PhUSE EU Connect 2018

Paper PP15

Stop Copying CDISC Standards

Craig Parry, SyneQuaNon, Diss, England

Abstract

We repeatedly see repositories which require a large amount of front loading, a lot of duplicating of the Clinical Data Interchange Standards Consortium (CDISC) documentation into a custom metadata repository (MDR). These MDRs are becoming a hot topic and they all claim to be these clever automation tools to do all your mapping for you. Which is rarely true. The paper will detail how basic software like EXCEL and the standards already provided by the CDISC and the National Cancer Institute (NCI) teams can be used to create a better and smarter starting point for your mapping specifications. It will also illustrate the importance of creating stronger specifications from the beginning by producing a Define.XML with a click of a button in seconds. It will also emphasize that users should stop duplicating the CDISC standards into these MDRs. My goal is to promote the use of basic and readily available software, like EXCEL, in a smarter way.

Introduction

The way source data is handled and structured over time has dramatically changed, rules and regulations have massively affected how clinical and non-clinical data is reported. Groups like Clinical Data Interchange Standards Consortium (CDISC) and the National Cancer Institute (NCI) have been key drivers of clinical and non-clinical standards. They have both been working endlessly to draft standards and drive the industry to become more standardized.

For companies to convert their source data into this standardized format and for reviewers to understand the step between source data and standardized data, there needs to be a multitude of documents explaining the mapping process from source to standardized. These documents can consist of a Mapping Specification (typically a Microsoft EXCEL document), the Annotated Case Report Form (CRF) and a Define.xml output. The Mapping Specifications will detail how the source data will be mapped into the standard format. The Annotated Case Report Form (CRF) illustrates a more visual link between the source data and the standardized data. Then, there is the Define.XML output, the main reference document for a reviewer, this document contains links to the standardized outputs, as well as links to multiple associated documents like the annotated CRF and Study Data Reviewer's Guide (SDRG).

The creation of these documents can take quite some time, even if there is a template or internal process to create these standardized data sets, there is always room for greater efficiency, i.e. there is always a way to make the process faster, smoother and or more dynamic. Imagine if a process could semi-automate large portions of these manual tasks, such as the creation of the Mapping Specifications or annotation of the CRF or even the production of the Define.XML output itself? Automating these manual tasks in some way would not only increase efficiency it would also create consistency between studies. It would also illustrate a clear link between these documents, i.e. the mapping specifications, CRF and Define.XML. However, for this automation to be truly efficient, it would need to be future proof and handle changes to the industry standards. For example, the NCI team release new terminology documents every three months and CDISC are due to release a new Implementation Guide soon. These new documents introduce new domains and updated rules, which companies need to adopt.

This paper explains how a simple Microsoft EXCEL based tool, built with ever changing standards as the focal point can work and in-turn encourage innovation and 'out of the box' thinking, but simultaneously reduce the duplication of metadata and the need to copy standards into static metadata repositories. The tool or utility uses Visual Basic for Applications (VBA) within Microsoft EXCEL, however the idea could work with any programming language. The approach to use EXCEL was taken for the universal accessibility the industry will have to Microsoft EXCEL. This utility would form the basis for the mapping specifications, annotated CRF and Define.XML output.

Utilizing the Industry Standards

Most companies have likely heard of Study Data Tabulation Model (SDTM) and or Standards for Exchange of Nonclinical Data (SEND) standards. The SDTM and SEND standards are produced by the CDISC teams, they detail how clinical and non-clinical source data should be standardized. These standards have also become an FDA requirement for new studies post December 2016. CDISC produces several documents to guide the industry on how to make source data CDISC compliant, be it SDTM or SEND compliant. The main standards document CDISC

produce for any standard is known as the Implementation Guide (IG), the IG is a long document, which is well detailed and updated frequently. The IG documents the rules, guidelines and examples around how to handle and convert the source data into CDISC compliant datasets, such as SDTM datasets. The IG is the document a programmer will use throughout the study. It will be the starting point for all mapping specifications and without this document data submission would be a free for all. Furthermore, to compliment the CDISC IG are the terminology documents known as Controlled Terminology (CT) lists, these lists are produced and maintain by the NCI team. The CT is a document defined on the NCI website as "the set of CDISC-developed or CDISC-adopted standard expressions (values) used with data items within CDISC-defined datasets.". These documents are vital in the ensuring CDISC compliance, as they detail allowed values for certain data fields, such as allowable race values.

RACE (Race)

NCI Code: C74457, Codelist extensible: No

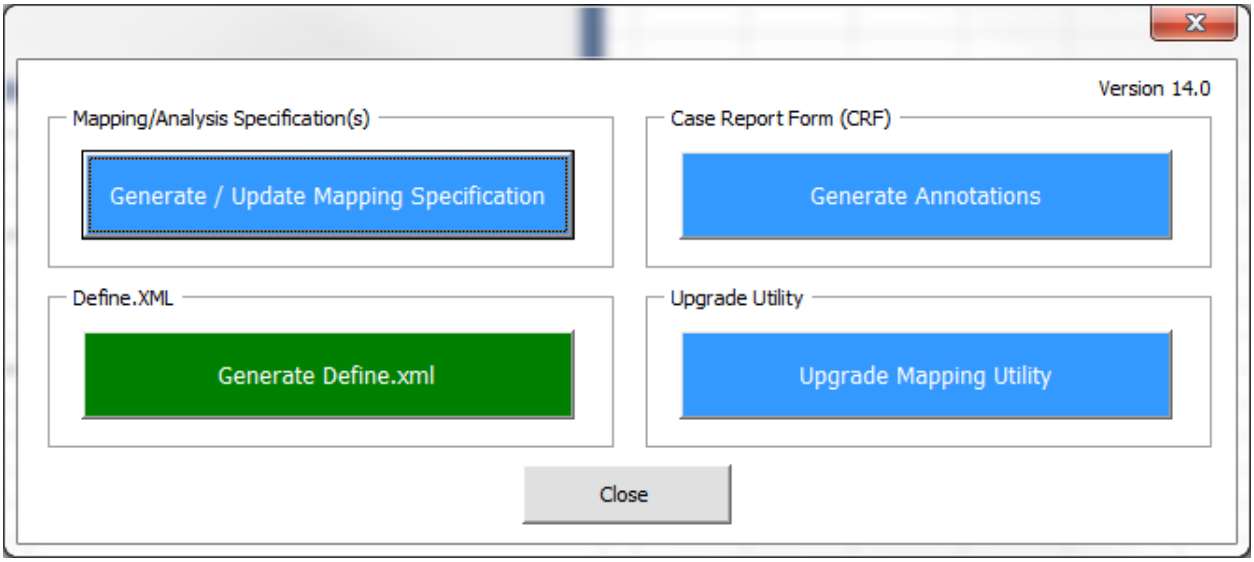
C74457	RACE			
NCI Code	**CDISC Submission Value**	**CDISC Synonym**	**CDISC Definition**	**NCI Preferred Term**
C41259	AMERICAN INDIAN OR ALASKA NATIVE		A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment. (FDA)	American Indian or Alaska Native
C41260	ASIAN		A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam. (FDA)	Asian
C16352	BLACK OR AFRICAN AMERICAN		A person having origins in any of the black racial groups of Africa. Terms such as "Haitian" or "Negro" can be used in addition to "Black or African American." (FDA)	Black or African American
C41219	NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER		Denotes a person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands. The term covers particularly people who identify themselves as part-Hawaiian, Native Hawaiian, Guamanian or Chamorro, Carolinian, Samoan, Chuu. (FDA)	Native Hawaiian or Other Pacific Islander
C41261	WHITE		Denotes a person with European, Middle Eastern, or North African ancestral origin who identifies, or is identified, as White. (FDA)	White

Both the IG and CT documents are ever evolving and constantly being updated and improved. As an example of this, in 2016 the NCI released four versions of the CT, with each revision adding or updating code lists and values. Since the IG and CT documents are the starting point for all mapping specifications, these constant updates make it quite difficult to maintain an up-to-date "specification template". This is because with each release of either the IG or CT document any static template would become out of date. That said companies need to be pragmatic and avoid developing these "static templates" (also known as a Metadata Repository (MDR)), simply because the manual effort used to maintain the template after each new update can be cumbersome, i.e. time wasted.

The Utility

Being dynamic to adapt to ever changing standards was the original starting point for the utility, a tool which needed to produce a template with change in mind. Therefore, why not build the template directly from the IG and CT documents? Now, both the IG and CT documents are available primarily in PDF format, which realistically is not very useful to a machine, i.e. not readily machine-readable. Trying to make a system, like SAS or EXCEL interact with a document in a PDF format is a nonstarter. Thankfully CDISC and the NCI teams are aware of this "flaw" with PDF and provide both documents in an array of different formats, such as EXCEL, XML and ODM from the CDISC **S**hare **H**ealth **a**nd **R**esearch **E**lectronic (SHARE) library and NCI website. Unlike PDF these formats are machine-readable, which any company can freely utilize. The utility in this paper uses the EXCEL format, since as previously mentioned, EXCEL was chosen due to its abundance within the industry.

The utility, in basic terms, is an EXCEL workbook with one worksheet and with some hidden VBA code and nothing else. Most people may never use or understand VBA; therefore, a user interface was created to make the tool more user friendly. The utility can be broken into three major parts, **Mapping Specification Creation**, **Annotated Case Report Form (CRF) Creation** and **Define.XML Output Creation**. Below is the user interface created which displays the three main features of the tool as well as one additional process which will not be detailed in this paper.



