PhUSE 2010

Paper CD13

Clinical Data Interchange Standards Consortium (CDISC) integration into the Oracle Clinical/Remote Data Capture® (OC/RDC) clinical data management system

Peter Van Reusel, Business & Decision Life Sciences, Brussels, Belgium
Nico De Leeuw, Business & Decision Life Sciences, Brussels, Belgium

ABSTRACT

Submissions of New Drug Applications (NDAs) to the Food and Drug Administration (FDA) in the United States require submitting clinical study data according to the CDISC standards. Currently, most of the pharmaceutical companies convert their clinical study data at the end of the clinical data management process cycle which may lead to significant NDA submission delays and an increased overall cost. Upfront integration of the CDISC standards into the sponsor's OC/RDC clinical data management system offers many advantages and is considered to be the longer term solution.

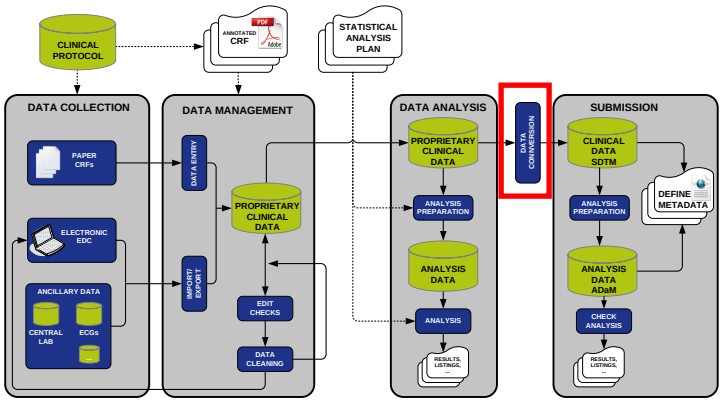
INTRODUCTION

The Clinical Data Interchange Standards Consortium (CDISC) develops and supports global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. According to the PhRMA-Gartner-CDISC project (1) when standards are implemented in the study start-up stage, there is a per study cycle time reduction of 8.1 months and a per study cost saving of $9 million. The FDA has embraced the CDISC SDTM and CRT-DDS as the standard data model for submitting tabulation data in an electronic format. FDA reviewers are requesting sponsors to use the SDTM model as it vastly improves their review efficiency and accuracy.

The presented CDISC integration approach applies the Clinical Data Acquisition Standards Harmonization (CDASH), Study Data Tabulation Model (SDTM) and Case Report Tabulation Data Definition Specification (CRT-DDS) standards. CDASH governs data collection standards and guidelines for Case Report Forms, SDTM standards guide the organization, structure and format of the clinical trial tabulation datasets, and CRT-DDS provides a standard XML format for submission.

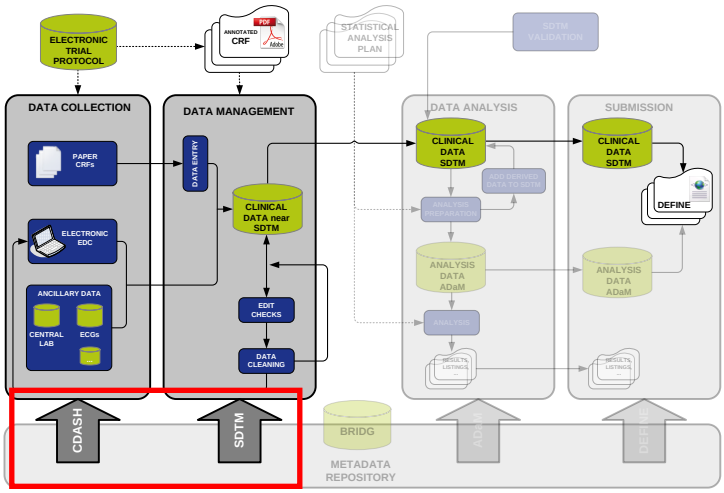
The advantages of the legacy data conversions (Figure #1) are that it has no impact on the clinical data management system and can easily be outsourced to a vendor. The drawbacks are the scalability limitations; the legacy conversion for the different studies may result in incompatibilities at a data and metadata level. It also takes a lot of time to convert data at the end of a study and this can be crucial, especially, for the clinical studies on the critical path. Finally, the statistical analysis and ADaM datasets are not generated from the SDTM datasets which will lead to difficulties in traceability between SDTM and the statistical output.

