PhUSE 2013

Paper TS03

A How-To Guide for Extending Controlled Terminology Using SAS Clinical Data Integration

Melissa R. Martinez, SAS Institute Inc., Cary, North Carolina, United States

ABSTRACT

SAS® Clinical Data Integration offers the ability to use additional controlled terminology beyond that provided by CDISC in order to standardize and validate data values. Terminology commonly used in this way includes MedDRA codes and the WHO Drug Dictionary, but even customized, company-specific terminology can be implemented. Users can also register newer versions of CDISC-published terminology that is not included in their version of SAS® Clinical Standards Toolkit. This paper fully describes the steps necessary to import extended controlled terminology into SAS Clinical Data Integration, use the terminology for compliance checks and lookups, and include or exclude the terminology from the define.xml document.

INTRODUCTION

SAS Clinical Data Integration provides a user-friendly graphical user interface that facilitates the transformation of raw clinical data into standardized, submission-ready data sets and related files. Use of SAS Clinical Data Integration tends to focus on the transformation aspect, but this application also makes using controlled terminology simple if you understand the underlying processes. This paper will first give an overview of controlled terminology and how it is used within the application. Then the underlying files and folder structures that are relevant to modifying controlled terminology are described in detail. Finally, detailed instructions on extending controlled terminology are given, from importing new versions to customizing a version for a specific study. This paper refers to SAS Clinical Data Integration 2.3, which uses SAS Clinical Standards Toolkit 1.4.

OVERVIEW OF THE PROCESS OF USING CONTROLLED TERMINOLOGY

WHAT IS CONTROLLED TERMINOLOGY?

Controlled terminology is a list of standardized possible values for a variable. There are several dozen controlled terminology codelists that SAS Clinical Standards Toolkit supports; a full list of them as applicable to each data standard can be found in the SAS Clinical Standards Toolkit User's Guide (see References). SAS Clinical Data Integration leverages the functionalities of SAS Clinical Standards Toolkit to provide validation and conformance checking of data that is transformed into CDISC standard domains. SAS Clinical Standards Toolkit includes several types of CDISC controlled terminology, such as ADaM, CDASH, and SDTM. Each controlled terminology type may include several versions, which are identified by their release date. The CDISC controlled terminology is used to verify that the values of CDISC domain columns subject to controlled terminology are valid. It is also used when building the define.xml file to populate the Controlled Terms or Formats column in the domain definition section and to populate the related codelists in the Controlled Terminology (Code Lists) Section.

Another type of controlled terminology is data dictionaries, such as the Medical Dictionary for Regulatory Activities (MedDRA) and the World Health Organization Drug Dictionary (WHO Drug Dictionary). These highly specific and standardized terminologies govern terms related to medical events and medical substances. SAS Clinical Standards Toolkit does not include any data dictionaries because they must be licensed from the governing organizations, but your organization can make use of its existing licenses of these dictionaries and use them to perform validation of the values of variables that should be coded according to the dictionaries.

IMPORT CONTROLLED TERMINOLOGY

On the Clinical Administration tab, right click on the Data Standards item and select Import. This opens the Import Wizard, which allows you to import data standards and controlled terminology from SAS Clinical Standards Toolkit. The second screen allows you to select CDISC-TERMINOLOGY as the data standard type. The third screen displays all of the available data standard versions from which you may choose. After completing the wizard, the controlled terminology is converted to a new data set named TERMS and will appear in the controlled terminology data set folder you specified. Figure 1 shows some of the screens from the import process for controlled terminology.

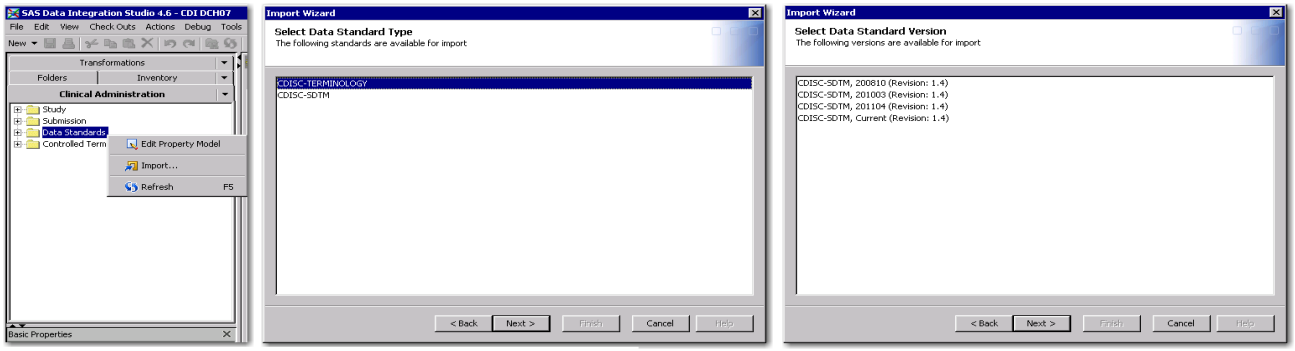


Figure 1: Importing Controlled Terminology

CREATE A CONTROLLED TERMINOLOGY PACKAGE

On the Clinical Administration tab, right click on Controlled Terminology and select New Terminology Package. This opens the New Terminology Package wizard. In the first screen, you define the name of the package, and optionally add a version and description. The second screen allows you to add controlled terminology data sets that you previously imported. You may add multiple controlled terminology data sets, including multiple versions of the same controlled terminology type (for example, two different releases of CDISC SDTM terminology). The controlled terminology files are used by SAS Clinical Data Integration functions in the order in which they are listed in the controlled terminology package. SAS Clinical Data Integration checks the data sets incrementally to search for the needed codelist and uses the codelist from the first data set where it is found. If a subsequent controlled terminology data set has the same codelist with different values, that subsequent codelist will not ever be used. Figure 2 below shows some of the screens from creating a controlled terminology package.