The Mapping Specifications

To create SDTM and or SEND data sets a programmer needs a set of instructions. These instructions allow the programmer to understand how the raw data maps to the SDTM or SEND data, typical these instructions are held within an EXCEL file called a Mapping Specification. The content of these mapping specifications primarily consists of the domain and variable information which comes directly from the IG and CT documents. Therefore, the CT and IG documents are the best starting point to begin drafting the mapping specifications. That said, the utility itself will build the mapping specifications directly from the CDISC and NCI documents, these documents are not altered in anyway and are merely downloaded off the respectable websites. Of course, there are pros and cons to using the CT and IG documents as a starting point.

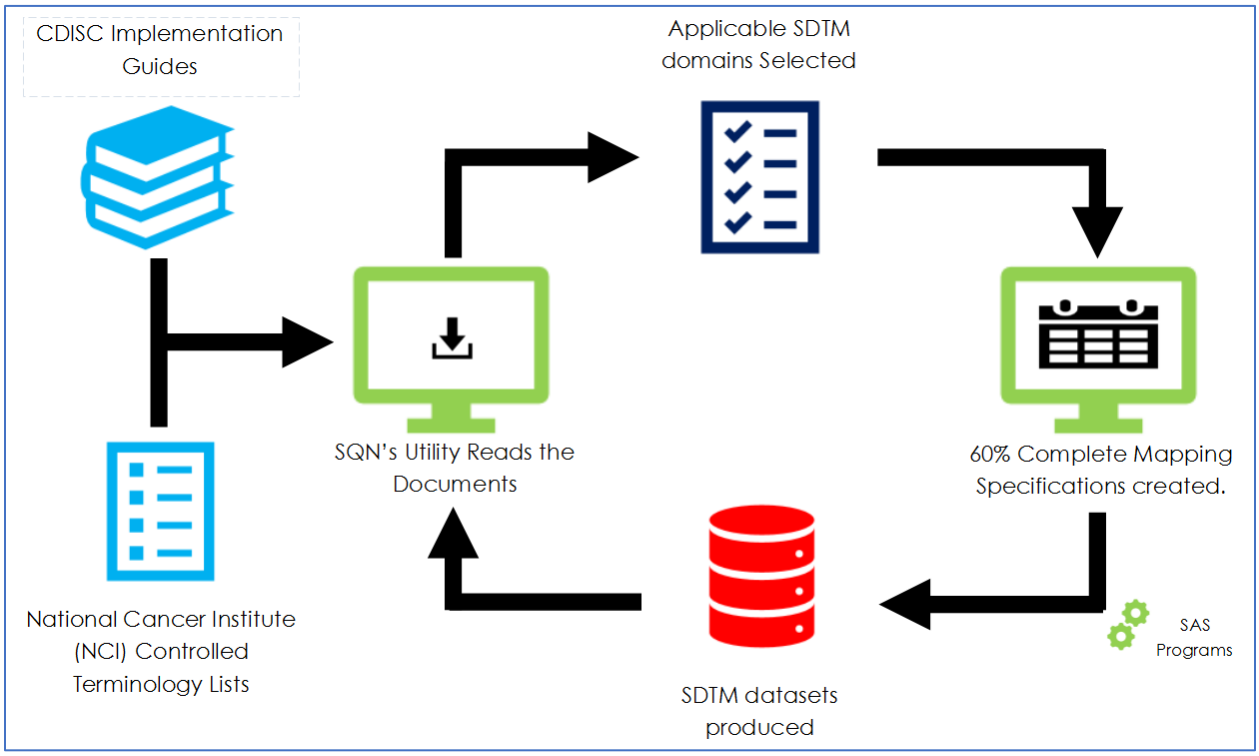
Pros	Cons
Lowers the risk of human error.	Relies on the CDISC and NCI teams not changing their document structures.
Compatible with any current version of the CDISC IG or NCI Controlled Terminology documents.	Part of the process remains hidden to users leaving them with less control.
Huge efficiency boost in initial draft.	
Future proof with new versions of the CDISC IG or NCI Controlled Terminology documents.	

The process is very simple and consists of two prerequisites.

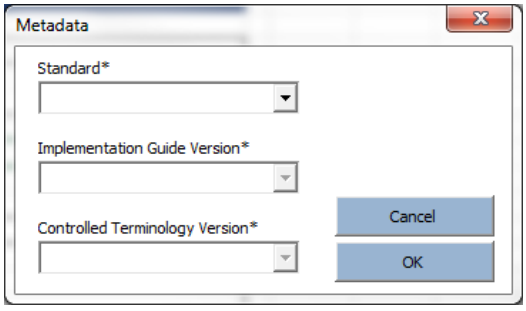
1. Microsoft EXCEL is available.
2. Member of CDISC.

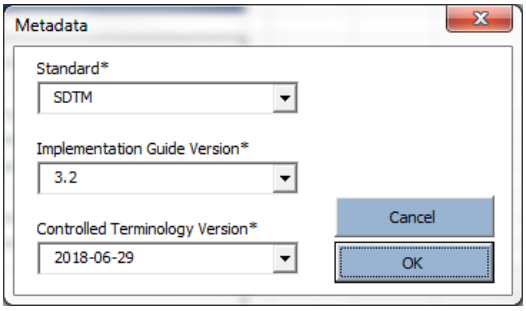
The reason for requiring membership to CDISC is because all members get access to CDISC SHARE, which is a metadata repository which contains machine readable formats of the CDISC documentation, e.g. EXCEL version of the IG.

Below is a high-level image of the utility and how the process works to create the mapping specifications.

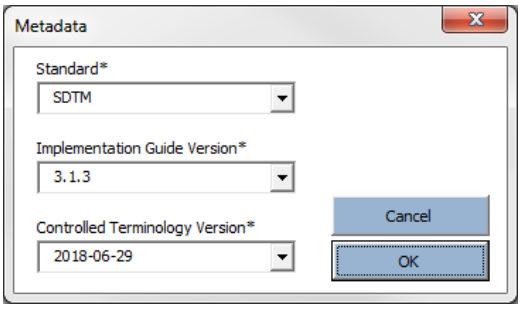


To generate a basic mapping specification template takes no more than 2 minutes by using the utility and only consists of the user clicking a few buttons. The mapping specification starts off as a blank EXCEL workbook with one worksheet. This worksheet merely houses a button the user can click to open the user interface pictured earlier. Upon clicking the "*Generate / Update Mapping Specification*" button the user will be prompted with a window asking for the version of the CDISC IG and NCI Controlled Terminology document they wish to use, and they can select any version from the dropdown list.