Figure #1: The legacy data conversion process



Oracle Clinical provides the life science industry with an integrated Clinical Data Management (CDM) and Remote Data Capture (RDC) application. This Clinical/Remote Data Capture (or OC/RDC) system is widely used across the pharmaceutical industry. Therefore, integrating the CDASH and SDTM data standards and guidelines upfront into OC/RDC was considered to be a relevant business case. (Figure #2) The advantages of one integrated system are application of identical nomenclature, use of a system wide data dictionary and reduction of data inconsistencies. In addition, it can easily drive the definition of the data collection tools in RDC and as such significantly reduce the time to clinical study set-up.

Figure #2: CDISC integration into clinical data management system



In summary, the main objectives of the CDASH/SDTM integration into OC/RDC are:

- Use identical nomenclature across the global Clinical Data Management system
- Apply CDISC controlled terminology
- Avoid additional data conversion steps
- Allow CDISC consistency checking early in the process
- Produce real-time SDTM deliverables at study set-up
- Facilitate generation of ADaM datasets based on SDTM

The implemented methodology is using the following approach:

- Instead of using the sponsor specific variables, a combination of CDASH and SDTM variables for data collection is applied whenever possible
- CDISC code lists are used instead of sponsor dictionaries wherever possible
- The variables not matching CDASH and SDTM are moved automatically into a SUPPQUAL domain
- A generic SDTM generator is developed, whereby all SDTM conversions occur outside OC/RDC. No additional programming is required.
- Finally, real-time SDTM datasets are generated ready for analysis

In the meantime, this methodology has been proven to be successful. This paper describes in 10 different steps (grouped together in four different phases) how to automatically generate submission ready SDTM datasets at study set-up.

Adverse Event data was used as an example to illustrate the SDTM conversion process.

PHASE 1: CDASH/SDTM INTEGRATION IN OC/RDC

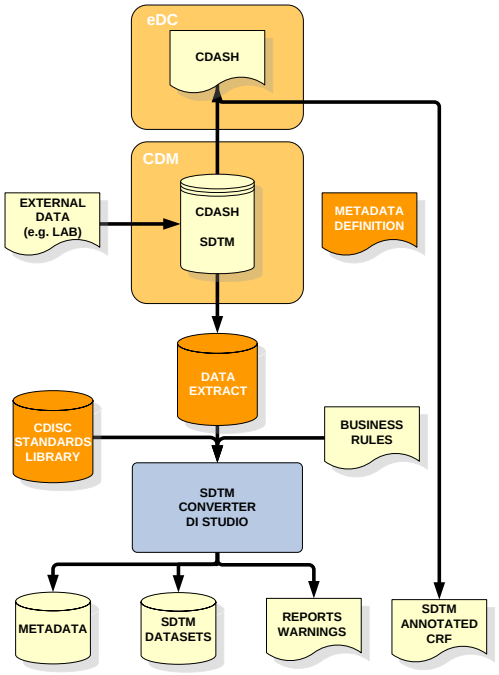
When integrating CDASH and SDTM into the OC/RDC system, the following system development steps need to be taken (Figure #3):

- Defining the metadata specifications
- Creating the global library and study definition
- Developing the CDASH/SDTM annotated CRF

In addition, a minimum of set-up rules and naming conventions are defined to make the entire process as efficient as possible, e.g.:

- Use all non-CDASH and non-SDTM variables to populate the SUPP domains
- Apply the extended attribute DVG_LONG_VALUE to populate TEST
- Use extended attributes for numeric variables to produce SDTM+

Figure #3: The SDTM data conversion flow



STEP 1: METADATA SPECIFICATIONS

Metadata was defined on a generic specification sheet per domain and then transmitted into the global library of Oracle Clinical. All CDASH and SDTM variables possible are uploaded together with the appropriate metadata information. Herewith a subset of the AE specification sheet:

Question name	Question Type	Length	SAS name	SAS label	DVG attached	Enterable	Displayed
AEYN	Char	2			NY	x	x
AETERM	Char	40	AETERM	Reported Term for Adverse Event		x	x
Etc.							

