

CDISC Variable Mapping and Control Terminology Implementation Made Easy

Balaji Ayyappan, Ockham Group, Cary, NC
Manohar Sure, Ockham Group, Cary, NC

ABSTRACT:

CDISC SDTM (Study Data Tabulation Model) standards have become a widely accepted form of representation of the primary component of CRT - data tabulation datasets. The core challenge in converting non-CDISC-compliant clinical data from ongoing studies and legacy studies is mapping of data to CDISC compliant domains and variables. It can be time consuming as the information lies in multiple places – clinical datasets, metadata, CRFs, protocol, SAP, SDTM implementation guide, etc. If only there was a tool bringing most of the required components on one screen, enabling you to dynamically map raw data components to SDTM compliant components using these pieces of information. In quest to achieve this we have created a utility that makes this process less frustrating and almost enjoyable. In less time, documentation of how the CDISC variables are mapped to the raw dataset variable(s) and implementing CDISC control terminology to raw dataset variable values and SAS format programs are created using this tool. The tool generated documents are used for creating define.xml file as a byproduct. By automating the process, this tool helps eliminate human error. The technologies used in this utility are SAS, Microsoft Excel, VBA and SAS IOM.

INTRODUCTION:

The CDISC Variable mapping document describes how the CDISC variables are mapped to the raw dataset variable(s) for each domain at study level. Standardizing the unique values of certain variables in standard CDISC domains are documented in Control Terminology (CT).

SAS INTEGRATED OBJECT MODULE (IOM):

The SAS IOM is a data provider that supports access to SAS data sets that are managed by SAS Integrated Object Model (IOM) servers. From Excel, the ADO Connection Object is used to create an open connection to a data source using SAS IOM. Through this connection, one can access and manipulate a SAS database. The utility setup is made in Excel Workbook.

Using the above technology, Excel and SAS IOM servers are connected; codes are written in Visual Basic Access (VBA). Excel-VBA Program is referenced to SAS IOM to execute SAS code using ADO (ActiveX Data Objects) DB connections in the windows environment.

CDISC CONTROLLED TERMINOLOGY (CT):

Latest CT is downloaded from www.cdisc.org and converted into individual format named SAS datasets. Dataset name and format name in the Pre-defined Data Definition will be same.

PROCESS:

PRE-DEFINED DATA DEFINITION SPREADSHEET:

This Excel workbook will have SETUP spreadsheet to fill the Project Directory paths (Directory like: Raw Dataset, CDISC Dataset, SAS Program, etc.) at study level. Screenshot of SETUP spreadsheet is shown below.

| | A | B | C |
|---|-------|-------------------------|---------------------------------------|
| 1 | STUDY | PROJ_DATA | PROJ_VALUE |
| 2 | N105 | PROTOCOL | N105 |
| 3 | N105 | RAW DATASETS LOCATION | D:\Projects\Data\Master Datasets\N105 |
| 4 | N105 | MACROS LOCATION | D:\Projects\Setup\Macros |
| 5 | N105 | SAS PROGRAM LOCATION | D:\Projects\StatProg\Analysis\N105 |
| 6 | N105 | CDISC LOCATION | D:\Projects\Data\CDISC\N105 |
| 7 | N105 | VALIDATION LOG LOCATION | D:\Projects\Data\VAL\N105 |

In INDEX spreadsheet, the information of related raw datasets for each domain is mapped at study level. Screenshot of INDEX spreadsheet is show below.

| STUDY | CDISC Dataset | CDISC Dataset Label | CDISC Category | Data1 | Data2 | Data3 | Data4 |
|-------|-------------------------|-------------------------|-------------------------|-------------|-------------|--------|-------|
| N105 | DM | Demographics | Special-Purpose Domains | DEMOGRAPHIC | RANDOM | | |
| N105 | CM | Concomitant Medications | Interventions | CMT | | | |
| N105 | EX | Exposure | Interventions | RANDOM | MED | SMISMI | |
| N105 | AE | Adverse Events | Events | AESAES | | | |
| N105 | DS | Disposition | Events | END | | | |
| N105 | EG | ECG Test Results | Findings | AESAES | | | |
| N105 | LB | Laboratory Test Results | Findings | LABORATORY | URIN | | |
| N105 | VS | Vital Signs | Findings | VS | DEMOGRAPHIC | | |
| N105 | SC | Subject Characteristics | Findings | DEMOGRAPHIC | | | |
| N105 | SUPQUAL | Supplemental Qualifiers | Supplementary Qualifier | | | | |