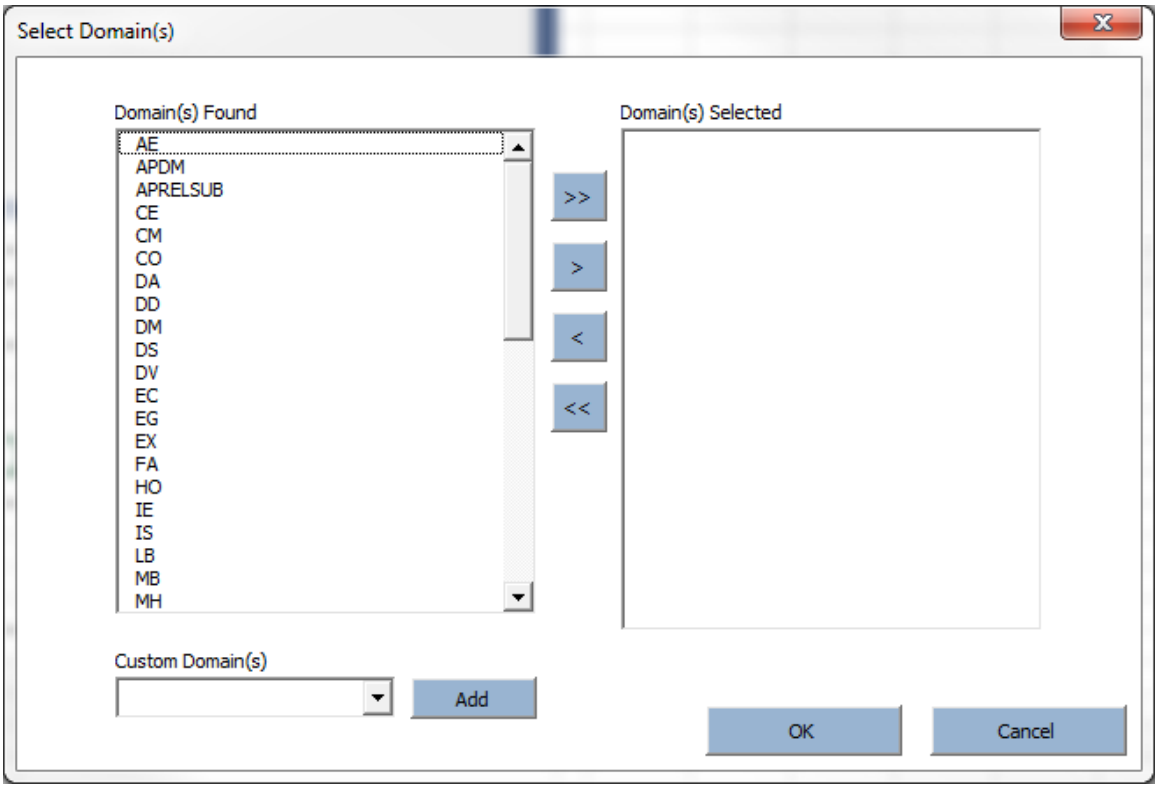




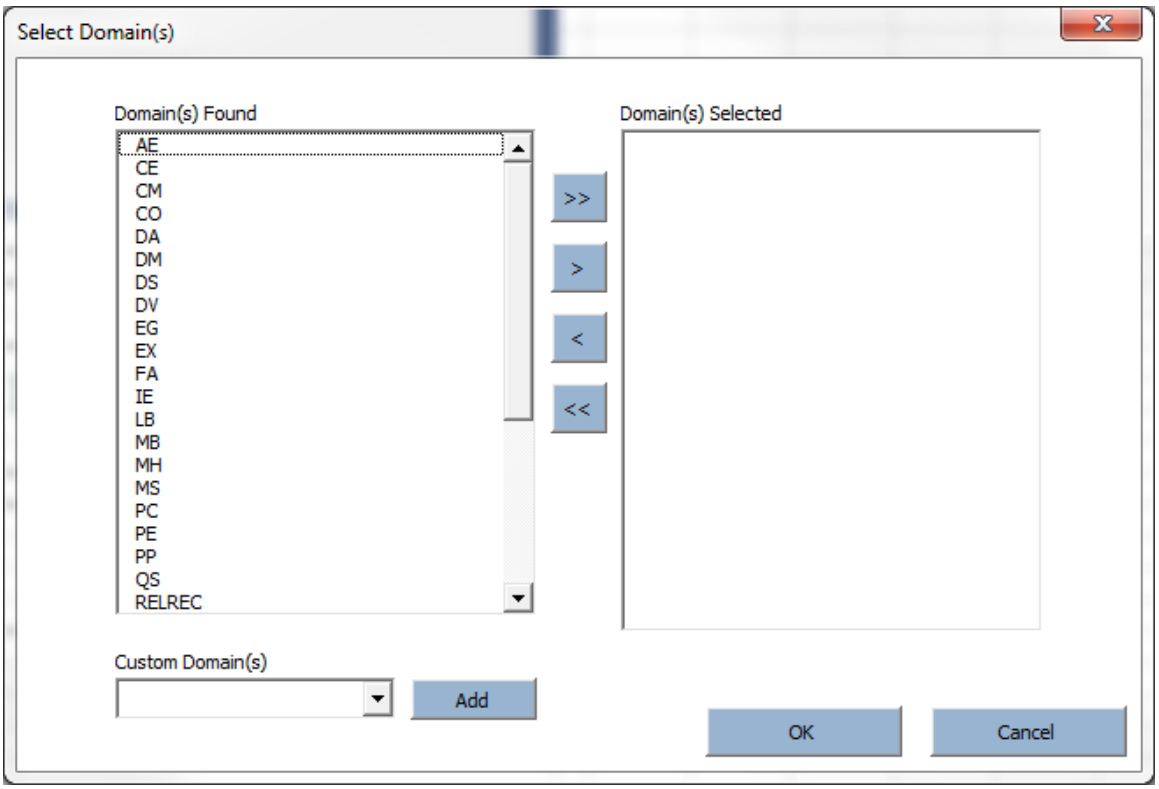
The versions of these documents are constantly updated, and the utility will handle this. For example, the user may have an old study which is part of a suit of legacy studies already submitted in SDTM format but using an older version of the SDTM IG. For consistency they may wish to generate the mapping specification template from an older version, e.g. version 3.1.3. Here are two examples.



Upon clicking OK the utility will begin to read the CDISC and NCI documents, searching for all domains available. The images below illustrate the differences in domains found between two versions of the IG document, this is because new domains are added with each revision of the IG document.

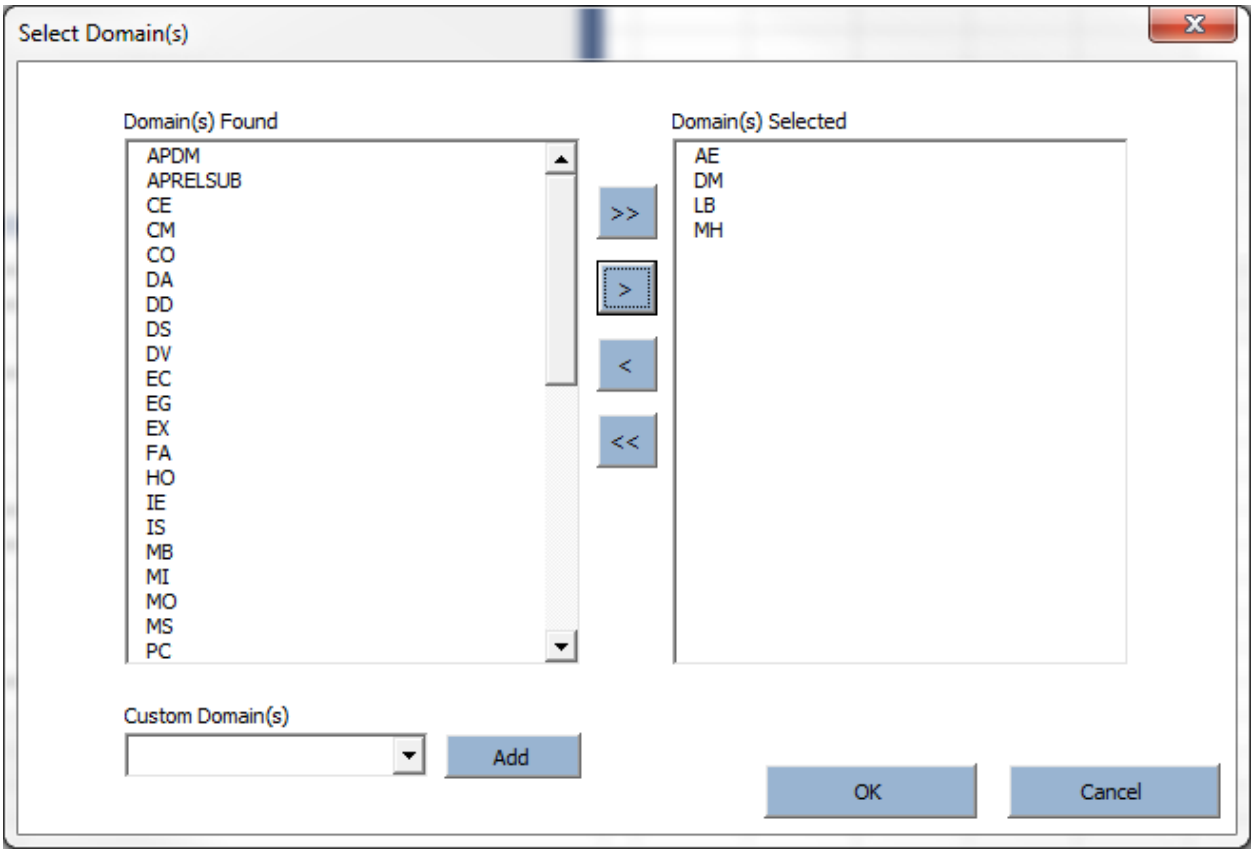


CDISC SDTM IG v3.2.

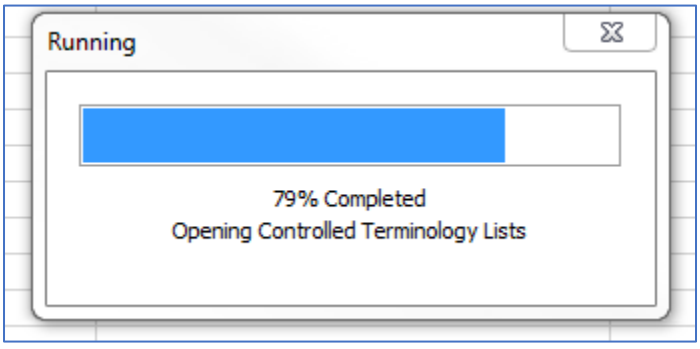


CDISC SDTM IG v3.1.3.

The user merely needs to select at least one domain for their study. For example, consider selecting the Demographics (DM), Adverse Events (AE), Laboratory Test Results (LB) and Medical History (MH) domains.



Note users can re-run this process and select more domains as needed at any time even. Upon clicking OK the utility will go to work, gathering the metadata for all the selected domains directly from the standards documents, which include all the domain variables, code lists, labels, data types, etc. Once complete the user is left with a 60% complete specification containing metadata pulled directly from the standards documents. Note, while the process is running the user will see a progress bar to help the user understand what is happening behind the scenes.



Consider the example detailed above, once the process has finished the user will be left with a mapping specification template, which has grown from one single worksheet to several. Each new worksheet will contain information pulled directly from the standards documents. Below outlines each new worksheet created by the process.

A *Domains* worksheet - this will hold the domain level metadata such as domain name, domain label, key variables, etc.

A	B	C	D
SDTM Implementation Guide: 3.2			
SDTM Controlled Terminology: 2018-06-29			
Domain Name	Domain Label	Class Name	Key Variables
AE	Adverse Events	Events	STUDYID USUBJID AETERM AESTDTC
DM	Demographics	Special-Purpose	STUDYID USUBJID
LB	Laboratory Test Results	Findings	STUDYID USUBJID LBTESTCD LBDTC
MH	Medical History	Events	STUDYID USUBJID MHTERM MHSTDTC

A worksheet for each selected domain, e.g. DM, AE, LB and MH worksheet. These worksheets contain the variable level metadata such as variable name, variable label, data type, derivation details, etc.