The CDASH variables will be used to collect and validate data but will never end up into the final SDTM datasets.

STEP 2: GLOBAL LIBRARY AND STUDY DEFINITION

The Oracle Clinical Global Library consists of four important components:

Questions

Any data element collected on a CRF in a clinical trial is called a “question”. Logically they map to SAS® variables. So CDASH, SDTM & other non-standard variables are defined as Questions. For example, the Questions for dataset Adverse Event include:

- CDASH only variables: AEYN, AESTDAT, AESTTIM, AEENDAT, AEENTIM, AEONGO
- SDTM only variables: STUDYID, DOMAIN, USUBJID, AESEQ, AEDECOD, AEBODSYS, AESTDTC, AEENDTC, AEENRF, AECONTRT
- Both CDASH & SDTM variables: AESPID, AETERM, AESEV, AETOXGR, AESER, AESCONG, AESDISAB, AESDTH, AESHOSP, AESLIFE, AESMIE, AEREL, AEACN, AEOUT
- Non-Standard variables: AEFUPOUT, AE_COVAL

CDASH is essentially a Data Collection Model and SDTM is a Data Representation Model so only implementing CDASH in Oracle is not a complete solution. To fully generate SDTM Datasets from Oracle, SDTM as well as Non-Standard variables need to be included in Oracle.

Question groups

A Question Group is a group of questions with a logical relationship to each other. A Question Group corresponds to a SAS Table.

For each Data Collection Module (DCM), Question Groups need to be defined which include questions logically linked to each other. So, for AE DCM, a Question Group AE which includes all the Questions for Adverse Events as defined above is created. The SDTM variables end up in the SDTM AE dataset and the non-standard variable AEFUPOUT ends up in SUPPAE. AE_COVAL goes to the CO (Comments) dataset.

SDTM Code list or Discrete Value Groups (DVGs)

A Discrete Value Group is a set of valid responses to a Question. Logically they correspond to a SDTM Code list, in SAS it can be regarded as a format.

For example, for Questions for AE Question group, a DVG NY was defined which had as valid responses: YES & NO. So for the Questions which have DVG NY only YES & NO can be entered.

Form Layout Template (FLT)

The Form Layout Template is needed to generate the Graphic Layout for RDC. It describes the Work Area which is the area between header & footer.

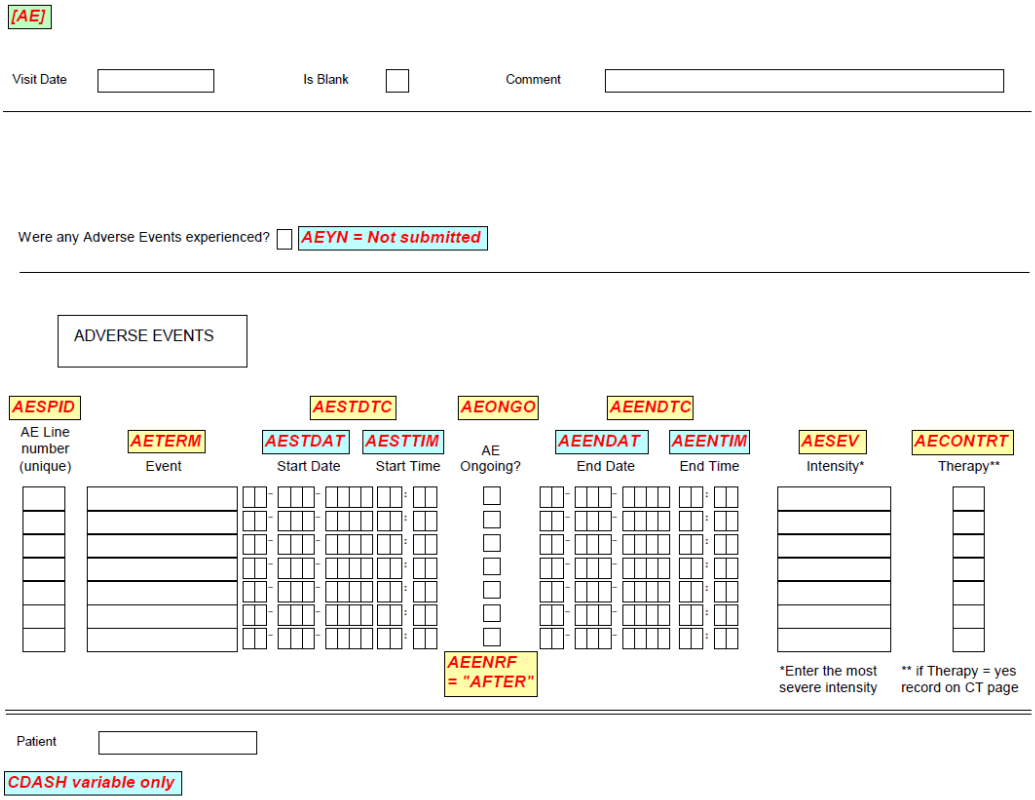
These four components, i.e. Questions, Question Groups, DVGs & FLT are defined in the Global Library to enable their re-use across studies.