Separate sheets for each Domain which are to be mapped will be present in this Workbook. For each domain there will be pre-defined columns like CDISC variable name, label, type, length, CT format if applicable (AE Domain Pre-defined spreadsheet is shown below.) Two blank columns in which we will be mapping the Raw Dataset variables and Comments, dynamically.

| | A | B | C | D | E | G | I | J | K | L |
|----|----------|---------------------------------|------|-------|---------|--------------------|------|-----------|--------------|---------------|
| 1 | VARNAM | VARLBL | TYPE | LEN | FMT | ROLE | CORE | VARS_KEEP | N105_DATAMAP | N105_COMMENTS |
| 2 | STUDYID | Study Identifier | Char | \$6 | | Identifier | Req | X | | |
| 3 | DOMAIN | Domain Abbreviation | Char | \$2 | AE | Identifier | Req | X | | |
| 4 | USUBJID | Unique Subject Identifier | Char | \$16 | | Identifier | Req | X | | |
| 5 | AESQ | Sequence Number | Num | 6 | | Identifier | Req | X | | |
| 6 | AEGRPID | Group ID | Char | \$10 | | Identifier | Perm | X | | |
| 7 | AEREFID | Reference ID | Char | \$10 | | Identifier | Perm | | | |
| 8 | AESPID | Sponsor-Defined Identifier | Char | \$10 | | Identifier | Perm | | | |
| 9 | AETERM | Reported Term for Adverse Event | Char | \$100 | | Topic | Req | X | | |
| 10 | AEMODIFY | Modified Reported Term | Char | \$100 | | Synonym | Perm | | | |
| 11 | AEDECOD | Dictionary-Derived Term | Char | \$100 | * | Synonym Qualifier | Req | X | | |
| 12 | AECAT | Category for Adverse Event | Char | \$50 | * | Grouping Qualifier | Perm | | | |
| 13 | AESCAT | Subcategory for Adverse Event | Char | \$50 | * | Grouping Qualifier | Perm | | | |
| 14 | AEPRESP | Pre-specified Adverse Event | Char | \$1 | (NY) | Record Qualifier | Perm | | | |
| 15 | AEBODSYS | Body System or Organ Class | Char | \$100 | (SOC) | Record Qualifier | Exp | X | | |
| 16 | AELC | Location of the Reaction | Char | \$50 | * | Record | Perm | | | |
| 17 | AESEV | Severity/Intensity | Char | \$20 | (AESEV) | Record Qualifier | Perm | X | | |
| 18 | AESER | Serious Event | Char | \$1 | (NY) | Record Qualifier | Exp | X | | |
| 19 | AEACN | Action Taken with Study | Char | \$30 | (ACN) | Record | Exp | X | | |

After mapping Project Directory Locations and Raw Dataset then following process are performed.

- Using SAS IOM Technology, connection between the Excel Spreadsheet VBA and SAS is established.
- Library References are made to Raw Dataset Location and CT dataset. Excel VBA is connected to SAS through ADO DB Access.
- Parameters are passed to SAS MACROS from VBA and executed.
- SAS OUTPUT DATASETS are read from VBA and displayed in VBA screen.

