

Automating the blankcrf.pdf and define.xml with your database build

John McDade, Phastar
john.mcdade@phastar.com

Background

At study set-up we receive a raw annotated CRF and database design excel workbook detailing the structure of the raw database with every variable collected on the study, these documents describe the Database Build. The role of these documents are usually confined to aiding the SDTM specification process which then allows us to create the SDTM annotated blankcrf.pdf. What if we could use the database design as the driving force for dynamically creating the define.xml before programming begins?

1 Database Specifications

Database specifications (DB specs) are often an underused asset on a study. Although their primary function is to map out the structure of the database in terms of raw datasets and variables, this document can be a key driver in define.xml creation. Variables in a database are designed to accept free text fields or have an associated codelist.

Below is an example DB spec for a raw AE module, variables which have the ‘Valid Values’ column populated have associated codelists. The variable AETERM has no values populated as this is a free text field but we can see AEACN has finite values. These can be extracted and used to generate the necessary Value Level Metadata and Codelist information for our define.xml.

Variable Name	Variable Label	Type	Length	Collection Length	Format	Valid Values
STUDY	Study Code	C	11	11		
SUBJECT	Enrolment Code	C	8	8		
LINE	Line Number	N	3	3		
AEANY	Any Adverse Events	C	2	2	SN0YES	0=No 1=Yes
AENO	AE Number	N	8	3		
AETERM	Adverse Event	C	200	200		
AESTART	AE Start Date	C	10	10		
AEENDAT	AE Stop Date	C	10	10		
AESER	Serious AE	C	2	2	SN0YES	0=No 1=Yes
AEACN	Action Taken, Investigational Product	C	2	2	SAEACN	0=Dose Not Changed 1=Dose Increased 2=Dose Reduced 3=Drug Interrupted 4=Drug Permanently Discontinued 98=Not Applicable
AEREL	Reasonable Possibility AE Caused by IP	C	2	2	SN0YES	0=No 1=Yes

2 Codelists and Value Level Metadata

The challenge in creating high quality SDTM specifications and define.xml is not in which variables may or may not be present – many companies automate generic specifications to this point which still require much manual intervention. The challenge is automating study specific codelists and value level metadata which is the most labour intensive part of define.xml creation and most poorly executed.

Common problems with codelists include:

- Codelist completely missing when relevant
- Shared codelists across domains when should be specific to domain e.g. for xxORRESU
- Dumping of entire NCI codelist into define.xml

3 Codelist Example

The variables CMDOSU and ODTDOSU below are from distinct raw datasets but both feed into the SDTM variable CMDOSU. The Valid Values column are the possible values when data is entered in this field. We can easily extract this information and combine to build our CMDOSU codelist. The same process will be followed for all variables with codelists.

Note that the NCI codes are included with this database specification, this is not always the case but we can merge to a specific version of the NCI controlled terminology to obtain these.

Variable Name	Variable Label	Type	Length	Collection Length	Format	Valid Values
CMDOSU	Dose Unit	C	6	6	SN0YES	0=No 1=Yes
OTDOSU	Total Dose Unit	C	6	6	SN0YES	0=No 1=Yes

Variable Name	Variable Label	Type	Length	Collection Length	Format	Valid Values
CMDOSU	Dose Unit	C	6	6	SN0YES	0=No 1=Yes
OTDOSU	Total Dose Unit	C	6	6	SN0YES	0=No 1=Yes

4 Value Level Example

The level of detail required in the define.xml value level is somewhat left to the discretion of the creator but some general guidance is given in the Define.XML Specification v2.0 and example define file, this can be used as a basis in the automation process.

For findings domains, the value of xxORRES should be given in terms of xxTESTCD where different metadata attributes are expected per test code value. The other minimum requirement is the SUPP variables described in terms of QNAM value.

Taking the Vital Signs domain as an example, below is a list of possible tests in the DB spec. These can be extracted and will be used for a dual purpose, this will generate value level information and also the VSTEST and VSTESTCD codelists.

Variable Name	Variable Label	Type	Length	Collection Length	Format	Valid Values
VSTEST	Test	C	7	7	SVSTEST	C25206=Temperature C25208=Weight C25298=Systolic Blood Pressure C25299=Diastolic Blood Pressure C49676=Pulse Rate C49678=Respiratory Rate

Variable	Where	Type	Length / Display Format	Controlled Terms or Format	Origin	Derivation/Comment
VSORRES	VSTESTCD EQ TEMP (Temperature)	text	200		CRF Pages 28 31 35 43	
VSORRES	VSTESTCD EQ WEIGHT (Weight)	text	200		CRF Pages 28 31 35 43	
VSORRES	VSTESTCD EQ SYSBP (Systolic Blood Pressure)	text	200		CRF Pages 28 31 35 43	
VSORRES	VSTESTCD EQ DIABP (Diastolic Blood Pressure)	text	200		CRF Pages 28 31 35 43	
VSORRES	VSTESTCD EQ HEIGHT (Height)	text	200		CRF Pages 28 31 35 43	
VSORRES	VSTESTCD EQ PULSE (Pulse Rate)	text	200		CRF Pages 28 31 35 43	
VSORRES	VSTESTCD EQ RESP (Respiratory Rate)	text	200		CRF Pages 28 31 35 43	