In a next step, a selection of the variables will be applied to the specific study; this will consist of a selection of CDASH, SDTM and non-CDASH/non-SDTM variables. In the very last step, only the SDTM, SUPPQUAL and COMMENT variables will be kept. No conversions are applied to the clinical data at this level.

STEP 3: DEVELOPING CDASH/SDTM ANNOTATED ECRF

The forms in combination with the visit schedule result in the development of the eCRF. An eCRF in .pdf format can easily be generated from the system; Adobe Acrobat Professional® is used to annotate the eCRF (Figure #4) with the corresponding CDASH and SDTM variables.

Figure #4: Annotated AE eCRF



PHASE 2: STDM CONVERSION INPUT

Upon completion of phase 1, the SDTM conversion input can be prepared. The input consists of the following:

- O*C extracts
- CDISC standards library
- Business rules, default values and parameter paths

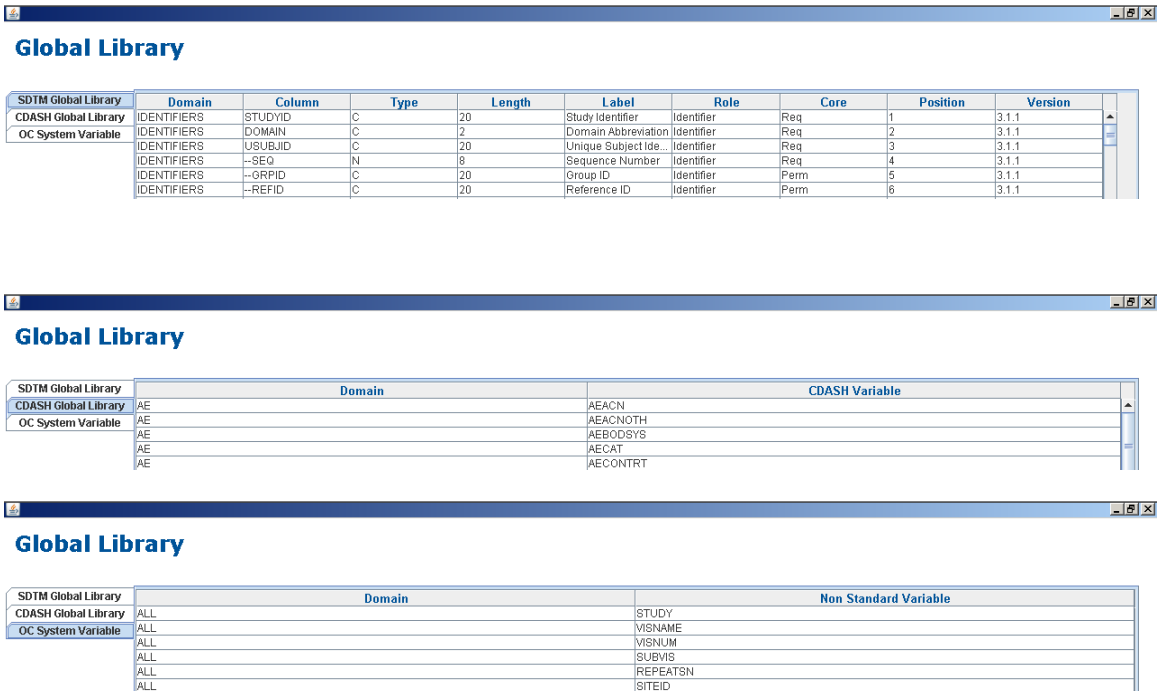
STEP 4: O*C EXTRACTS

Once OC/RDC is set-up for a specific study, the O*C extract views containing only the entered patient data can be created. The variables created in the view include CDASH, SDTM as well as non-CDASH/non-SDTM variables; the standard O*C system variables are included in all views.

STEP 5: CDISC STANDARDS LIBRARY

The CDISC standards library (Figure #5) is used by the SDTM generator to interpret standard source or system variables i.e., SDTM or CDASH or standard O*C system variables.

Figure #5: CDISC standards library interface



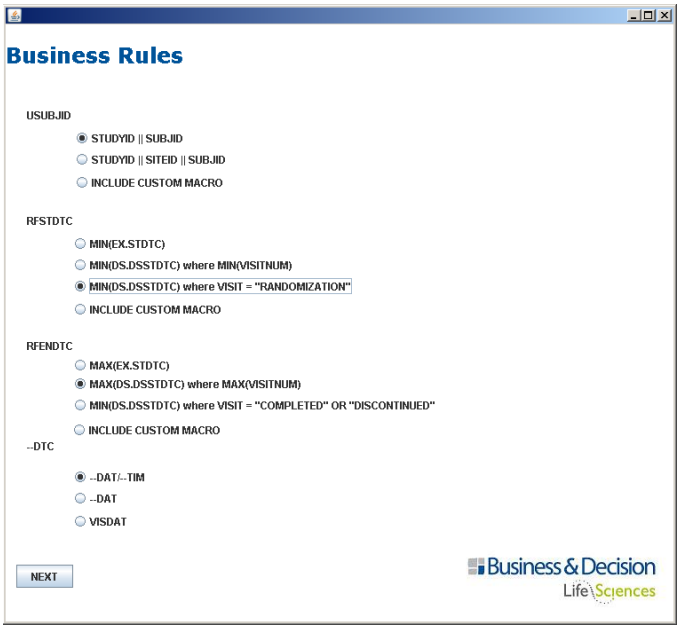
STEP 6: BUSINESS RULES, DEFAULT VALUES AND PARAMETER PATHS