C	D	E	F	G	
Variable	Label	Type	ODM Data Type	Origin	Mapping Details
JDYID	Study Identifier	Char	text	Protocol	
DOMAIN	Domain Abbreviation	Char	text	Assigned	DM
USUBJID	Unique Subject Identifier	Char	text	Derived	

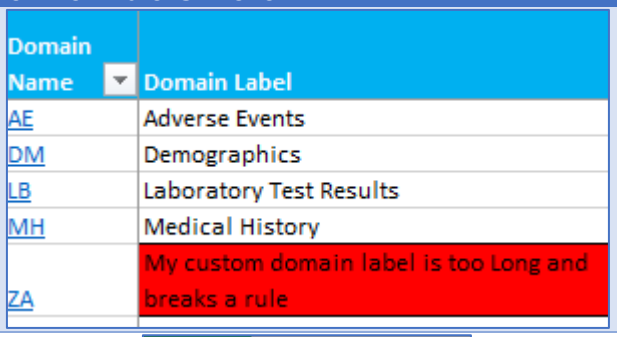
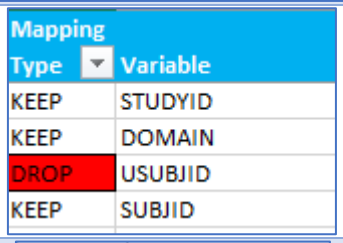
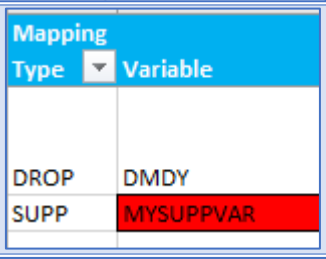
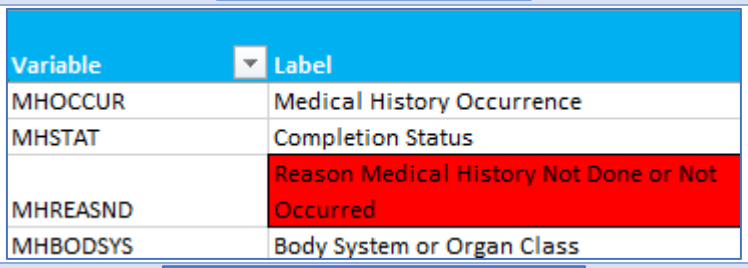
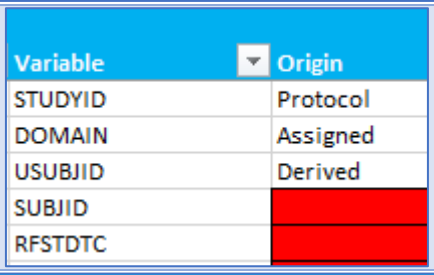
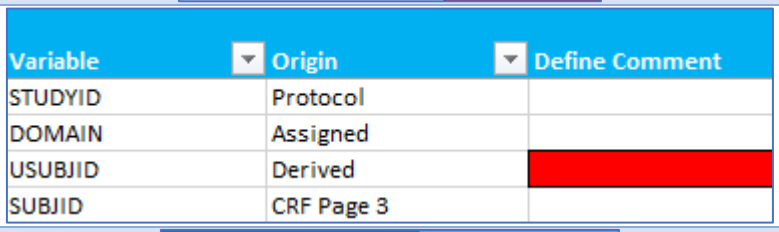
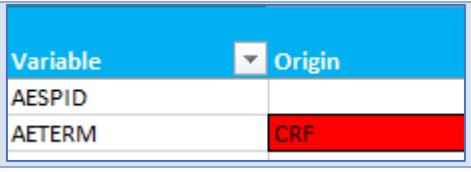
A worksheet which will eventually hold the Code Lists and associated Value Lists related to the study.

A	B	C	E
Code List Look-up	Code List Refresh		
Code List Name	Code List Label	Code List Code	Value List Name
ACN	Action Taken with Study Treatment	C66767	DOSE INCREASED
ACN	Action Taken with Study Treatment	C66767	DOSE NOT CHANGED
ACN	Action Taken with Study Treatment	C66767	DOSE RATE REDUCED
ACN	Action Taken with Study Treatment	C66767	DOSE REDUCED
ACN	Action Taken with Study Treatment	C66767	DRUG INTERRUPTED
ACN	Action Taken with Study Treatment	C66767	DRUG WITHDRAWN
ACN	Action Taken with Study Treatment	C66767	NOT APPLICABLE

Lastly, a worksheet holding the test / parameter code and value list information. This worksheet is like the code list worksheet pictured above however, it gathers more information about each value list, such as such as test / parameter origin and data type.

A	B	C	E	F
Value List Look-up	Value List Refresh			
Code List Name	Code List Label	Code List Code	Value List Name (--TESTCD)	Value List Label (--TEST)
LBTESTCD	Laboratory Test Code	C65047		

As already stated, at this point the mapping specification is about 60% complete. The reason for this percentage is the utility is not a “one click does all” tool, it is a stepping stone from the standards documentation into something more familiar to a programmer. This 60% complete template is also created with built in rules to aid the user complete the remaining 40% but also ensuring compliance is maintained. These built in “rules” use the conditional formatting feature of EXCEL which means when a rule is broken the cell breaking the rule is turned red automatically, this prompts users to investigate the problem cell straight away. By having rules built into the document allows the programmer to correct the problem in the beginning. Note, when a problem arises later in a study there is a much higher risk of a bottleneck. Below is a list of **some** of the rules incorporated into the template by default.

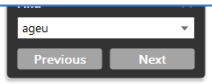
Rule	Example when Rule is Broken
Domain label cannot exceed 40 characters.	
Required or Expected variables cannot be dropped.	
Supplemental QNAM values cannot exceed 8 characters.	
Variable labels cannot exceed 40 characters.	
Variable Origin cannot be left null.	
Define comment must be provided when variable origin is set to derived.	
Case Report Form (CRF) page numbers must be included when origin is CRF.	

The remaining 40% of the mapping specification may sound like a lot of work, especially when a lot of companies may have their own templates which may be closer to 90% complete. However, remember the aim of this utility was never to create a completed specification. It was to create a dynamic and adaptable utility which is flexible to work with ever changing standards. This “90% complete mapping specification” may rely on the study being a very standard parallel early phase study for a specific indication, which is a very niche tool. That said, this remaining 40% is not all manual work and the utility does provide further tools to assist in the completion.

Consider the Code List and Value List worksheets, these worksheets will eventually contain all the controlled terms and test metadata related to the study. Adding all this metadata manually may sound like a lot of work however, the utility was built in parallel with SQN’s SAS programs. Consider a SAS program which will create the demographics (DM) output, during development of the SAS program the mapping specifications will be regularly imported and stored as datasets, i.e. each worksheet from the mapping specification will be stored as a SAS dataset. The reason for this is we can use the metadata from the specification to semi-automate certain aspects when developing any SDTM output. A very simple and common use of these imported datasets is to assign variable labels dynamically. The SQN SAS programs will gather the variable labels from the imported mapping specification datasets and assign the labels dynamically to the SDTM output instead of the user writing out SAS code. This is good practice because metadata should never be duplicated.

A more complex process would be using these imported datasets to run checks against controlled terminology values. Consider the DM program again, there are certain variables in DM which are controlled by CDISC terminology, e.g. the age unit’s variable (AGEU) is controlled by the age units code list, also called AGEU. What this means is any values within the AGEU variable must adhere to terms found in the AGEU code list. Below is the AGEU allowed terms from the NCI Controlled Terminology document.

AGEU (Age Unit)



NCI Code: C66781, Codelist extensible: No

C66781	AGEU			
NCI Code	**CDISC Submission Value**	**CDISC Synonym**	**CDISC Definition**	**NCI Preferred Term**
C25301	DAYS		A unit of measurement of time equal to 24 hours.	Day
C25529	HOURS	Hours;h;hr	A unit of measurement of time equal to 60 minutes.	Hour
C29846	MONTHS	Month	One of the 12 divisions of a year as determined by a calendar. It corresponds to the unit of time of approximately to one cycle of the moon's phases, about 30 days or 4 weeks. (NCI)	Month
C29844	WEEKS	Week	Any period of seven consecutive days. (NCI)	Week
C29848	YEARS	Year	The period of time that it takes for Earth to make a complete revolution around the sun, approximately 365 days; a specific one year period. (NCI)	Year

Within the mapping specifications there is a Code List worksheet, which will eventually hold all code lists (e.g. AGEU) and the applicable code list values. In the beginning this worksheet will simply list any code lists found in the domains the user selected in the beginning, e.g. if only DM was selected in the beginning then the Code List worksheet will contain a line for each code list associated with the DM domain, which are SEX, RACE, AGEU, NY and DOMAIN.

1	Code List Name	Code List Label	Code List Code	Value List Name
2	AGEU	Age Unit	C66781	
3	DOMAIN	SDTM Domain Abbreviation	C66734	
4	NY	No Yes Response	C66742	
5	RACE	Race	C74457	
6	SEX	Sex	C66731	

Note by default only expected (EXP) or required (REQ) variables are kept, and only kept variables will have their code list added to the Code List worksheet in the beginning, therefore users will not see the ETHNIC code list present on the Code List worksheet in the beginning as the associated variable (also called ETHNIC) is permissible and not kept by default. However, the utility does have a feature to update the Code List worksheet for the user automatically. Once the DM worksheet has been completed, i.e. all permissible variables have been dropped and or kept the user can run this process by clicking a button which will update the Code List worksheet and add any permissible variables which have a code list associated to the code list worksheet. For example, if the ETHNIC variable was updated on the DM worksheet to be kept the process will add the ETHNIC code list to the Code List worksheet as a new line.