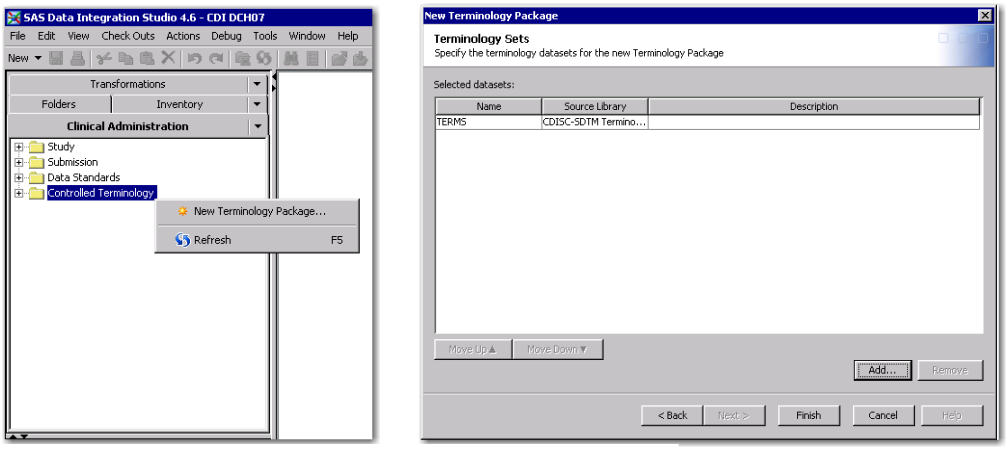


Figure 2: Creating a Controlled Terminology Package

DEFINE A CONTROLLED TERMINOLOGY PACKAGE FOR A STUDY

On the Folders tab, select the folder under which you would like to create a study, right click, and select New → Study. On the final screen of the New Study wizard, you can add a controlled terminology package. A study may only have one controlled terminology package associated with it. You can also manage the controlled terminology package after a study has been created. Find the study on the Clinical Administration tab under Study → Instances, right click on the study, and select Properties. On the Study tab, there is a field at the bottom called Terminology Package. After the study is created, you can add or remove a controlled terminology package from this location. Figure 3 below shows screens from both defining a controlled terminology package within the New Study wizard and adding a controlled terminology package after a study has already been created.

CREATE JOBS TO POPULATE CDISC DOMAINS

Use your company-specific processes for transforming raw data into CDISC domains.

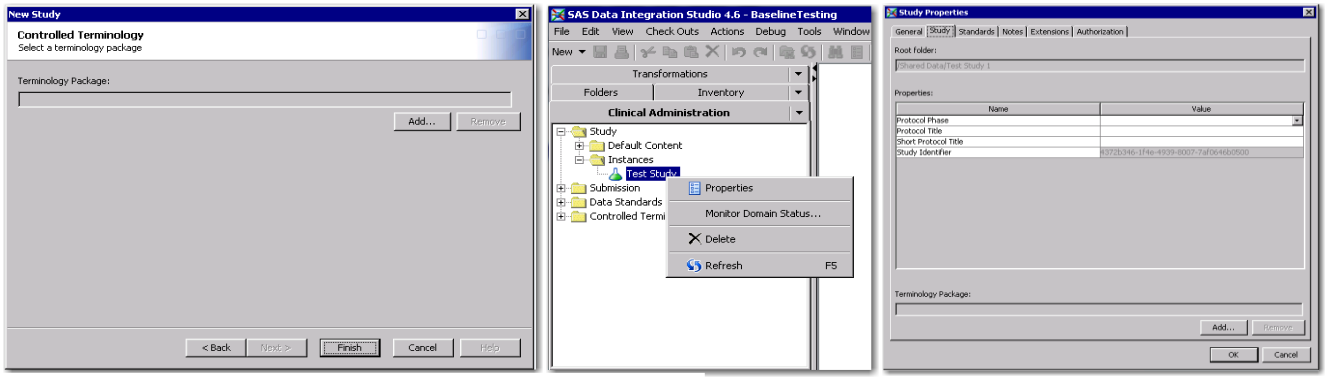


Figure 3: Defining a Controlled Terminology Package for a Study

RUN COMPLIANCE CHECKS ON CDISC DOMAINS

Create a job and use the CDISC-SDTM Compliance transformation to run compliance checks. Within this transformation, you define the data standard to use, the domains to check, the specific compliance checks to run, and the output location for the reports. Several of the compliance checks included for each data standard compare values for columns against the controlled terminology associated with the study and will give an error or warning in the report if a non-null value is not in the controlled terminology codelist. Figure 4 below shows the Checks tab of the CDISC-SDTM Compliance transformation.

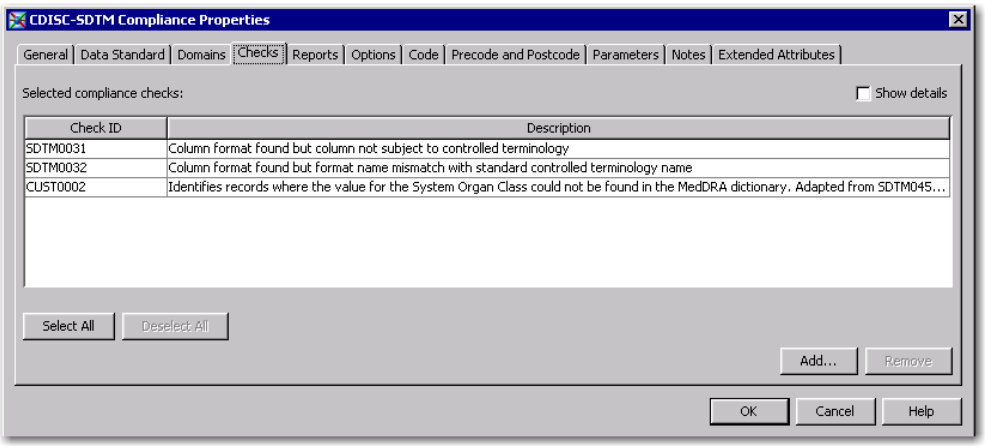
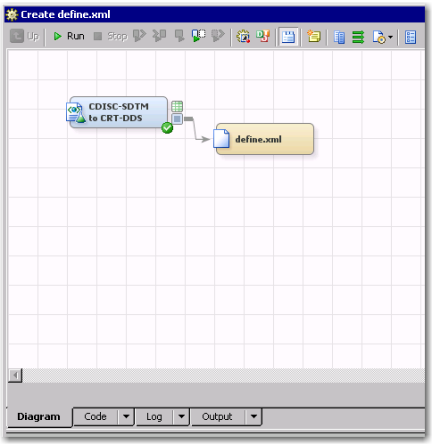


Figure 4: Running Compliance Checks on CDISC Domains

CREATE DEFINE.XML

Create a job and use the CDISC-SDTM to CRT-DDS to create the define.xml file. The Controlled Terms or Formats column, visible to users when viewing the define.xml file using a style sheet, is populated with the related codelist for some of the domain variables. This is controlled by the clinical properties of the column, which can be seen by opening the properties of the domain, then right clicking on a specific domain column and selecting Properties. The Clinical Column tab contains metadata about the column. If a column is governed by controlled terminology, the codelist name will be populated in the XML Codelist field, and in the Term field, enclosed in parenthesis. You can modify the XML Codelist field to either remove, add, or change the controlled terminology codelist associated with the column, which will affect the Controlled Terms or Formats column in define.xml.