The business rules (Figure #6) define the generic SDTM conversion choices. E.g.; to generate USUBJID, three options have been defined. The first one concatenates STUDY with SUBJID, the second one concatenates STUDY, SITEID and SUBJID and the last option allows including a custom macro. Next to USUBJID, other variables like RFSTDTC, RFENDTC, --DTC are also defined upfront.

The default values are entered in an editable table used by the SDTM generator. E.g.; if the DOSE FORM of the study medication is not captured in O*C; then it can be entered in this default values sheet with the appropriate value (e.g., INJECTION).

The parameter paths are defining the project specific paths in a Central Parameter Table.

Figure #6: Business rules interface



PHASE 3: STDM CONVERSION

During this phase the O*C data output is converted into SDTM datasets based on a set of SAS programs.

STEP 7: LIST OF SAS PROGRAMS

The core of the SDTM generator exists out of a list of Base SAS code (located in SAS DI Studio®). As input, the SDTM generator uses the previously described CDISC standards library, O*C extracts, and the Business rules, defaults values and parameter paths.

In the following example, it uses the AE O*C extract and applies the following conversions:

- **Domain specific transformation:** converts the QG name to the DOMAIN name; i.e. AE

- **Global transformation based on business rules:** this rule computes values based on rules specified in the business rule sheet

E.g.: To generate USUBJID; a number of default business rules are specified; including a custom macro. For this specific study, the STUDY with the SUBJID was selected to be concatenated.