(Snippets of VBA code is described above are shown below.)

```

Dim obWS As SAS.Workspace
Dim obWSMgr As New SAS.WorkspaceManager.WorkspaceManager
Dim obDS As SAS.DataService
Dim obLibRef As SAS.Libref
Dim obConnection As New ADO.DB.Connection
Dim obRecordSet As New ADO.DB.Recordset

Public Function Submit_SAS

    Set obWS = obWSMgr.Workspaces.CreateWorkspaceByServer _
        ("", visibilityProcess, Nothing, "", "", xmlInfo)
    ..

    Set obDS = obWS.DataService
    'Assign Library Reference to the Directory
    Set obLibRef = obDS.AssignLibref("Raw", "", _
        dataloc, "")

    ' Print the Library Name
    Debug.Print obLibRef.Name
    ..

    Call Submit_SAS("proc sql; create table tds as select memname from sashelp.vtable_
        where upcase(libname)='RAW' and upcase(memtype)='DATA'; quit;")

    obConnection.Open "provider=sas.iomprovider.1; SAS workspace ID=" + obWS.UniqueIdentifier

    Set fld = obRecordSet.Fields
    ..
    ReDim vars(RecCnt, VarCnt) As String

    ' Move to the First Record
    obRecordSet.MoveFirst

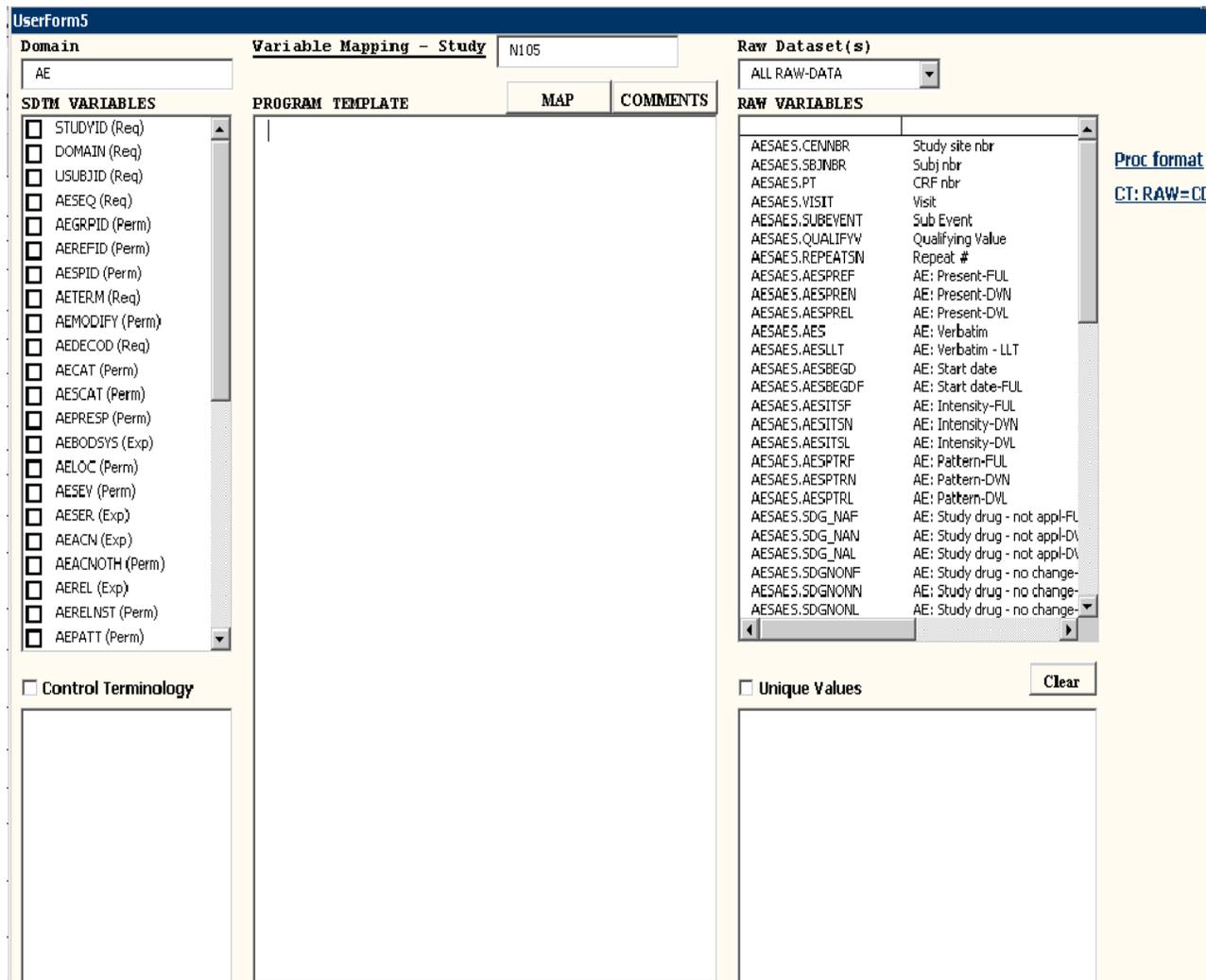
    UserForm3.ComboBox4.Text = ActiveSheet.Range("H" & ar).Value
    UserForm3.ComboBox5.Text = ActiveSheet.Range("I" & ar).Value
    ..|