Another section of define.xml is the Controlled Terminology (Code Lists) Section. This only includes codelists that are associated with at least one column in at least one domain included in your define.xml file. It pulls the complete codelist out of the first terminology file that contains the codelist in your controlled terminology package associated with your study and lists it in define.xml. The values in the Controlled Terms or Formats column in the domain definitions part of define.xml are hyperlinked to this section, and take you directly to the specific codelist in the Controlled Terminology (Code Lists) Section. Figure 5 below shows screens from a job created to make the define.xml file and from two section of a define.xml file: the domain definition portion and the Controlled Terminology (Code Lists) section.



Subject Characteristics Dataset (SC)

Variable	Label	Type	Controlled Terms or Format	Origin	Role	Comment
STUDYID	Study Identifier	text			Identifier	
DOMAIN	Domain Abbreviation	text			Identifier	
USUBJID	Unique Subject Identifier	text			Identifier	
SCSEQ	Sequence Number	integer			Identifier	
SCTESTCD	Subject Characteristic Short Name	text	SCCD		Topic	

Controlled Terminology (Code Lists) Section

Code Value	Code Text
ACN, Reference Name (ACN)	
DOSE INCREASED	DOSE INCREASED
DOSE NOT CHANGED	DOSE NOT CHANGED
DOSE REDUCED	DOSE REDUCED

Figure 5: Creating define.xml

UNDERLYING STRUCTURE

LOCATION OF METADATA FILES/DATA SETS/ETC

In order to understand how to extend controlled terminology, it is important to understand the underlying structure of the SAS Clinical Standards Toolkit files and how they relate to the processes involving controlled terminology. All of the metadata and files related to data standards are in the Global Standards Library. By default, this is the cstGlobalLibrary folder, installed at the C:\ level when on a Microsoft Windows computer. You can modify this location during installation of SAS Clinical Data Integration, but this paper assumes the default file structure is used. All of the controlled terminology-related files are in the C:\cstGlobalLibrary\standards\cdisc-terminology-1.4 level of the hierarchy. There is a folder for each type of controlled terminology (cdisc-adam, cdisc-cdash, cdisc-sdtm), a control folder, and a programs folder.

Underneath each of the controlled terminology type folders there is a subfolder for each release of the controlled terminology and a folder named current. The current folder is a convenience; when the software is installed it contains a copy of the most recent release of the controlled terminology files included with the software. An organization may replace these files with whatever controlled terminology release it considers “current”, which may or may not be the latest release of the controlled terminology. Underneath each of these release folders is one more folder, called formats, and this is where the actual controlled terminology data set and format catalog are stored. In Figure 6 to the right, the cdisc-sdtm and its current folder are expanded to display this structure.

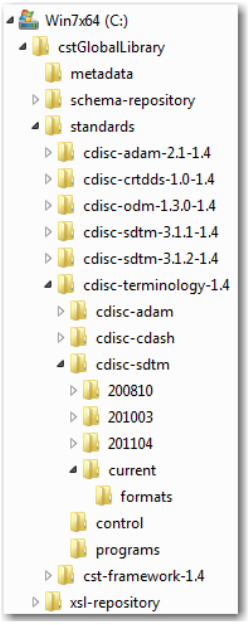


Figure 6: Global Standards Library Folder Structure

CTERMS DATA SET

Within the formats folder is a SAS data set and a SAS format catalog, both named cterms, as well as the .xpt file that thedata set and format catalog were extracted from, during the installation of SAS Clinical Data Integration. The cterms data set contains complete CDISC controlled terminology records in a specific format required by SAS Clinical Standards Toolkit. The cterms format catalog is created from the cterms data set, where the codelist value is the format name and the cdisc_submission_value terms are the format entries.

As mentioned before, SAS Clinical Data Integration leverages the functionalities of SAS Clinical Standards Toolkit. When used independently, SAS Clinical Standards Toolkit uses the cterms SAS format catalogs when it is performing controlled terminology-related tasks. However, SAS Clinical Data Integration uses the cterms files a little differently. SAS Clinical Data Integration provides an Import Wizard that imports controlled terminology. The Import Wizard asks you to define the location of the cterms data set and a location within SAS Clinical Data Integration to place the terminology file. When finished, the Import Wizard copies the cterms data set from the formats folder, saves it as a new data set named terms in the same formats folder, and registers the metadata for the terms data set in SAS Clinical Data Integration. From that point forward, SAS Clinical Data Integration uses the terms data set when performing all terminology-related tasks. It does not pick up any changes to the physical cterms data set unless the Import Wizard runs again. SAS Clinical Data Integration does not use the cterms format catalog during the import because its underlying code uses the terms data set to create a temporary format catalog whenever it is using the controlled terminology for a task. Therefore, this paper only discusses working with the cterms SAS data set, and not the cterms format catalog, for the purposes of extending controlled terminology.

You can see the structure of the imported terms data set in Figure 7 below. When performing controlled terminology compliance checks, if a domain column has its Clinical Column properties set so that it is associated with an XML Codelist, SAS Clinical Data Integration looks within the terms data set codelist field for the specified XML Codelist value. Then the actual values in the CDISC domain column are compared to the cdisc_submission_value possible values from the terms data set. Figure 7 shows the AGESPAN codelist and its possible values in the cdisc_submission_value column.