INPUT			OUTPUT	
STUDY	**SUBJID**	**STUDY**	**SUBJID**	**USUBJID**
ABC	1	ABC	1	ABC-1
ABC	2	ABC	2	ABC-2
ABC	3	ABC	3	ABC-3

- **ISO 8601 data conversion:** this rule converts the variables ending with DAT & TIM in ISO 8601 date format e.g.:

INPUT				OUTPUT	
AESTDAT	**AESTTIM**	**AEENDAT**	**AEENTIM**	**AESTDTC**	**AEENDTC**
20090101	1200	20090601		2009-01-01T12:00	2009-06-01
200901	1200	20090601	1200	2009-01--T12:00	2009-09-01T12:00
2009			1400	2009	-----T14/00

- **Populate SEQ variable:** this rule generates SEQ variable based on Keys specified in Table Level Metadata in the CDISC Standards Library

TABLE LEVEL METADATA					
Dataset	**Description**	**Class**	**Structure**	**Purpose**	**Keys**
AE	Adverse Events	Events	One record per adverse event per subject	Tabulation	STUDYID, USUBJID, AETERM, AESTDTC
Etc.					

Note that for custom domains, a rule for events, interventions and findings has been defined.

- **Create Supplementary datasets:** this rule retrieves the non-CDASH, non-SDTM, and non-O*C system variables to populate the supplementary datasets.

INPUT		OUTPUT					
AESEQ	**AEFUPOUT, LABEL, AE FU TO OUTCOME**	**RDOMAIN**	**IDVAR**	**IDVARVAL**	**QNAM**	**QLABEL**	**QVAL**
1	FOLLOW-UP SUFFICIENT	AE	AESEQ	1	AEFUPOUT	AE FU TO OUTCOME	FOLLOW-UP SUFFICIENT
2	LOST TO FOLLOW-UP	AE	AESEQ	2	AEFUPOUT	AE FU TO OUTCOME	LOST TO FOLLOW-UP
3	FOLLOW-UP SUFFICIENT	AE	AESEQ	3	AEFUPOUT	AE FU TO OUTCOME	FOLLOW-UP SUFFICIENT

- **Create Comment dataset:** this rule collects comments in O*C in the &DOMAIN._COVAL variable.

INPUT		OUTPUT			
AESEQ	**AE_COVAL**	**RDOMAIN**	**IDVAR**	**IDVARVAL**	**COVAL**
1	Comment1	AE	AESEQ	1	Comment1
2	Comment2	AE	AESEQ	2	Comment2
3	Comment3	AE	AESEQ	3	Comment3

- **Create finally the SDTM dataset:** this rule uses variable level attributes from the CDISC standards library.

VARIABLE LEVEL METADATA				
SASTableName	**SASColumnName**	**SASColumnType**	**SASColumnLength**	**SASLabel**
AE	STUDYID	C	20	Study Identifier
AE	DOMAIN	C	2	Domain Abbreviation
AE	USUBJID	C	20	Unique Subject Identifier

In summary, below is the list of functionalities used to convert the O*C extract for Adverse Event into a 100% SDTM compliant AE dataset.

Functionality	Input	Output
Domain specific transformations based on default values	(QG NAME)	DOMAIN
Global transformation based on business rules	STUDY STUDY, SUBJID AEONGO	STUDYID USUBJID AEENRF
ISO 8601 data conversion	AESTDAT, AESTTIM AEENDAT, AEENTIM	AESTDTC AEENDTC
Populate SEQ variable		AESEQ
Create supplementary datasets	"AEFUPOUT" AEFUPOUT LABEL OF AEFUPOUT	QNAM QVAL QLABEL
Create Comment dataset	DCMDATE AE_COVAL	CODTC COVAL
Create SDTM dataset		Only SDTM variables are kept; the others are dropped

Note that the SDTM generator includes other functionalities like e.g.:

- automatically transposing finding domains
- extracting TEST from DVG to generate the TESTCD
- Etc.

PHASE 4: STDM CONVERSION OUTPUT

Once phase 3 is completed, and the SDTM generator has run, the following output is produced

- Metadata (in report, define.xml and/or define.pdf format)
- SDTM datasets
- Data Profiling and Warning reports

STEP 8: METADATA

Metadata are retrieved at a table, variable and value level; including controlled terminology.

A metadata report, define.xml and/or define.pdf can be retrieved from the SDTM generator.

Table Metadata	Dataset Name, Dataset Label, Class, Structure, Purpose, Keys
Variable Metadata	Variable Name, Variable Label, Type, Length, Controlled Terminology, Role
Value Level Metadata	Source variable, Value, Label, Type, Controlled Terminology
Controlled Terminology	Code Value, Code Label

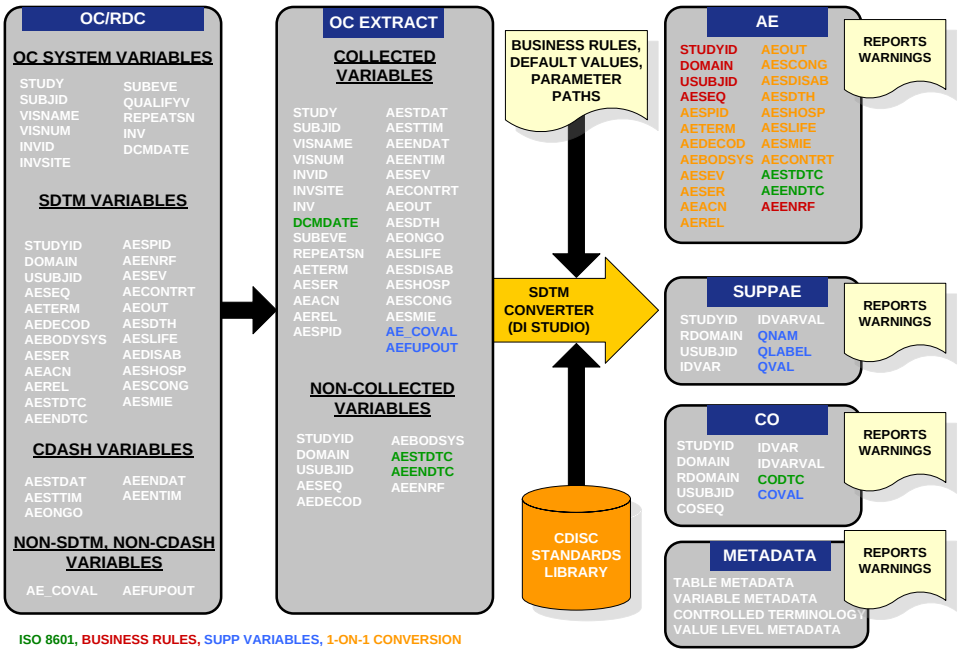
About 80% of the define.xml metadata is automatically produced by the system.

To generate 100% of the define.xml metadata, it is suggested to create a company library that captures Origin and Comments.

STEP 9: SDTM DATASETS

Based on the O*C extracts, the CDISC standards library, and the set of business rules and default values, the SDTM generator generates the SDTM domains. Figure #7 displays the complete data conversion flow from data entry in OC/RDC up to the automatic generation of the SDTM domains for AE (Adverse Events). The original Adverse Event data is split into a generic AE domain, a SUPPAE and CO domain.

Figure #7: SDTM data conversion flow



STEP 10: DATA PROFILING/WARNINGS REPORT

From the SDTM generator system, a data profiling and warnings report is generated.

The data profiling report (Figure #8) provides a list of variables with the corresponding label, type and length. It also displays, next to the distinct values, the mean, minimum, maximum and missing values.

The warning reports include basic checks for conformance with the CDISC domain models as specified in the 3.2.2. Conformance section of the CDISC SDTM Implementation Guide (Version 3.1.2).

Figure #8: AE data profiling report

DATAPATH.AE *Adverse Events*
Created: 18FEB10:15:37:52
obs : 384
Keys :