```

Once the connection between VBA and SAS is established, VBA macros are executed which opens the pre-defined screen for the corresponding domain (Suppose if we are mapping for AE domain, displayed screen will be filled with relevant AE information's.)

- CDISC variable for the corresponding domain which has to be mapped are listed in one section of the screen. This information is fetched from spreadsheet using VBA code.
- Second section will list the CT values for the selected CDISC variable. Getting this information from CT SAS datasets directory which is created using CDISC CT spreadsheet.
- According to the raw dataset mapped for this domain, all the corresponding raw variables, labels and formats will be listed in the third section.
- Fourth section will list the unique of both format and informat values for selected raw variable.
- SAS format is created dynamically by selecting the raw value and CT values in the middle of the program template screen.

Please check the snapshot of the screen shown below with the CDISC variables and mapped Raw Dataset variables for AE Domain. Initially CT and Raw value listbox will be blank.



When the CDISC variable is selected, if formats are applicable then its CT values are listed down and unique raw values are listed for the selected raw variable. Select the required values and click "Proc format" key, it appears in the middle section and automatically documents the unique values of raw variable with the applied control terminology values in the CT spreadsheet. In the program template section the SAS format code is created automatically, which can be converted into permanent SAS Catalogue.

Select unique CDISC variable and the corresponding Raw Variable to be mapped, click the "MAP" command, it automatically documents the mapping information for the selected CDISC variable along with the selected raw variable(s) in the backend of Domain sheet. The SAS Format which is created in the program template section will be saved as a SAS program and then

we can create SAS catalogue using the format program. Snapshot of the screen is shown below.

The screenshot shows the 'Userform5' application window. On the left, a 'Domain' list contains various CDISC variables like STUDYID, DOMAIN, USUBID, AESEQ, AEGRPID, AEREFID, AESPID, AETERM, AEMODIFY, AEDECOD, AECAT, AECAT, AEREL, AEPRESP, AEBODSYS, AELOC, AESEV, AESER, AEACH, AEACHOTH, and AEREL. The 'VARNAM' column is highlighted. In the center, the 'Variable Mapping - Study' section shows a 'PROGRAM TEMPLATE' with a 'MAP' button. On the right, the 'Raw Dataset(s)' section shows a list of 'RAW VARIABLES' with their corresponding labels and a 'Unique Values' section below it. Annotations with arrows point to specific elements: 'CDISC variables from Pre-defined Data Definition File' points to the 'VARNAM' column; 'CT values from CT Directory for selected CDISC variable' points to the 'AELOC' variable's 'SEVERE' value; 'Raw Dataset Variable Names and Labels' points to the 'RAW VARIABLES' list; and 'Unique Raw values for Selected Raw Variable' points to the 'Unique Values' section.

CDISC Variable Mapping Document will be created as shown below for each domain.