	codelist	codelist_extensible	Code	Codelist_Code	Codelist_Name	CDISC_Submission_Value	CDISC_Synonym_s_	CDISC_Definition	NCI_Preferred_Term
16	AGESPAN	Yes	C27954	C66780	Age Span	ADOLESCENT (12-17 YEARS)		A juvenile between the onset of puberty and maturity; in the state of development between puberty and maturity. (NCI)	Adolescent
17	AGESPAN	Yes	C49685	C66780	Age Span	ADULT (18-65)		A person from 18 years to 65 years of age. (NCI)	Adult 18-65 Years Old
18	AGESPAN	Yes	C49683	C66780	Age Span	CHILDREN (2-11 YEARS)		A person from 2 years to 11 years of age. (NCI)	Children 2-11 Years Old
19	AGESPAN	Yes	C16268	C66780	Age Span	ELDERLY (> 65)		An age group comprised by people 65 years of age and older. (NCI)	Elderly
20	AGESPAN	Yes	C49641	C66780	Age Span	IN UTERO		The period of time during which the embryo or fetus is present in the uterus of the female. Also describes the location of the embryo or fetus as being in the uterus in contrast to outside the uterus (ex utero). (NCI)	In Utero
21	AGESPAN	Yes	C49643	C66780	Age Span	INFANT AND TODDLER (28 DAYS - 23 MONTHS)		A person from 28 days to 23 months of age. (NCI)	Infant And Toddler
22	AGESPAN	Yes	C16731	C66780	Age Span	NEWBORN (0-27 DAYS)		An infant during the first month after birth. (NCI)	Newborn
23	AGESPAN	Yes	C49642	C66780	Age Span	PRETERM NEWBORN INFANTS		An infant born prior to completion of the normal gestation period. (NCI)	Preterm Newborn Infant

Figure 7: terms data set

STANDARDSUBTYPES DATA SET

The standardsubtypes data set contains metadata about all of the controlled terminology data sets available to SAS Clinical Data Integration. If you choose to add controlled terminology data sets, whether newer releases of CDISC terminology or company-specific terminology, you need to update the standardsubtypes data set. Figure 8 below shows the structure of the standardsubtypes data set. In the root path, each path begins with the macro &_cstGRoot, which is the path to the Global Standards Library folder and populates at run time. You can also enter an absolute path in this field.

	Name of standard	Standard version	Name of standard subtype	Version of standard subtype	Root path for the standard	Is this the default version for the subtype (Y/N)?	The revision of the subtype	Description of the subtype
1	CDISC-TERMINOLOGY	CDISC-ADAM	NCI_THESAURUS	201101	&_cstGRoot./standards/cdisc-ter	Y	1.4	Controlled Terminology released by NCI on 2011-01-07
2	CDISC-TERMINOLOGY	CDISC-ADAM	NCI_THESAURUS	Current	&_cstGRoot./standards/cdisc-ter	N	1.4	Current Controlled Terminology (copy of the latest versio
3	CDISC-TERMINOLOGY	CDISC-CDASH	NCI_THESAURUS	201104	&_cstGRoot./standards/cdisc-ter	Y	1.4	Controlled Terminology released by NCI on 2011-04-08
4	CDISC-TERMINOLOGY	CDISC-CDASH	NCI_THESAURUS	Current	&_cstGRoot./standards/cdisc-ter	N	1.4	Current Controlled Terminology (copy of the latest versio
5	CDISC-TERMINOLOGY	CDISC-SDTM	NCI_THESAURUS	200810	&_cstGRoot./standards/cdisc-ter	N	1.4	Controlled Terminology released by NCI on 2008-09-24
6	CDISC-TERMINOLOGY	CDISC-SDTM	NCI_THESAURUS	201003	&_cstGRoot./standards/cdisc-ter	N	1.4	Controlled Terminology released by NCI on 2010-03-08
7	CDISC-TERMINOLOGY	CDISC-SDTM	NCI_THESAURUS	201104	&_cstGRoot./standards/cdisc-ter	N	1.4	Controlled Terminology released by NCI on 2011-04-08

Figure 8: standardsubtypes data set

BEST PRACTICES

MASTER COPY OF TERMINOLOGY VERSION IN THE GLOBAL STANDARDS LIBRARY

Your organization may want to import new versions of CDISC controlled terminology, import external data dictionaries, or permanently modify an existing CDISC controlled terminology list to extend some codelists. In these cases, a master copy of these controlled terminology files as released by the governing organization should be stored within the cdisc-terminology level folder, following the existing structure. For permanent modifications to a controlled terminology list that you want to be available to all studies, create an additional folder for the modified version within the cdisc-terminology area. In Figure 9 to the right, there are two 201104 folders, one of them named 201104-Extended. The files in the 201104 folder are the original controlled terminology files as released by CDISC, but the files in the 201104-Extended folder include some additional terms in some of the extensible codelists.

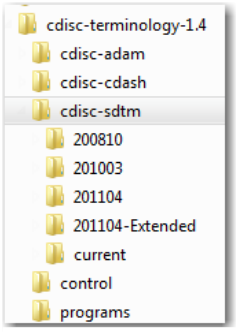


Figure 9: Master Copy of Extended Controlled Terminology Files

If your organization installs an upgrade to SAS Clinical Standards Toolkit or SAS Clinical Data Integration, there is a risk that standard controlled terminology files included with the product can be overwritten, depending on options selected during the upgrade process. For example, the cterms data set and format catalog in the cdisc-sdtm/201104/formats folder may be imported fresh and overwritten if the option to reinstall the current version is selected. In this case, if you had made changes to those files to add extensible terminology, those changes would be lost when the upgrade occurs. If you had saved off your modified version into a separate folder, you would not lose those modifications during the upgrade. Upgrades do not typically wipe out the entire folder structure, so additional files you have saved or folders you have made are not erased.

STUDY-SPECIFIC TERMINOLOGY IN A SEPARATE AREA

Controlled terminology is a standardized list of all possible values for a variable. It is important to remember that only *possible* values should be in your controlled terminology list for any given study. For example, if your study was conducted only within the United States, you should not have over 200 country codes in your COUNTRY codelist for that study. However, the codelist values should not be data-driven. Your study would likely have all three possible values for AESEV (MILD, MODERATE, SEVERE). If you only happen to have mild and severe adverse events occur in your study, that does not mean the MODERATE term should be removed from your codelist.

These rules for controlled terminology mean that you will typically need a controlled terminology list specific for each study (or perhaps each compound, or family of studies) that is based off of one of the CDISC released versions, but modified to include only the possible terms for your study. These study-specific controlled terminology data sets should be saved in a location separate from the master copies. They do not necessarily need to be within the cdisc-terminology level folder, or even within the Global Standards Library structure. They can be maintained within the folder structure for the study. Determine the location that best suits your organization's needs, but follow the structure of the master copy by always having a subfolder named "formats," and save the controlled terminology files within that formats folder. When running the Import Wizard, the underlying code in SAS Clinical Standards Toolkit will append the root path in the standardsubtypes data set to add the formats folder, so if the controlled terminology data set is not stored in a folder named formats, SAS Clinical Data Integration will not be able to find it.