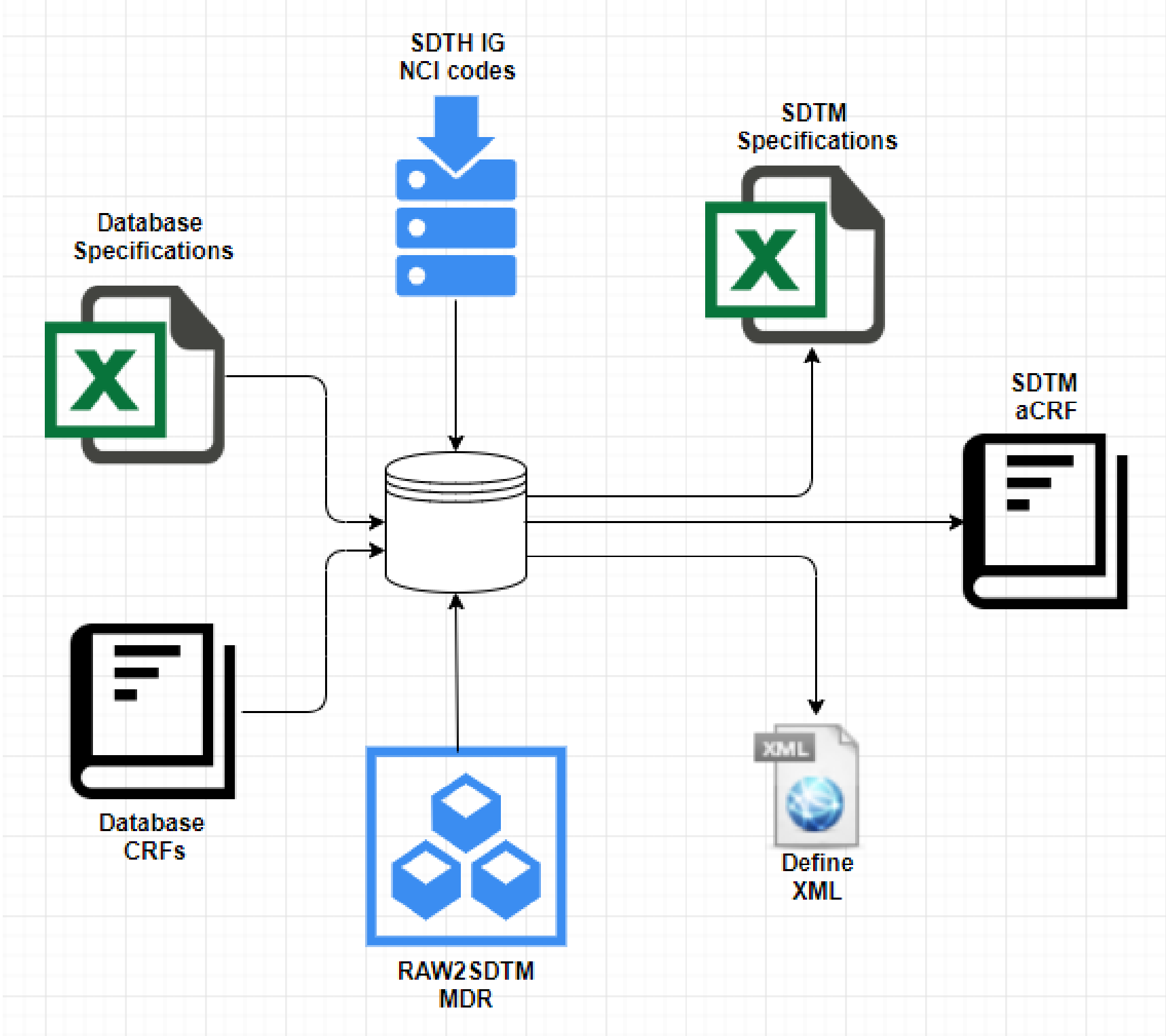
5 Blankcrf Automation

Depending on the CRF type required to annotate, the database annotated CRF and blank CRF can be used as inputs to the process to extract page numbering and positional information for SDTM aCRF.

The CRF on the left below has hyperlinks from which we can extract their y-coordinate on the PDF page. Combining this with an allocated x-coordinate, page numbers and the SDTM annotation can build an FDF file which is simply imported on the blank CRF.

6 Summary

An overview of the full process is given below, in addition to the Database specifications and CRFs previously mentioned, there is also a link to CDISC and raw mapping metadata. Important to note this process is independent of SDTM data so is produced before programming starts.



The use of CDISC metadata gives us a dynamic link to any version of the IG and controlled terminology.

The raw mapping file is simply a file linking database variables to their SDTM counterparts, this is commonly available as a corporate standard file within Pharma companies to ensure consistent mapping across vendors and studies.