| | A | B | C | D | E | G | I | J | K | L |
|---|----------|---------------------------------|------|-------|---------|-----------------------|------|-----------|---|---------------|
| | VARNAM | VARLBL | TYPE | LEN | FMT | ROLE | CORE | VARS_KEEP | H105_DATAMAP | H105_COMMENTS |
| 1 | STUDYID | Study Identifier | Char | \$6 | | Identifier | Req | X | | "N105" |
| 2 | DOMAIN | Domain Abbreviation | Char | \$2 | AE | Identifier | Req | X | | "AE" |
| 3 | USUBID | Unique Subject Identifier | Char | \$16 | | Identifier | Req | X | AESAES.CENNR(\$10.), AESAES.SBJNR(\$15.) | |
| 4 | AESEQ | Sequence Number | Num | 5 | | Identifier | Req | X | | DERIVED |
| 5 | AEGRPID | Group ID | Char | \$10 | | Identifier | Perm | X | | |
| 6 | AEREFID | Reference ID | Char | \$10 | | Identifier | Perm | | | |
| 7 | AESPID | Sponsor-Defined Identifier | Char | \$10 | | Identifier | Perm | | | |
| 8 | AETERM | Reported Term for Adverse Event | Char | \$100 | | Topic | Req | X | AESAES.AES(\$200.) | |
| 9 | AEMODIFY | Modified Reported Term | Char | \$100 | | Synonym | Perm | | | |
| 0 | AEDECOD | Dictionary-Derived Term | Char | \$100 | * | Synonym Qualifier | Req | X | AESAES.AESPT_(\$150.) | |
| 1 | AECAT | Category for Adverse Event | Char | \$50 | * | Grouping Qualifier | Perm | | | |
| 2 | AESCAT | Subcategory for Adverse Event | Char | \$50 | * | Grouping Qualifier | Perm | | | |
| 3 | AEPRESP | Pre-specified Adverse Event | Char | \$1 | (NY) | Record Qualifier | Perm | | | |
| 4 | AEBODSYS | Body System or Organ Class | Char | \$100 | (SOC) | Record Qualifier | Exp | X | AESAES.AESSOC(\$150.) | |
| 5 | AELOC | Location of the Reaction | Char | \$50 | * | Record Qualifier | Perm | | | |
| 6 | AESEV | Severity/Intensity | Char | \$20 | (AESEV) | Record Qualifier | Perm | X | AESAES.AESITSF(\$15.) | |
| 7 | AESER | Serious Event | Char | \$1 | (NY) | Record Qualifier | Exp | X | AESAES.SAEPREF(\$NO_YES.) | |

CT document created while mapping the raw values to CT values is shown below. This CT sheet will have raw dataset format and informat values and applied control terminology values from all the domains.

| A | B | C | D | E | F | G | H | I | J |
|-------|--------|----------------------|------------------------|--|-----------------|---------------------------------|-------------------------------|---------------------------|---------|
| Study | Domain | CDISC Format Name | CDISC Variable Name | Raw Variable Name (Label) | Raw Variable | Raw Variable Informat Values | Raw Variable Format Values | Controlled Terminology | Comment |
| N105 | AE | (NY) | AESER | AESAES.SAEPREF (AE: SAE - present-FUL) | char | NO | No | N | |
| N105 | AE | (NY) | AESER | AESAES.SAEPREF (AE: SAE - present-FUL) | char | YES | Yes | Y | |
| N105 | AE | (AESEV) | AESEV | AESAES.AESITSF (AE: Intensity-FUL) | char | MIL | Mild | MILD | |
| N105 | AE | (AESEV) | AESEV | AESAES.AESITSF (AE: Intensity-FUL) | char | MOD | Moderate | MODERATE | |
| N105 | AE | (AESEV) | AESEV | AESAES.AESITSF (AE: Intensity-FUL) | char | SEV | Severe | SEVERE | |

CONCLUSION:

This tool is useful for creating CDISC variable mapping and CT mapping, with no typing errors. This tool will reduce the time drastically for preparing documentation. These documents can be used as source for creating DEFINE.XML file. We can also create SAS program templates using the tool and can add many features to this tool.

REFERENCES:

Using VBA and BASE SAS to Get Data from SAS to Excel without Data Integrity Issues
<http://www.lexjansen.com/phuse/2005/as/as11.pdf>

Converting CDISC Controlled Terminology to SAS Formats
<http://www.lexjansen.com/pharmasug/2009/ad/ad11.pdf>

ACKNOWLEDGEMENTS:

We would like to thank Paul Ragland and Nivethitha Manickavasagar for providing us excellent ideas and helping to implement this utility.

CONTACT INFORMATION:

Your comments and questions are valued and encouraged. You can contact us at:

Balaji Ayyappan
 Ockham Inc.
 8000 Regency Parkway, Suite # 360
 Cary, NC 27518
 Phone: (919) 653 6090
 Email: bavyappan@ockham.com

Manohar Sure
 Ockham Inc.
 8000 Regency Parkway, Suite # 360
 Cary, NC 27518
 Phone: (919) 380 5665
 Email: msure@ockham.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration