Amount of SUTRT consumed. Not populated if SUDOSTXT is populated. Units for SUDOSE SUDOSTXT or SUDOSTXT. Examples: OUNCES, CIGARETTE EQUIVALENTS, GRAMS.
3	DMOG	ALC_UNIT	2	15	Alcohol Amount Unit	$	CRF	6	SU	SUDOSU			SUTRT="ALCOHOL"				Consumption Units	
4	DMOG	ALC_USR	1	8	Alcohol User	Y/N	CRF	6	SU	SUOCCUR		SU.SUTRT	SUTRT="ALCOHOL"				SU Occurrence	When the use of specific substances is solicited, SUOCCUR is used to indicate whether or not (Y/N) a particular pre-specified substance was used. Values are null for substances not specifically solicited.
5	DMOG	A_NO	2	6	Volunteer A Number	$	CRF	5	DM						ANUMBER	Volunteer A Number		
6	DMOG	COHORT	2	5	Cohort	$	CRF	5	DM	ARM								
7	DMOG	DOB	1	8	Date of Birth	DATE	CRF	5	DM	BRTHDTC							Date/Time of Birth	Date/time of birth of the subject.
8	DMOG	DTSTOP	1	8	Date Stopped Smoking	DATE	CRF	6	SU	SUENDTC			SUTRT="TOBACCO"				End Date/Time of Substance Use	
9	DMOG	DTSTOPT	2	9	Date Stopped Smoking (text)	$					NOT SUBMITTED - Same information as other field NOT SUBMITTED - Empty in all cases and not on CRF							
10	DMOG	ENTRYCOM	2	300		$												
11	DMOG	RACE	1	8	Race	RACE	CRF	6	DM	RACE							Race	Race of the subject. Sponsors should refer to "Collection of Race and Ethnicity Data in Clinical Trials" (FDA, September 2005) for guidance regarding the collection of race (http://www.fda.gov/RegulatoryInformation/Guidances/ucm126340.htm) See Assumption below regarding RACE

FEEEDBACK ON THE SPREADSHEET

We found that the spreadsheet proved to be a quick way for a single experienced programmer to review all raw data and record their judgement on mapping.

The spreadsheet was flexible enough so that other important metadata can be recorded at the same time.

It was efficient to use the same person across a number of studies to perform the same task. This led to more quickly achieving consistency of mapping decisions across studies.

MAPPING TO SAS CST DATASETS

The next step was to use the raw to SDTM mapping info to generate SAS CST metadata datasets from the CST Toolkit.

DESCRIPTION OF THE CST TOOLKIT

SAS CST Toolkit is a set of metadata definitions and macros to aid reporting clinical trials in SAS. It covers a range of reporting activities e.g. define.xml generation, TFL definition, helpful utility macros such as Replacing Extended ASCII Characters in a SAS Data Set.

SAS CST supports CDISC standards and regulatory requirements for defines. The ability to generate the define.xml was the main driver for investigating the use of SAS CST in this project.

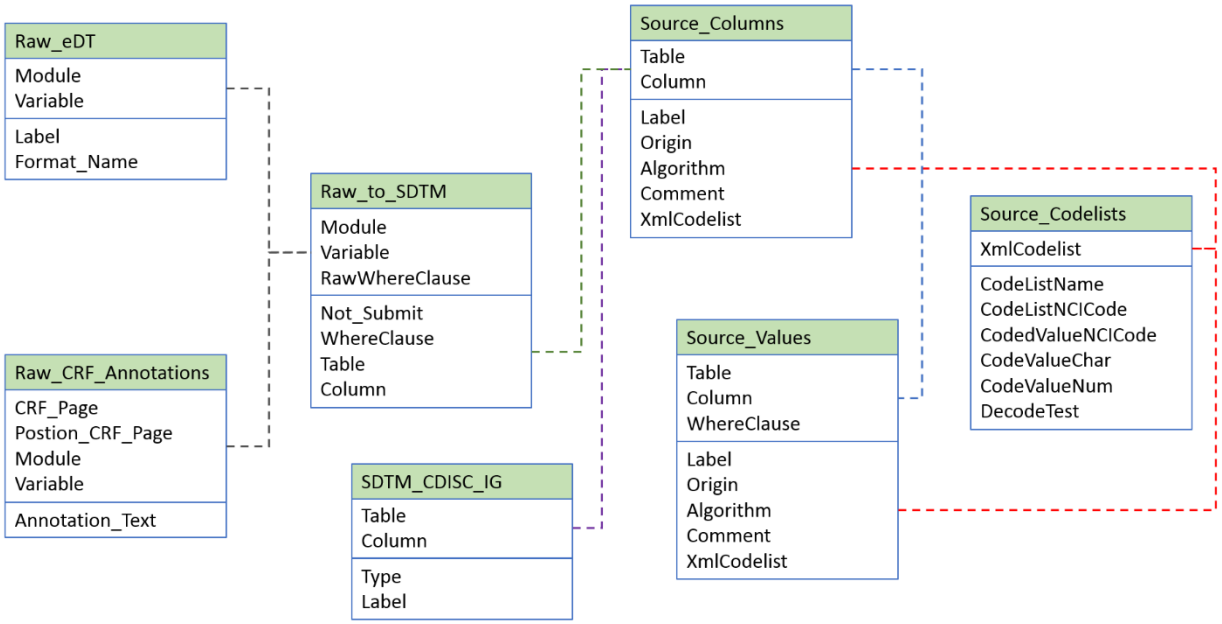
The SAS CST metadata for SDTM and ADaM has six standard SAS CST metadata datasets:

- SOURCE_STUDY – one record containing study description
- SOURCE_TABLES – one record per SDTM dataset
- SOURCE_COLUMNS – one record per variable in each SDTM
- SOURCE_VALUES – value level information required for variables
- SOURCE_CODELISTS – one record per code list item
- SOURCE_DOCUMENTS – document link (e.g. CRF page) info for each variable referencing a document

The following entity relationship diagram shows the relationship between the three key SAS CST metadata datasets involved in the raw to SDTM mapping (SOURCE_COLUMNS, SOURCE_VALUES and SOURCE_CODELISTS) and the metadata collected in the spreadsheet described in the earlier section.

Raw to SDTM mapping is mentioned in the SAS CST but no specific metadata is described. There is a direct relationship between the raw to SDTM mapping and the Source_Columns and Source_Values CST metadata.

ENTITY RELATIONSHIP OF CST AND RAW TO SDTM MAPPING



GENERATING SAS CST METADATA

SAS code was written to combine the raw to SDTM mapping metadata spreadsheet with other raw and SDTM metadata source. This code then went on to output the six SAS CST metadata dataset.

The code used the following metadata inputs:

- RAW to SDTM mapping
- Additional variable metadata information not coming directly from raw data. This covered assigned and derived variables.
- SDTM datasets themselves – list of variables, lists of code list values
- SDTMIG information
- CDISC Controlled terminology – NCI codes

SDTMIG information provided the link between SDTM variables and CDISC Controlled Terminology. This allowed SOURCE_CODELISTS to be populated.

For variables with non-standard code lists these were initially populated from the SDTM datasets that had been generated. For example, QSTESTCD/QSTEST possible values were generated from identifying the unique values of QSTESTCD/QSTEST actually present in the SDTM datasets. If needed additional code list values were input manually in the code.

For cases where code lists were available from both CDISC Controlled Terminology and the SDTM datasets themselves it was possible to force the code to use one or the other. This needed programmer review of the annotated CRF.

Once the SAS CST metadata were created it was found necessary to run a series of edit checks (e.g. ensure all required variables are populated) on the resulting datasets. This allowed the CST metadata to be in a fit state to allow the define generation to occur. Review of the resulting define would produce further updates to the input metadata.

CST METADATA DATASETS TO DEFINE

With the SAS CST metadata ready it was relatively simple to set up and run the SAS CST driver program (create_definexml_from_source.sas) to generate the define.xml. The additional user inputs to the code are:

- update macro parameters to identify standards and versions to be used
- provide location of the CST meta data
- provide location for the reference stylesheet
- provide locations for the output define.xml
- provide location for results information (e.g. error log)

The only place user input was needed was in the start of the driver program. The remainder of the driver program and all macros called were unchanged.

There is large amounts of documentation but we found that practical use of the code was the more informative.

It took a while to get any output produced. This related either to the input datasets or to the SAS environment and not to the SAS CST drivers or macros.

The error tracking information output was useful to find where the issue was occurring and identify possible causes.

The main issue related to our specific SAS environment set up. Once run in a SAS environment with the standard autoexec the majority of issues resolved.

The define.xml validation available with the SAS CST Toolkit was not used and instead Pinnacle 21 validation was used as an added level of independence. The level of errors present was consistent with that of a define.xml created using other processes.

CONCLUSION

Metadata to support raw to SDTM mapping can be relatively easily produced and stored in a way that can be combined with the pre-existing study metadata standards present in SAS CST. This opens up the use of the functionality available from the SAS CST Toolkit.

Having one person generating all the raw to SDTM mapping metadata for a study helped with consistency of mapping decisions across the project.

Standardising raw to SDTM mapping metadata provided great benefits in this project and will continue to help in future projects. Having said that though a great deal of the work of creating the standard metadata would not be necessary if raw data was collected in a standard format from client to client and study to study. The use of CDASH standards in raw data collection would go a long way to removing the work needed in mapping raw data

Standard raw to SDTM metadata in SAS opens up other potential options in the future:

- Automatic generation of SDTM creation SAS code
- Combining mapping information with raw CRF annotation information to automatically generate SDTM annotated CRFs

Recording the raw to SDTM mapping information in this format that allows an expert to more quickly review and compare mapping intentions across studies to ensure consistency across studies.

CONTACT INFORMATION

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

Andrew Bullivant
Phastar
Unit 2, 2A Bollo Ln, Chiswick, London W4 5LE, United Kingdom
City / Postcode
Phone: +44 (0)20 7183 7062
Email: andrew.bullivant@phastar.com
Web: www.phastar.com

Brand and product names are trademarks of their respective companies.