#	Variable	Label	Type	Length		Contents
1	STUDYID	Study Identifier	CHAR	20	+	1 distinct values, 0 missing values, all shown
2	DOMAIN	Domain Abbreviation	CHAR	2	+	1 distinct values, 0 missing values, all shown
3	USUBJID	Unique Subject Identifier	CHAR	20	+	172 distinct values, 0 missing values, 30 shown
4	AESEQ	Sequence Number	NUM	8	+	9 distinct values, 0 missing values, all shown Mean: 2.141 Minimum: 1 Maximum: 9 Missing values: 0
5	AETERM	Reported Term for the Adverse Event	CHAR	200	+	284 distinct values, 0 missing values, 30 shown
6	AEDECOD	Dictionary-Derived Term	CHAR	200	+	195 distinct values, 0 missing values, 30 shown
7	AEBODSYS	Body System or Organ Class	CHAR	200	+	23 distinct values, 0 missing values, all shown
8	AESEV	Severity/Intensity	CHAR	20	-	3 distinct values, 0 missing values, all shown MILD (241x) MODERATE (127x) SEVERE (16x)
9	AESER	Serious Event	CHAR	20	+	2 distinct values, 0 missing values, all shown
10	AEACN	Action Taken with Study Treatment	CHAR	40	+	6 distinct values, 0 missing values, all shown
11	AEREL	Causality	CHAR	20	+	2 distinct values, 0 missing values, all shown
12	AEOUT	Outcome of Adverse Event	CHAR	40	+	3 distinct values, 0 missing values, all shown
13	AESCONG	Congenital Anomaly or Birth Defect	CHAR	1	+	All values are missing
14	AESDISAB	Persist or Signif Disability/Incapacity	CHAR	1	+	All values are missing
15	AESDTH	Results in Death	CHAR	1	+	All values are missing
16	AESHOSP	Requires or Prolongs Hospitalization	CHAR	1	+	2 distinct values, 367 missing values, all shown
17	AESLIFE	Is Life Threatening	CHAR	1	+	2 distinct values, 382 missing values, all shown
18	AESMIE	Other Medically Important Serious Event	CHAR	1	+	2 distinct values, 383 missing values, all shown
19	AECONTRT	Concomitant or Additional Trtmnt Given	CHAR	1	+	2 distinct values, 0 missing values, all shown
20	AESTDTC	Start Date/Time of Adverse Event	CHAR	20	+	222 distinct values, 0 missing values, 30 shown
21	AEENDTC	End Date/Time of Adverse Event	CHAR	20	+	197 distinct values, 87 missing values, 30 shown
22	AESTDY	Study Day of Start of Adverse Event	NUM	8	+	128 distinct values, 9 missing values, 30 shown Mean: 29.784 Minimum: -72 Maximum: 109 Missing values: 9
23	AEENDY	Study Day of End of Adverse Event	NUM	8	+	124 distinct values, 89 missing values, 30 shown Mean: 48.275 Minimum: -43 Maximum: 218 Missing values: 89
24	AEENRF	End Relative to Reference Period	CHAR	20	+	2 distinct values, 297 missing values, all shown

These conformance checks are a subset of the complete set of CDISC SDTM compliance checks generated from the SDTM generator output. The in-house developed Data Model Compliance Checker (DMCC) can be linked into this system and used to ensure the entire set of SDTM data and metadata checks is submitted.

CONCLUSION

It is clear that upfront integration of the CDISC standards into the sponsor's OC/RDC clinical data management system offers many advantages and is considered to be the longer term solution. The advantages of one integrated system are application of identical nomenclature, use of a system wide data dictionary and reduction of data inconsistencies. In addition, it can easily drive the definition of the data collection tools in RDC and as such significantly reduce the time to clinical study set-up.

The use of CDASH and SDTM variables within OC/RDC should be maximized; including use of CDISC controlled terminology. Current alternatives are:

- programming SDTM requirements at OC/RDC view level; but this is difficult to automate
- back-end data conversions; but this is resource-intensive

Through smart set-up in OC/RDC, SDTM development is automated via an SDTM generator. No SDTM conversion is required in OC/RDC; but occurs outside this environment. This eventually results in real-time SDTM datasets and metadata, ready for analysis. SDTM conversion occurs at a <push of the button>.

REFERENCES

1 http://www.cdisc.org/pdf/eurointerchange2007/TheBusinessCaseforCDISCStandards-Kush.pdf

CONTACT INFORMATION

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

Peter Van Reusel
Business Unit Director
Business & Decision Life Sciences
Sint-Lambertusstraat 141 Rue Saint Lambert
1200 Brussels
Tel +32 2 774 11 00
Mobile +32 476 54 59 17
Fax +32 2 774 11 99
peter.vanreusel@businessdecision.com
www.businessdecision-lifesciences.com