EXTENDING CDISC TERMINOLOGY

IMPORTING NEWER VERSION OF CDISC TERMINOLOGY

CDISC releases new controlled terminology versions more frequently than SAS releases new versions of SAS Clinical Data Integration or SAS Clinical Standards Toolkit. Your organization may want to adopt a new version of CDISC terminology before it is available through an upgrade of your SAS software. CDISC releases controlled terminology in several formats, including Excel and text files. The file structure needs some modifications to get it into a structure that SAS Clinical Standards Toolkit expects.

Figure 10 below shows a snapshot of the ACN codelist from the raw CDISC SDTM terminology Excel file. The first row is highlighted, and indicates the beginning of a new codelist. The Codelist Code column is not populated for this row, but the Codelist Extensible (Yes/No) column is. Notice that for the columns beneath, the Codelist Code column is populated and the value is the same as the Code value in the highlighted row. In the highlighted row, the CDISC Submission Value is ACN, but in the following rows, the values are things like DOSE INCREASED and DOSE NOT CHANGED. The CDISC Submission Value in the highlighted row is actually the Codelist name. The values in the remaining rows are the possible values for ACN.

	A	B	C	D	E	F	G	H
1	Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)	CDISC Definition	NCI Preferred Term
2	C66767		No	Action Taken with Study Treatment	ACN	Action Taken with Study Treatment	Terminology specifying changes to the study treatment as a result of an adverse event.	CDISC SDTM Action Taken with Study Treatment Terminology
3	C49503	C66767		Action Taken with Study Treatment	DOSE INCREASED		An indication that a medication schedule was modified by addition; either by changing the frequency,	Dose Increased
4	C49504	C66767		Action Taken with Study Treatment	DOSE NOT CHANGED		An indication that a medication schedule was maintained. (NCI)	Dose Not Changed
5	C49505	C66767		Action Taken with Study Treatment	DOSE REDUCED		An indication that a medication schedule was modified by subtraction, either by changing the	Dose Reduced
6	C49501	C66767		Action Taken with Study Treatment	DRUG INTERRUPTED		An indication that a medication schedule was modified by temporarily terminating a prescribed	Drug Interrupted
7	C49502	C66767		Action Taken with Study Treatment	DRUG WITHDRAWN		An indication that a medication schedule was modified through termination of a prescribed regimen	Drug Withdrawn
8	C48660	C66767		Action Taken with Study Treatment	NOT APPLICABLE	NA; Not Applicable	Determination of a value is not relevant in the current context. (NCI)	Not Applicable
9	C17998	C66767		Action Taken with Study Treatment	UNKNOWN	U; Unknown	Not known, not observed, not recorded, or refused. (NCI)	Unknown

Figure 10: Raw CDISC SDTM Controlled Terminology Excel File (ACN codelist)

Figure 11 below shows a snapshot of the same ACN codelist in the format that SAS Clinical Standards Toolkit needs. The differences are the addition of the codelist column and the population of the codelist_extensible column. The codelist column takes the Codelist name, which in this case is ACN, and explicitly defines it in each row of the data set instead of having it available only in a header row like in the raw file. The value for codelist Extensible (Yes/No) from the header row in the raw table is populated for every record in the codelist. Finally, the header row is removed, since the pertinent values from it have been filtered down to the related records. This new structure for the data set contains complete information about the associated codelist in every row.

	codelist	codelist_extensible	Code	Codelist_Code	Codelist_Name	CDISC_Submission_Value	CDISC_Synonym_s_	CDISC_Definition	NCI_Preferred_Term
1	ACN	No	C49503	C66767	Action Taken with Study Treatment	DOSE INCREASED		An indication that a medication schedule was modified by addition; either by changing the frequency, strength or amount. (NCI)	Dose Increased
2	ACN	No	C49504	C66767	Action Taken with Study Treatment	DOSE NOT CHANGED		An indication that a medication schedule was maintained. (NCI)	Dose Not Changed
3	ACN	No	C49505	C66767	Action Taken with Study Treatment	DOSE REDUCED		An indication that a medication schedule was modified by subtraction, either by changing the frequency, strength or amount. (NCI)	Dose Reduced
4	ACN	No	C49501	C66767	Action Taken with Study Treatment	DRUG INTERRUPTED		An indication that a medication schedule was modified by temporarily terminating a prescribed regimen of medication. (NCI)	Drug Interrupted
5	ACN	No	C49502	C66767	Action Taken with Study Treatment	DRUG WITHDRAWN		An indication that a medication schedule was modified through termination of a prescribed regimen of medication. (NCI)	Drug Withdrawn
6	ACN	No	C48660	C66767	Action Taken with Study Treatment	NOT APPLICABLE	NA; Not Applicable	Determination of a value is not relevant in the current context. (NCI)	Not Applicable
7	ACN	No	C17998	C66767	Action Taken with Study Treatment	UNKNOWN	U; Unknown	Not known, not observed, not recorded, or refused. (NCI)	Unknown

Figure 11: CDISC SDTM Controlled Terminology Transformed into cterms SAS Data Set

An important note about adding new versions is that the length of cdisc_submission_value is limited by the underlying SAS Clinical Data Integration code that converts the terms data set to a format catalog. In SAS Clinical Data Integration 2.3, this field is limited to 80 characters (in version 2.4 it has been increased to 200 characters). In the December 2012 CDISC controlled terminology, there are some controlled terms that are longer than 80 characters. SAS Clinical Data Integration truncates those values at 80 characters, which results in there being non-unique values within some of the codelists. This causes an error because each possible value in a format must be unique when using PROC FORMAT.This issue needs to be addressed in order to use the December 2012 or later controlled terminology in SAS Clinical Data Integration 2.3.

This issue only affects 12 terms within three codelists (CVPRCIND, DISCHDX, LOC) in the December 2012 version. One option is to remove the terms whose cdisc_submission_value terms are too long. Another option is to save the original terms as a new variable, such as cdisc_submission_value_original, and truncate the values in the cdisc_submission_value column that are greater than 80 characters. In this case, a best practice is to truncate those values to 78 characters and add a unique two-digit identifier to the end of the text to ensure the values remain unique. Either of these options causes compliance checks using that codelist to generate an error if any of the too-long terms are in your domain. Any users working with this codelist need to be aware of this limitation so that they do not mistakenly think the error for these terms is valid. You can also modify the error message for the compliance check to alert users of the codelist values that may generate a false error and prompt them to check the term manually. A sample program that converts the raw CDISC SDTM text file to a SAS Clinical Standards Toolkit-ready data set using the truncation method for too-long terms can be found in the Sample Code section at the end of the paper (Sample 1).