The Value List Name column in the image above will eventually be populated with values found within the study data. Users can fill this out manually by opening data sets and copy and pasting, but this is tedious and slow. With the correct setup and SAS programs this process can be semi-automated. Consider the DM SAS program which just mapped the SDTM variables straight from the raw variables and the Code List worksheet looked like the above image, i.e. empty. Upon submitting the SAS program, checks would fire, and the SAS log would display some useful information and the controlled variables.

```
WARNING: (MSDTERMINOLOGY_01) AGEU Code List not found or is empty (see below), please check the CodeLists tab.
WARNING: (MSDTERMINOLOGY_01) FINAL.AGEU value(s) missing from AGEU code list.
WARNING: (MSDTERMINOLOGY_01) YR
WARNING: (MSDTERMINOLOGY_01) FINAL.AGEU value not found on AGEU code lists, please check. AGEU=YR
WARNING: (MSDTERMINOLOGY_01) DOMAIN Code List not found or is empty (see below), please check the CodeLists tab.
WARNING: (MSDTERMINOLOGY_01) FINAL.DOMAIN value(s) missing from DOMAIN code list.
WARNING: (MSDTERMINOLOGY_01) DM
WARNING: (MSDTERMINOLOGY_01) FINAL.DOMAIN value not found on DOMAIN code lists, please check. DOMAIN=DM
WARNING: (MSDTERMINOLOGY_01) NY Code List not found or is empty (see below), please check the CodeLists tab.
WARNING: (MSDTERMINOLOGY_01) FINAL.DTHFL value(s) missing from NY code list.
WARNING: (MSDTERMINOLOGY_01) Y
WARNING: (MSDTERMINOLOGY_01) FINAL.DTHFL value not found on NY code lists, please check. DTHFL=Y
WARNING: (MSDTERMINOLOGY_01) RACE Code List not found or is empty (see below), please check the CodeLists tab.
WARNING: (MSDTERMINOLOGY_01) FINAL.RACE value(s) missing from RACE code list.
WARNING: (MSDTERMINOLOGY_01) Caucasian
WARNING: (MSDTERMINOLOGY_01) FINAL.RACE value not found on RACE code lists, please check. RACE=Caucasian
WARNING: (MSDTERMINOLOGY_01) SEX Code List not found or is empty (see below), please check the CodeLists tab.
WARNING: (MSDTERMINOLOGY_01) FINAL.SEX value(s) missing from SEX code list.
WARNING: (MSDTERMINOLOGY_01) Male
WARNING: (MSDTERMINOLOGY_01) FINAL.SEX value not found on SEX code lists, please check. SEX=Male
```

The messages above aids the user in populating the Code List worksheet far quicker than doing it by looking at the raw datasets. The user just needs to navigate to the Code List worksheet and locate the code lists listed above and add the values (also listed above) as appropriate, e.g. the AGEU code list and the value YR.

1	Code List Name	Code List Label	Code List Code	Raw Value (Optional)	Value List Name
2	AGEU	Age Unit	C66781	YR	YEARS
3	DOMAIN	SDTM Domain Abbreviation	C66734		DM
4	NY	No Yes Response	C66742		Y
5	RACE	Race	C74457	Caucasian	WHITE
6	SEX	Sex	C66731	Male	M

Note the additional column “Raw Value (Optional)” above, this column indicates to the user if the raw value has been decoded with the correct value which adds traceability for the user and sponsor, e.g. see RACE, SEX or AGEU lines above. The SAS programs use this “raw” column to perform automapping for the user, this replaces the need to write out IF THEN statements to map variables in turn reducing the risk of human error. Once the mapping specification has been updated, the DM program can be resubmitted, and the SAS log should now be clean. The output will be produced as expected, take note of the raw values of Caucasian, YR and Male, with no extra SAS code they have been remapped automatically.

RAW.DM

AGEU	SEX	RACE
YR	Male	Caucasian
YR	Male	Caucasian

SDTM.DM

AGEU	SEX	RACE
YEARS	M	WHITE
YEARS	M	WHITE

Similar to the Code Lists worksheet is the Test / Parameter worksheet, the tests / parameters are also controlled, but they require additional metadata when it comes to creating the Define.XML output, so they have been split into a separate worksheet. With the example above, where only the DM, AE, LB and MH domains were selected, then the Value List worksheet would only contain the LBTESTCD line, however if domains like ECG Test Results (EG) or Physical Examinations (PE) had been selected a line would exist for each of their test / parameter variables, e.g. EGTESTCD and PETESTCD.

	Code List Name	Code List Label	Code List Code	Value List Name (--TESTCD)	Value List Label (--TEST)
2	LBTESTCD	Laboratory Test Code	C65047		
3					

The same rule applies, when the LB program is initially submitted with the SAS log with display messages about the terms not being present within the mapping specification.

```
WARNING: (MSDTERMINOLOGY_01) FINAL.DOMAIN value not found on DOMAIN code lists, please check. DOMAIN=LB
WARNING: (MSDTERMINOLOGY_01) FINAL.LBBLFL value not found on NY code lists, please check. LBBLFL=YES
WARNING: (MSDTERMINOLOGY_01) LBCAT Code List not found or is empty (see below), please check the CodeLists tab.
WARNING: (MSDTERMINOLOGY_01) FINAL.LBCAT value(s) missing from LBCAT code list.
WARNING: (MSDTERMINOLOGY_01) CHEMISTRY
WARNING: (MSDTERMINOLOGY_01) HEMA
WARNING: (MSDTERMINOLOGY_01) FINAL.LBCAT value not found on LBCAT code lists, please check. LBCAT=CHEMISTRY
WARNING: (MSDTERMINOLOGY_01) FINAL.LBCAT value not found on LBCAT code lists, please check. LBCAT=HEMA
WARNING: (MSDTERMINOLOGY_01) UNIT Code List not found or is empty (see below), please check the CodeLists tab.
WARNING: (MSDTERMINOLOGY_01) FINAL.LBORRESU value(s) missing from UNIT code list.
WARNING: (MSDTERMINOLOGY_01) kg
WARNING: (MSDTERMINOLOGY_01) mg/dl
```

However, for test / parameter variables additional checks will fire and the following messages will also be displayed within the SAS log indicating to the user that these values need adding to the Value List worksheet. These checks work the same way as the code list checks work.

```
WARNING: (MSDTEST_01) LBTESTCD (or LBPARMCD) Value List not found or is empty (see below), please check the Value
WARNING: (MSDTEST_01) FINAL.LBTESTCD value(s) missing from LBTESTCD value list.
WARNING: (MSDTEST_01) GLUC
WARNING: (MSDTEST_01) S-HCT
WARNING: (MSDTEST_01) LBTESTCD Value List contains blank values, please check the ValueLevelMetadata tab.
```

This indicates all the test / parameter values found within the raw data which should be added to the Value List worksheet. The user should use the SAS log to help populate the Test / Parameter worksheet.

	Code List Name	Code List Label	Code List Code	Raw Value (Optional)	Value List Name (--TESTCD)
2	LBTESTCD	Laboratory Test Code	C65047		GLUC
3	LBTESTCD	Laboratory Test Code	C65047	S-HCT	HCT
4					

Now resubmitting the LB program will remove some of the warnings but not all. This is where the Value List worksheet differs from the Code List worksheet. After resubmitting LB the following messages will remain in the SAS log.

```
WARNING: (MSDTEST_01) FINAL.LBTESTCD value not found / label not provided on LBTESTCD Code List, LBTESTCD=GLUC
WARNING: (MSDTEST_01) FINAL.LBTESTCD value not found / label not provided on LBTESTCD Code List, LBTESTCD=S-HCT
```

This is because on the Test / Parameter worksheet there is an additional column which corresponds to the coded test / parameter (e.g. LBTESTCD) value's label (e.g. LBTEST). This column will also need populating, with the correct value before submitting LB. This value comes directly from the controlled terminology document, which is currently sitting close to 2000 pages. To manually look through this document takes a lot of time, however, the utility can search the document for the user and return the correct value in a matter of seconds. The user can run a "*Value List Search*" process which will do the look-up automatically. The process will search through the controlled terminology documentation for the user by using a combination of code list in question (e.g. LBTESTCD) then locating each specific value listed e.g. GLUC or HCT in the example above, and then return the correct label e.g. LBTEST value for LBTESTCD=GLUC and LBTESTCD=HCT. The process begins by clicking a button found on the Value List worksheet.

Note this process also returns the value's CDISC code which is used in the production of the Define.XML output later. Upon a successful run the Value List worksheet will now look like this.

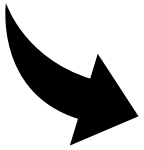
	Value List Look-up	Value List Refresh		Raw Value	Value List Name	Value List Label	Value List
1	Code List Name	Code List Label	Code List Co	(Optional)	(--TESTCD)	(--TEST)	Code
2	LBTESTCD	Laboratory Test Code	C65047		GLUC	Glucose	C105585
3	LBTESTCD	Laboratory Test Code	C65047	S-HCT	HCT	Hematocrit	C64796

Note this process can also be performed on the Code List worksheet. This Code List Search process can scan a 2000-page document and return the associated CDISC labels and or codes for every single value in about 5 seconds.

Once LB has been resubmit the macros will do their job and perform the automapping's like described for the Code Lists worksheet based on the information found in the specification. Additionally, they will also create the LBTEST variable for the user. This is because if the user derives LBTEST in their SAS program with IF THEN statements there is a chance for human error and this also creates duplicated metadata, because these test / parameter labels are already listed with in the mapping specification. Therefore, letting the macros use the mapping specification to generate automapping's is far more efficient.

RAW.LB

LBTESTCD	
S-HCT	2
GLUC	2



SDTM.LB

	LBTESTCD	LBTEST	
1	HCT	Hematocrit	H
1	GLUC	Glucose	G

The mapping specifications are just one part of the process and this illustrates how forward-thinking process can not only draft a 60% complete specification but also make the remaining 40% partially automated.

Annotated Case Report Form (CRF)

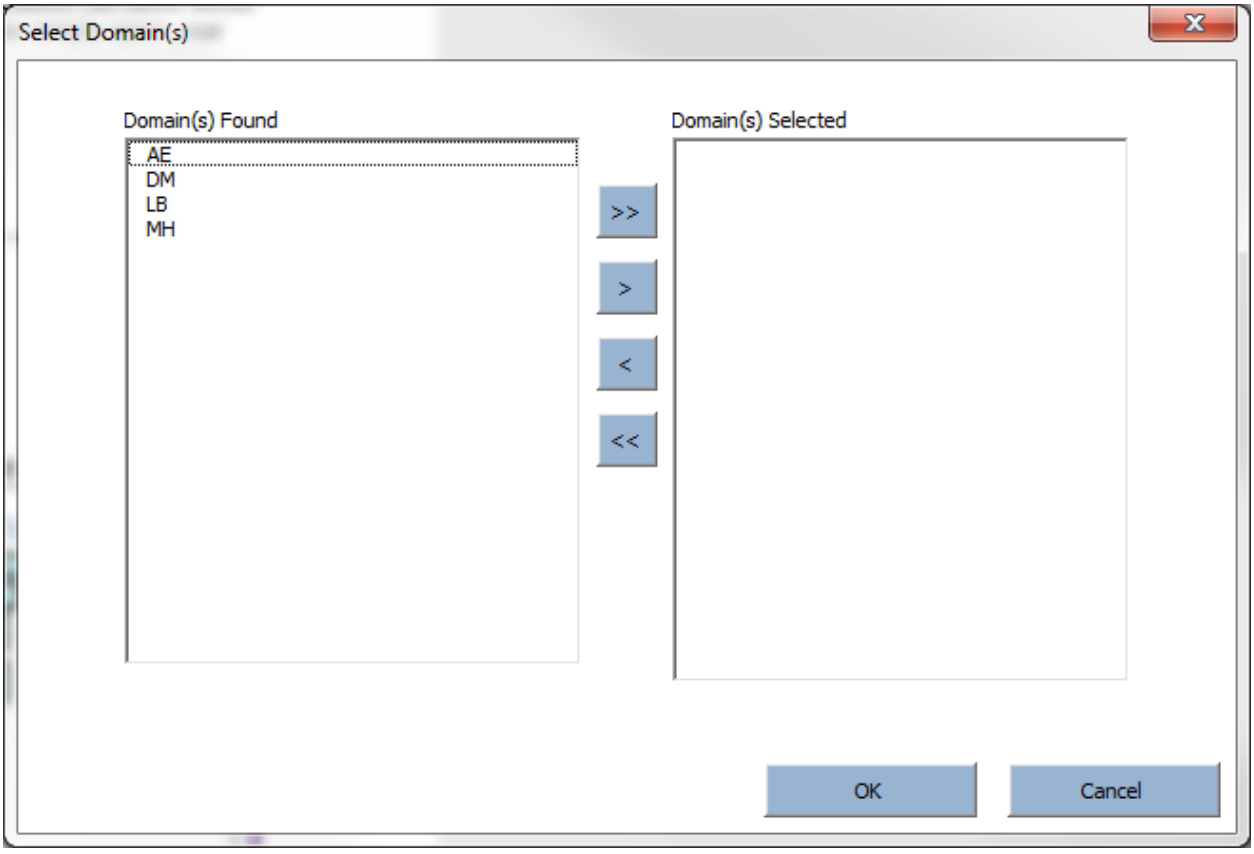
Following on from the mapping specifications the utility can also assist in creating the annotated Case Report Form (CRF). A CRF is a snapshot of the electronic Data Capture (eDC) system for a study. The CRF will contain all pages and modules from a single eDC system for a single study. For example, a CRF will show the demographic, lab and vital sign pages for a single study in a flat file format, such as PDF. Now an annotated CRF is a CRF with every data field labelled with the corresponding variable name, e.g. SDTM variable name. During the production of any SDTM dataset a large portion of the variables will be captured and collected within an eDC system, these variables will therefore come directly from the CRF. Programmers need to illustrate this mapping by creating an annotated CRF. This clearly indicates to a reviewer where an SDTM variable has come from and where the raw variable has been mapped to. The process of annotating the CRF is tedious because it typically consists of one person manually added text boxes into a PDF file one text box at a time.

This is where the second feature of the utility comes in. Provided the mapping specification has been drafted (even if only one domain has been drafted) the utility can generate the annotations for the user in-turn avoiding the manual creation of all the annotations. The idea is a mapping specification will contain an origin field for every single variable. This origin field will contain information about where the variable came from, examples of different origins within SDTM are Protocol, Derived, CRF, electronic Data Transfer (eDT) or Assigned.

Consider the Demographics (DM) domain, where several variables are coming directly from the CRF. Once the DM worksheet has been drafted by the user the DM worksheet will have the relevant CRF page numbers already listed, and these are the variables which will need to be annotated onto the CRF.

Variable	Origin
BRTHDTC	CRF Page 3
AGE	CRF Page 3
AGEU	Assigned
SEX	CRF Page 3
RACE	CRF Page 3

Since this CRF metadata already exists within the mapping specification the utility can use this metadata and create an annotation file for the user. The user can begin this annotation generation process by clicking "*Generate Annotations*" within the user interface. Upon doing so a window will appear listing all the domains available within their mapping specification, allowing the user to select the applicable domain(s).



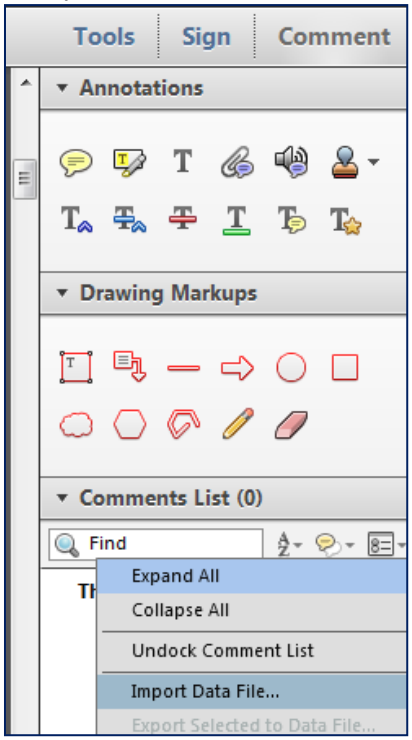
When at least one domain has been selected the utility will go to work gathering all the CRF metadata from each selected domain within the mapping specification, in less than 1 second. Once the process has finished an annotation file is created. This annotation file has the extension slightly unfamiliar and underused. The file extension is called an Acrobat Forms Data Format (FDF) file which is a file format specific to PDF files. This annotation file (FDF) contains the CRF annotations in a machine-readable format. All the user needs to do is import this annotation file (FDF) file into their CRF document (PDF) to create a partially complete annotated CRF.

1. Mapping Specifications

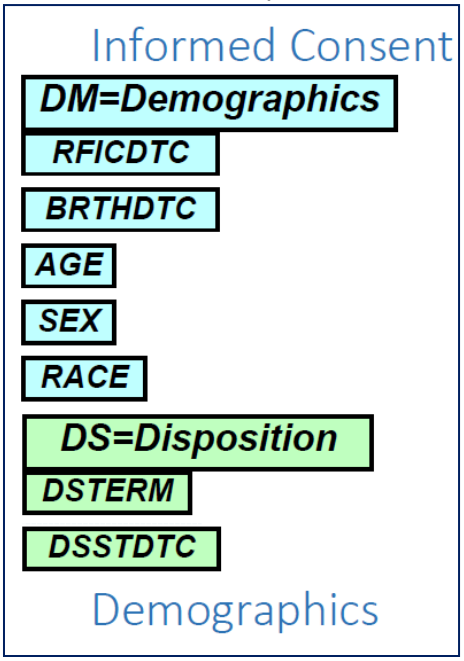
Variable	Origin
BRTHDTC	CRF Page 3
AGE	CRF Page 3
AGEU	Assigned
SEX	CRF Page 3
RACE	CRF Page 3