Now that there is a newer version of controlled terminology in your hierarchy, your organization may want this to be considered the current version of the controlled terminology. If so, copy the new cterms data set to the current folder within the cdisc-sdtm folder, replacing the existing cterms data set.

Next, update the standardsubtypes data set with the following modifications:

- Add a new record for the new controlled terminology version. If this version is considered the new default version, be sure to set that value to Y.
- Modify the existing record for the “Current” version to update the description of the subtype to the new release date.
- Modify the record for the controlled terminology version that had previously been the default version to set the default flag to N. You should only have one default version per standard subtype.

Note that when importing controlled terminology into SAS Clinical Data Integration, the wizard combines the standardversion and standardsubtypeversion values when displaying the description of the available controlled terminology. You can add more than just the date to the standardsubtypeversion value to make it more descriptive, if necessary. After these steps, the new controlled terminology version will be available to SAS Clinical Data Integration. You can find a sample program that updates the standardsubtypes data set in the Sample Code section at the end of the paper (Sample 2).

ADDING TERMS TO EXTENSIBLE CDISC CODELISTS

Some codelists in CDISC-supplied controlled terminology lists are extensible, meaning you can add additional terms to the list. You may simply insert additional rows into the cterms data set using basic data step programming. As mentioned in the section above, you may want to add terms to extensible codelists in a master copy, and then use that version as a starting point for study-specific controlled terminology. Be sure to add terms to the cterms data set and update the standardsubtrypes data set before running the Import Wizard in SAS Clinical Data Integration to register the extended controlled terminology.

USING EXTERNAL DATA DICTIONARIES (MEDDRA, WHO DRUG DICTIONARY)

Data dictionary data sets do not need to be in the same format as the CDISC controlled terminology data sets because they are not used by SAS Clinical Data Integration in the same way. Data dictionaries' terminology lists are not included in define.xml, and you need to customize compliance checks using data dictionaries. These compliance checks typically just compare the values of a column from a CDISC domain to the values of a variable in the controlled terminology data set, and alert you if a value in the domain is not found in the controlled terminology data set. To use external data dictionaries, just convert the raw files received from the governing organization into SAS data sets, store them similarly to CDISC controlled terminology files within the cdisc-terminology level folder, and register them in SAS Clinical Data Integration as you would any regular SAS data set. There is no need to use a formats subfolder for external data dictionary files, and the standardsubtypes data set does not need to be modified to include these files.

CREATING STUDY-SPECIFIC CONTROLLED TERMINOLOGY

Use a master copy of the appropriate version of CDISC controlled terminology as a starting point for study-specific controlled terminology, and note that this may be a version to which you have added extended terms. You can create a study-specific controlled terminology file using simple data step programming to remove unneeded terms and store the new file in the study-specific folder. Update the standardsubtypes data set, and then run the Import Wizard in SAS Clinical Data Integration to register the study-specific controlled terminology.

CREATING CUSTOM COMPLIANCE CHECKS FOR DATA DICTIONARIES

There is an existing compliance check for CDISC 3.1.2, SDTM0451, that is set up to check the values of the AEDECOD column in the AE domain to the Preferred Term values in the MedDRA terminology file. This compliance check assumes that the MedDRA terminology file is named pt and that the variable containing the Preferred Terms is named pt_name, which may not be true for your file. This compliance check is easily modified for its intended purpose as well as for using any data dictionary files and terms.

Start by selecting the compliance check and selecting Customize. Give the Check ID a new name, and modify the description as necessary (Figure 12).

The next screen specifies the domains that the check applies to, and is set to AE. To modify this, select the Direct Edit button and modify the text in the Domain Specification field. To specify more than one domain, put the + sign between the domain names (Figure 13).

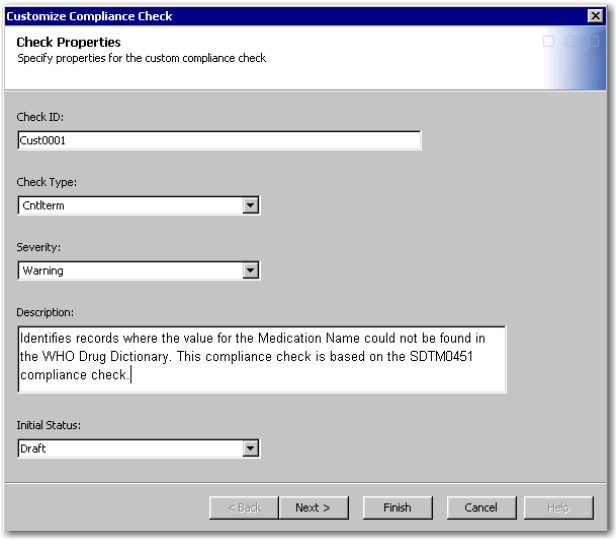


Figure 12

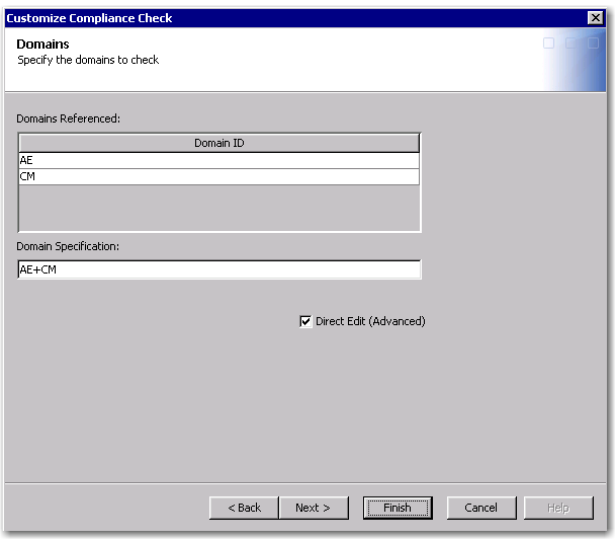


Figure 13

The next screen specifies the columns that the check applies to, and is set to AEDECOD. To modify this, select the Direct Edit button and modify the text in the Column Specification field. To specify multiple columns, enclose each one in [] brackets. When the check applies to multiple domains, you can put ** for the first two characters of the column name so that it will fill in the domain abbreviation. **DECOD would apply to AEDECOD and CMDECOD, for example. Figure 14 below shows using both of these features; note that the "example" column is included for demonstration of the [] brackets and not as a valid column.

The next screen contains the code that performs the compliance check. The code contains three instances of pt_name; if your controlled terminology data set variable has a different name, select the Direct Edit button and replace pt_name with the correct variable name (Figure 15).