2. FDF File

3. Import the FDF File.



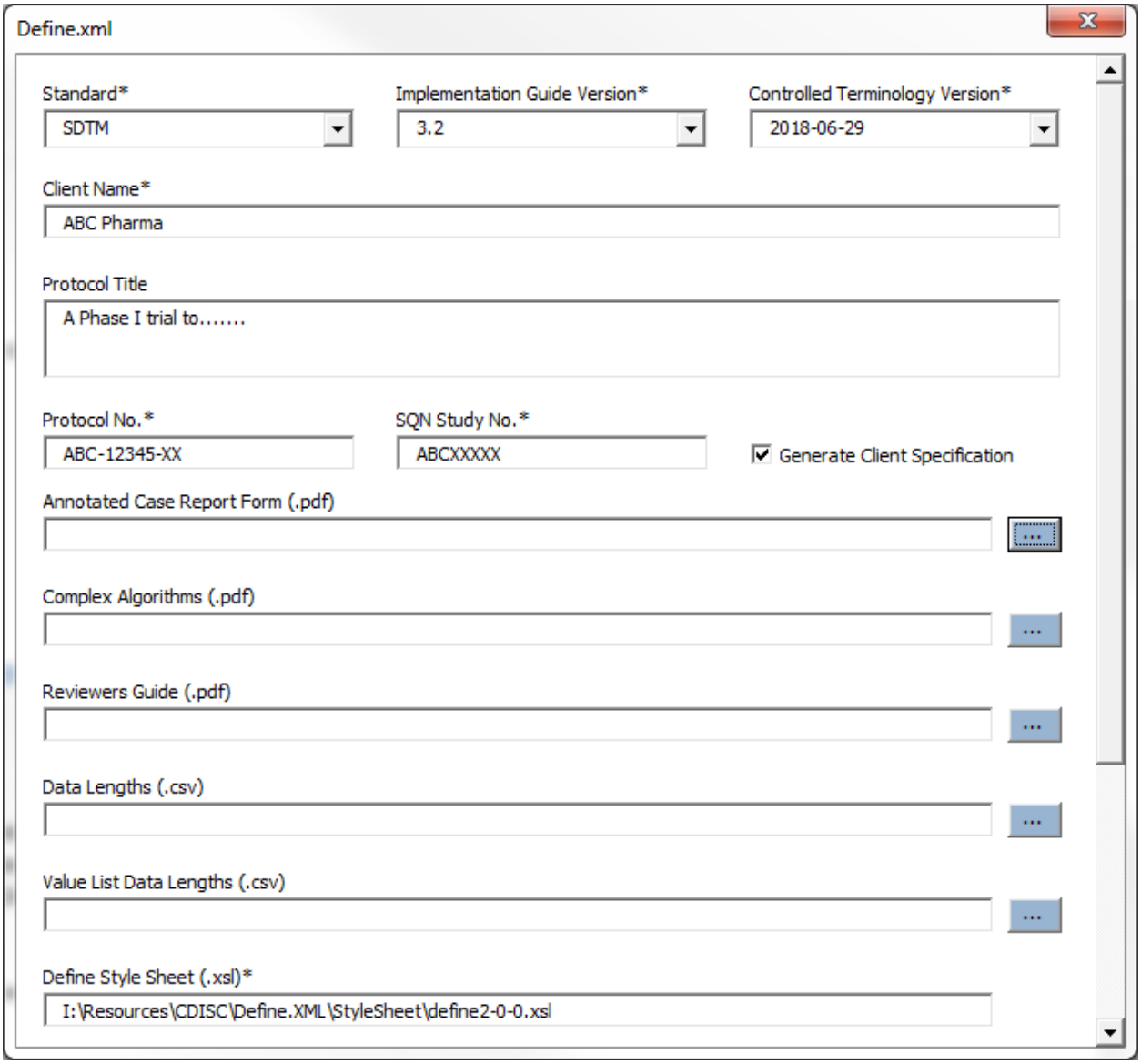
4. Annotated Case Report Form



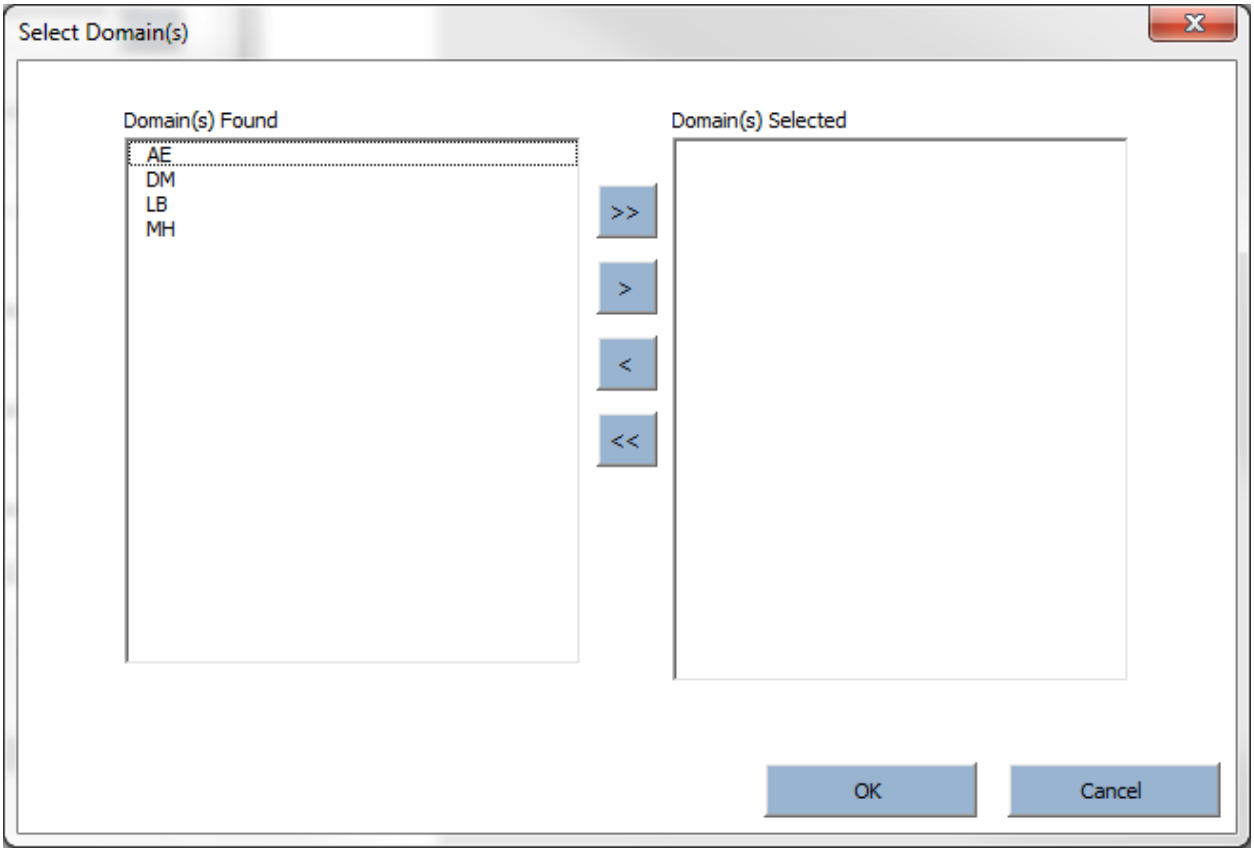
The utility cannot “magically” position an annotation perfectly onto a PDF document, hence the phrasing “partially complete annotated CRF”, but what it can do is insert the annotation on the correct page based on the mapping specification information and colour coordinate annotations when multiple domains occur on the same page, as above. Additionally, it can also handle repeat pages by printing a more suitable annotation instead of repeating the annotations, e.g. “see page X for annotations” where X is the page number for the first occurrence of the data within the CRF. This process hugely boost efficiency and quality when producing an annotated CRF.

Define.XML

The Define.XML is a file which consists of a study's full metadata, i.e. the study, domain, variable and value metadata. It gives the reviewer a better insight into the study, e.g. how variables were derived or where the variables have come form, etc. There are several approaches to create the Define.XML output, examples can consist of using the SAS data sets as the source, others will use a hybrid approach of using the data sets and mapping specifications as the source. This is where the utility differs, it relies solely on the mapping specifications as the source, no data sets. The big advantage of this is it allows the Define.XML to be produced prior to any data sets being ready or even output. It can give the reviewer insight into the study prior to any programming beginning. The utility was designed with the Define.XML output as a key driver, therefore the mapping specification generation process was created to retain all the information required to produce the Define.XML output. Therefore, the utility will not require any additional software and or data to produce the Define.XML output, the Define.XML can be produced purely from the mapping specification in EXCEL. This provides the user with a greater sense of control and traceability. Any updates made to the Define.XML are performed far quicker due to the source data being the mapping specification and not a SAS data set. The utility can gather the metadata within the mapping specification in a matter of seconds and produce a compliant CDISC Define.XML. The user can begin the process by clicking the "*Generate Define.XML*" button within the user interface. Once started, a window will appear which will allow the user to provide all the relevant file(s) and or location(s).

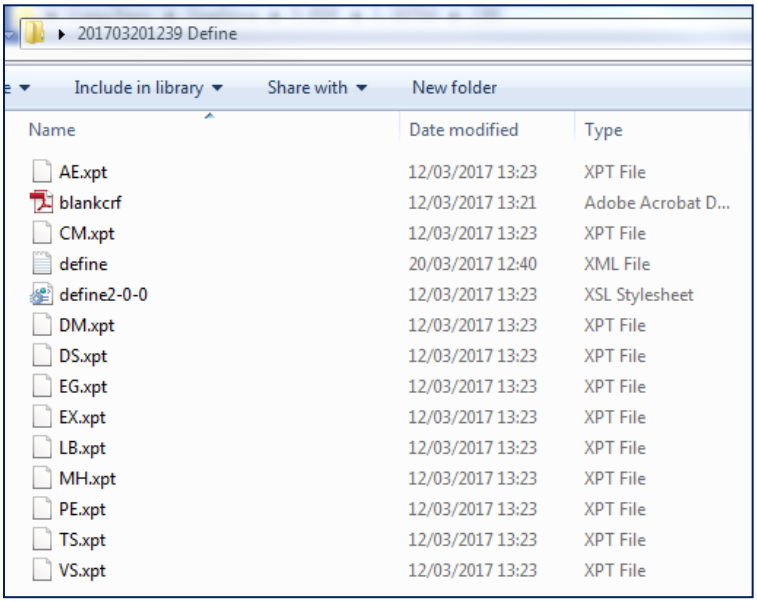


The window above provides fields for a plethora of documents, however if any of these documents have not been produced yet the Define.XML output can still be created. Similarly, if any domains within the mapping specification have not been complete they can be omitted using the domain selection window which will appear during the process and the Define.XML can still be created.