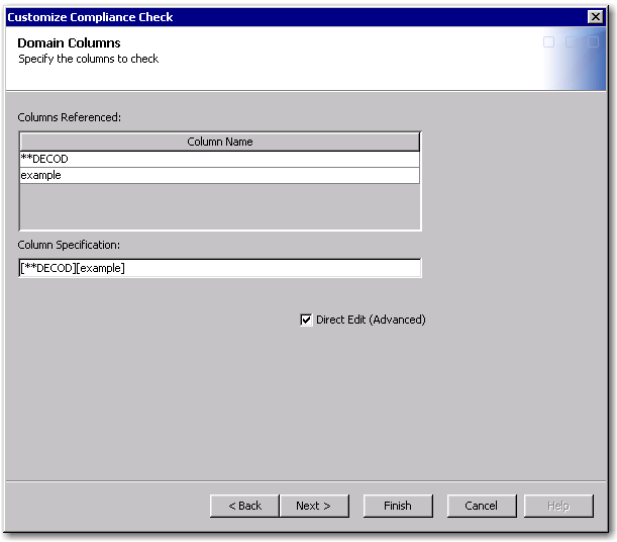


Figure 14

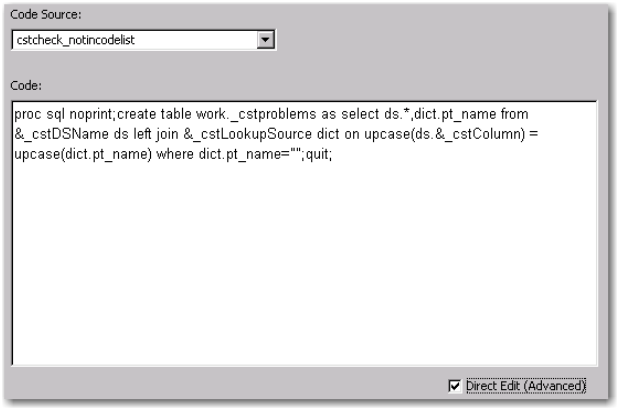


Figure 15

The next screen specifies which controlled terminology data set should be used for the compliance check, and is set to pt. If your data set has a different name, select the Direct Edit button and type in the correct data set name in the Lookup Source field (Figure 16).

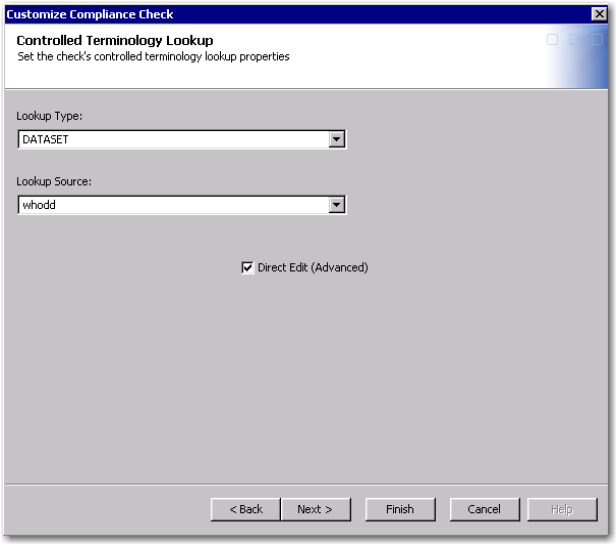


Figure 16

The final screen specifies the error message to be written to the report whenever the compliance check fails. To modify this, select the Direct Edit button and type the message you prefer. After selecting Finish, the new custom compliance check appears in the list of available checks, but it is in draft mode. To activate it, select the compliance check and select Make Active. At this point the compliance check is available to the CDISC-SDTM Compliance transformation. Custom compliance checks for other purposes can be easily created by finding an existing compliance check that is similar to what you want and modifying it to suit your needs.

MANAGING CLINICAL PROPERTIES TO INFLUENCE THE APPEARANCE OF CODELISTS IN DEFINE.XML

In the domain definition portion of the define.xml file, the value of the Controlled Terms or Formats column is determined by the properties of the individual domain columns. To view these properties, open the properties of a domain and go to the Columns tab. Select a column, right click and select Properties. Then select the Clinical Column tab. The figure below shows the properties for the AESEV column. The XML Codelist value is AESEV, and the Term value is (AESEV). The XML Codelist value is what controls the column's association with a codelist. The Term value is the same as the XML Codelist value, enclosed in parenthesis, but Term is not used by SAS Clinical Data Integration. If you no longer wanted this column associated with a controlled terminology list, you could delete the XML Codelist value and the Controlled Terms or Formats column in define.xml would be empty for that column. Even though the Term value is not used, delete its value also to

maintain consistency and avoid confusion. Figure 17 below displays the default Clinical Column properties for the AESEV column.

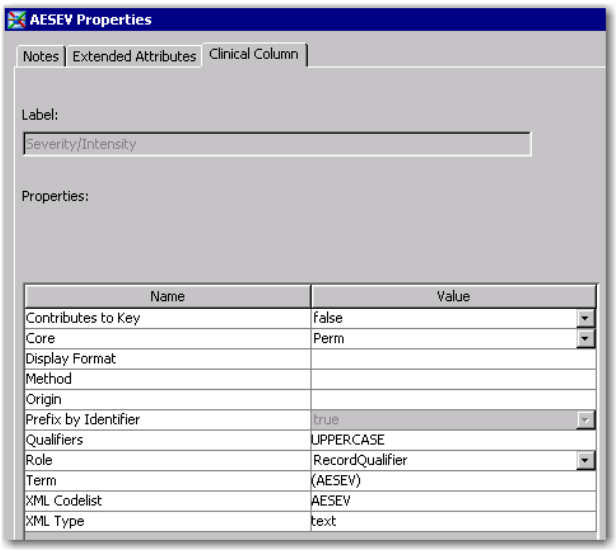


Figure 17

You may need to add a controlled terminology codelist to a domain column. In an earlier example, we imported a new version of CDISC controlled terminology from December 2012. This version includes a new codelist for the EPOCH column. However, SAS Clinical Data Integration 2.3 does not know that EPOCH should have a codelist associated with it, so the Term and XML Codelist values for this column are blank. To associate the EPOCH codelist with this column, go to its Clinical Column properties tab and enter EPOCH for XML Codelist. For consistency, you may also enter (EPOCH) for Term. The define.xml file now includes this in the Controlled Forms or Formats column for the EPOCH column. Note that you would need to take this step within each domain using the EPOCH column; its column properties do not carry over from one domain to another.

CONCLUSION

Controlled terminology is extendable in many ways: importing new versions, using external data dictionaries, adding terms to extensible codelists, and creating study-specific versions of existing standards. This paper presents all of the information and steps necessary to accomplish any of these tasks using SAS Clinical Data Integration 2.3. Remember that SAS Clinical Data Integration can maintain an unlimited number of controlled terminology packages, and you may find that your organization uses a different controlled terminology package for each study. Using extended controlled terminology helps increase data quality in your submission data sets and makes your define.xml file more complete and accurate.

REFERENCES

SAS Institute Inc. 2012. *SAS Clinical Data Integration 2.3: User's Guide, Second Edition*. Cary, NC: SAS Institute, Inc. http://support.sas.com/documentation/cdl/en/clindiug/65391/PDF/default/clindiug.pdf

SAS Institute Inc. 2011. *SAS Clinical Standards Toolkit 1.4: User's Guide*. Cary, NC: SAS Institute, Inc. http://support.sas.com/documentation/cdl/en/clinstdtktug/64439/PDF/default/clinstdtktug.pdf

SAS Institute Inc. 2011. *SAS Clinical Standards Toolkit 1.4: Getting Started*. Cary, NC: SAS Institute, Inc. http://support.sas.com/documentation/onlinedoc/clinical/14/clinstdtktgs.pdf

ACKNOWLEDGEMENTS

Thank you to Julie Maddox and Lex Jansen for tremendous support during my research that led to this paper's publication. Thank you to Angela Lightfoot, Gene Lightfoot, and Donna Dutton for supporting my efforts and providing valuable feedback. Finally, thank you to David Ramage and Bernd Doetzkies for the opportunity to strengthen my experience.

SAMPLE CODE

SAMPLE 1 – CONVERT RAW CDISC SDTM CONTROLLED TERMINOLOGY FILE TO CTERMS SAS DATA SET

```
libname sdtmdec "C:\cstGlobalLibrary\standards\cdisc-terminology-1.4\cdisc-
sdtm\201212\formats";

* Set the GUESSINGROWS to about as many records as the controlled terminology file contains ;
PROC IMPORT OUT= WORK.sdtmterms
   DATAFILE= "C:\cstGlobalLibrary\standards\cdisc-terminology-1.4\cdisc-sdtm\201212\SDTM
Terminology_2012_12_21.txt"
   DBMS=dlm REPLACE;
   DELIMITER='09'X;
   GETNAMES=YES;
   GUESSINGROWS=7000;
RUN;

* Creates a data set of what will be the codelist and codelist_extensible values. ;
data codelists;
   set sdtmterms(keep=code codelist_extensible__yes_no_ cdisc_submission_value);
   where codelist_extensible__yes_no_ ^= '';
run;

* Creates a data set of the remaining detailed codelist records.;
data otherterms;
   set sdtmterms;
   drop codelist_extensible__yes_no_;
   where codelist_extensible__yes_no_ = '';
run;

* Merges the codelists and otherterms data sets where the code value for the codelist values;
* matches the codelist_code value for the detailed records. Also sets the codelist value.;
proc sql;
   create table terms as
   select codelists.cdisc_submission_value as codelist,
codelists.codelist_extensible__yes_no_ as codelist_extensible, otherterms.*,
otherterms.cdisc_submission_value as original_cdisc_submission_value
   from codelists, otherterms
   where codelists.code=otherterms.codelist_code;
quit;

proc sort data=terms;
   by codelist cdisc_submission_value;
run;

* CDI 2.3 limits the length of cdisc_submission_value to 80 characters. This causes an error;
* when CDI creates a format catalog from terms data set because of non-unique values when;
* variable is truncated. This finds values>80 characters, retains first 78, adds numeric ;
* identifier. NOTE these values could cause a false error during compliance checking.;
data limitto80 (keep=codelist cdisc_submission_value original_cdisc_submission_value);
   set terms;
   by codelist;
   where length(original_cdisc_submission_value)>80;
   if first.codelist then count=0;
   count+1;
   cdisc_submission_value=cat(substr(original_cdisc_submission_value,1,78),count);
run;

* Merge the modified values back in with the terms data set;
data finalterms;
   merge terms limitto80;
   by codelist original_cdisc_submission_value;
run;

* Creates a physical copy of the terms data set;
```

```
data sdtmdec.cterms;
   set finalterms;
   if substr(codelist,length(codelist)) in ("0","1","2","3","4","5","6","7","8","9") then
codelist=cats(codelist,"F");
run;
```

SAMPLE 2 – UPDATE STANDARDSUBTYPES DATA SET

```
libname control " C:\cstGlobalLibrary\standards\cdisc-terminology-1.4\control";

* ***IMPORTANT*** After copying the current version of the standardsubtypes data set ONCE, ;
* this code is commented out.;
data control.standardsubtypesbackup;
   set control.standardsubtypes;
run;

* Create new record for 201212 release.;
data newsubtyperecords;
   standard='CDISC-TERMINOLOGY';
   standardversion='CDISC-SDTM';
   standardsubtype='NCI_THESAURUS';
   standardsubtypeversion='201212 Master Copy';
   rootpath='&_cstGRoot./standards/cdisc-terminology-1.4/cdisc-sdtm/201212';
   isstandarddefault='Y';
   productrevision='1.4';
   description='Controlled Terminology released by NCI on 2012-12-21';
   output;
run;

* Modify description for current records, change previous default version to not default.;
data modifyoldrecords;
   set control.standardsubtypes;
   if standardsubtypeversion="Current" and standardversion="CDISC-SDTM" then
description='Current Controlled Terminology (copy of the latest version: 2012-12-21)';;
   if standardsubtypeversion="201104" and standardversion="CDISC-SDTM" then
isstandarddefault="N";
run;

* Combine the updated records with the new records. ;
data updatedstandardsubtypes;
   set modifyoldrecords newsubtyperecords;
run;

* Sort the final data set. ;
proc sort data=updatedstandardsubtypes;
   by standardversion standardsubtypeversion;
run;

* Replace physical standardsubtypes data set with new version of the data set.;
data control.standardsubtypes;
   set updatedstandardsubtypes;
run;
```

CONTACT INFORMATION

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

Melissa R. Martinez
SAS Institute Inc.
720 SAS Campus Dr.
Cary, NC 27513, USA
Work Phone: +1 (919) 531-9277
E-mail: Melissa.Martinez@sas.com