Once the process has run and dependent on which files were provided within the selection window the utility will create a Define.XML package which will consist of the following.

- Define.XML output
- XPT data sets
- Annotated Case Report Form (CRF).
- Define.XML Stylesheet
- Reviewer's Guide
- Mapping Specifications (.XLSX).



What the utility is doing is converting the EXCEL mapping specifications into a Define.XML output, no SAS programs, no third-party software, just EXCEL, with some VBA and a mapping specification. Consider the following example mapping specification and Define.XML output.

Domain Level Metadata (Mapping Specification)

3	Domain Name	Domain Label	Class Name	Key Variables
4	AE	Adverse Events	Events	STUDYID USUBJID AETERM AESTDTC AEENDTC
5	CM	Concomitant Medications	Interventions	STUDYID USUBJID CMTRT CMSTDTC
6	CO	Comments	Special-Purpose	STUDYID USUBJID RDOMAIN IDVARVAL
7	DM	Demographics	Special-Purpose	STUDYID USUBJID
8	DS	Disposition	Events	STUDYID USUBJID DSSTDTC DSTERM
9	EX	Exposure	Interventions	STUDYID USUBJID EXTRT EXSTDTC
10	LB	Laboratory Test Results	Findings	STUDYID USUBJID LBTESTCD LBDTC LBTPTNUM LBSTRESC
11	MH	Medical History	Events	STUDYID USUBJID MHSCAT MHTERM MHDTC
12	PE	Physical Examination	Findings	STUDYID USUBJID PETESTCD PEDTC

Domain Level Metadata (Define.XML)

Tabulation Datasets for Study A (SDTM-IG 3.2)

Dataset	Description	Class	Structure	Purpose	Keys	Location
TA	Trial Arms	TRIAL DESIGN	One record per planned element per arm	Tabulation	STUDYID	TA.xpt
TE	Trial Elements	TRIAL DESIGN	One record per planned element	Tabulation	STUDYID	TE.xpt
TI	Trial Inclusion/Exclusion Criteria	TRIAL DESIGN	One record per I/E criterion	Tabulation	STUDYID	TI.xpt
TS	Trial Summary	TRIAL DESIGN	One record per trial summary parameter value	Tabulation	STUDYID, TSPARMCD	TS.xpt
TV	Trial Visits	TRIAL DESIGN	One record per planned visit per arm	Tabulation	STUDYID	TV.xpt
CO	Comments	SPECIAL PURPOSE	One record per comment per subject	Tabulation	STUDYID, RDOMAIN, USUBJID, IDVARVAL	CO.xpt
DM	Demographics	SPECIAL PURPOSE	One record per subject	Tabulation	STUDYID, USUBJID	DM.xpt

Variable Level Metadata (Mapping Specification)

Variable	Label	Type	ODM Data Type	Origin
STUDYID	Study Identifier	Char	text	Protocol
DOMAIN	Domain Abbreviation	Char	text	Assigned
USUBJID	Unique Subject Identifier	Char	text	Derived
SUBJID	Subject Identifier for the Study	Char	text	CRF Page 1
RFSTDTC	Subject Reference Start Date/Time	Char	date	CRF Page 46
RFENDTC	Subject Reference End Date/Time	Char	date	CRF Page 46
RFXSTDTC	Date/Time of First Study Treatment	Char	date	CRF Page 46

Variable Level Metadata (Define.XML)

Demographics (DM) [Location: DM.xpt]

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin
STUDYID	Study Identifier	1	text	9		Protocol
DOMAIN	Domain Abbreviation		text	2	SDTM Domain Abbreviation	Assigned
USUBJID	Unique Subject Identifier	2	text	16		Derived
SUBJID	Subject Identifier for the Study		text	6		CRF Page 1
RFSTDTC	Subject Reference Start Date/Time		date		ISO8601	CRF Page 46
RFENDTC	Subject Reference End Date/Time		date		ISO8601	CRF Page 46
RFXSTDTC	Date/Time of First Study		date		ISO8601	CRF Page 46

Supplemental Value Level Metadata (Mapping Specification)

Variable	Label	Type	ODM Data Type	Origin
CMROUTE2	Original route code	Char	text	CRF Page 44 45

Supplemental Value Level Metadata (Define.XML)

Value Level Metadata - SUPPCM [QVAL]

Variable	Where	Type	Length / Display Format	Controlled Terms or Format	Origin	Derivation/Comment
QVAL	QNAM = "CMROUTE2" (Original route code)	text	5		CRF Pages 44 45	

Test / Parameter Value Level Metadata (Mapping Specification)

	Value List Look-up		Value List Name (--			Value List ODM	Data
1	Code List Name	Code List Cod	TESTCD)	Value List Label (--TEST)	Value List Co	Data Type	Length
15	LBTESTCD	C65047	ALB	Albumin	C64431	integer	3
16	LBTESTCD	C65047	CREAT	Creatinine	C64547	integer	3
17	LBTESTCD	C65047	ALBCREAT	Microalbumin/Creatinine Ratio	C74761	float (2)	4
18	LBTESTCD	C65047	ALP	Alkaline Phosphatase	C64432	integer	3
19	LBTESTCD	C65047	ALT	Alanine Aminotransferase	C64433	integer	3

Test / Parameter Value Level Metadata (Define.XML)

Value Level Metadata - LB [LBORRES]

Variable	Where	Type	Length / Display Format	Controlled Terms or Format	Origin
LBORRES	LBTESTCD = "ALB" (Albumin)	integer	3		CRF Pages 6 7 14 15 17 18 22 25 26 28 32 36 37 40 41
LBORRES	LBTESTCD = "CREAT" (Creatinine)	integer	3		CRF Pages 6 7 14 15 17 18 22 25 26 28 32 36 37 40 41
LBORRES	LBTESTCD = "ALBCREAT" (Microalbumin/Creatinine Ratio)	float	4		CRF Pages 6 7 14 15 17 18 22 25 26 28 32 36 37 40 41
LBORRES	LBTESTCD = "ALP" (Alkaline Phosphatase)	integer	3		CRF Pages 6 7 14 15 17 18 22 25 26 28 32 36 37 40 41
LBORRES	LBTESTCD = "ALT" (Alanine Aminotransferase)	integer	3		CRF Pages 6 7 14 15 17 18 22 25 26 28 32 36 37 40 41

Test / Parameter Code Lists (Mapping Specification)

Value List Name (--TESTCD)	Value List Label (--TEST)	Value List Co
ALB	Albumin	C64431
CREAT	Creatinine	C64547
ALBCREAT	Microalbumin/Creatinine Ratio	C74761
ALP	Alkaline Phosphatase	C64432
ALT	Alanine Aminotransferase	C64433
TT	Thrombin Time	C80365
TTC	Thrombin Time Control	
APTT	Activated Partial Thromboplastin Time	C38462
APTTC	APTT Control	
AST	Aspartate Aminotransferase	C64467

Test / Parameter Code Lists (Define.XML)

Laboratory Test Code [CL.LBTESTCD, *C65047*]

Permitted Value (Code)	Display Value (Decode)
ALB [*C64431*]	Albumin
CREAT [*C64547*]	Creatinine
ALBCREAT [*C74761*]	Microalbumin/Creatinine Ratio
ALP [*C64432*]	Alkaline Phosphatase
ALT [*C64433*]	Alanine Aminotransferase
TT [*C80365*]	Thrombin Time
TTC [*]	Thrombin Time Control
APTT [*C38462*]	Activated Partial Thromboplastin Time
APTTC [*]	APTT Control
AST [*C64467*]	Aspartate Aminotransferase

Other Code Lists (Mapping Specification)

	Code List Look-up				
1	Code List Name	Code List Code	Value List Name	Value List Label	Value List Code
97	RACE	C74457	ASIAN		C41260
98	RACE	C74457	BLACK OR AFRICAN AMERICAN		C16352
99	RACE	C74457	OTHER	Other	
100	RACE	C74457	WHITE		C41261

Other Code Lists (Define.XML)

Race [CL.RACE, *C74457*]

Permitted Value (Code)	Display Value (Decode)
ASIAN [*C41260*]	ASIAN
BLACK OR AFRICAN AMERICAN [*C16352*]	BLACK OR AFRICAN AMERICAN
OTHER [*]	Other
WHITE [*C41261*]	WHITE

* Extended Value

Conclusion

In conclusion the paper suggests that with the correct setup and foresight an efficient and dynamic process can be created by simply using EXCEL, no third-party software. A process which can not only handle the mapping specification creation and annotate the CRF for a user but create a compliant Define.XML output before programming has even begun. A process that expects change and is adaptable to new standards. A process which reduces the duplication of metadata and uses the CDISC and NCI standards. A process which “Stops copying CDISC standards” and does not create a “static” MDR.

References

The National Cancer Institute (NCI) Terminology: https://www.cdisc.org/standards/semantics/terminology
Clinical Data Interchange Standards Consortium (CDISC): https://www.cdisc.org/

Contact Information

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

Author: Craig Parry
Company: SyneQuaNon
Address: Gostling House / Diss / IP22 4GT
Email: Craig.Parry@SyneQuaNon.com
Web: https://www.synequanon.com/