Paper AD10

Data Capture to Trial-Specific Metadata for Analysis Data Sets: Automation Along the Way

Sathish Sundaram, Tata Consultancy Services, Indianapolis, IN
Ganesh Munuswamy, Tata Consultancy Services, Indianapolis, IN
Mominul Islam, Eli Lilly & Company, Indianapolis, IN

ABSTRACT

Development of analysis data sets and associated metadata that conform to industry (CDISC) standards is one of the critical challenges in preparation and analysis of clinical trial data intended for FDA submission. Data captured at clinical trial sites through case report forms (CRFs) and stored in databases are often subjected to a sequence of processing steps eventually resulting in analysis data sets (ADS). Metadata of the data sets and variables in the analysis data sets form a vital part of the information (of any study / trial) that is submitted to FDA. We outline a strategy (based on a SAS macro we developed) towards automating some of the steps involved in generation of trial-specific ADS metadata.

INTRODUCTION

There are several steps involved in creation of Tables, Figures and Listings (TFL) from data captured in a clinical trial. The captured data stored in a database, is often processed to generate SAS data sets (of the raw data) which are eventually converted to Analysis Data sets. ADS contain variables used in the process of TFL creation and a carefully developed and thoroughly documented ADS significantly reduces the time and efforts involved in TFL generation and review.

Documenting the ADS include clearly defining the attributes of the data sets and variables within the data sets. Companies use metadata to describe the ADS as well as to control processes involved in creating the analysis data sets and the down stream applications / codes that use the data sets towards TFL development.

Pharmaceutical companies are increasingly ensuring that the ADS and the associated metadata developed internally conform to operating standards (often modeled on CDISC standards) of the company. Well-defined operational standards both at the stage of data capture and at ADS level go a long way in ensuring that the data submitted to FDA conforms to industry standards and follows the regulatory authority's guidelines. However, considerable efforts go into generating these ADS and their metadata from the raw SAS data sets and descriptive documents containing trail-specific variable derivations.

While there are several ways of generating ADS from the captured data, we outline a strategy to efficiently generate the metadata for the analysis data sets. The generated metadata can then be used to drive the development of ADS themselves. In this paper we specifically focus on the metadata creation aspect of the analysis data set generation process.

We describe a specific example where data is captured electronically through Inform architecture and extracted onto ClinTrial database consisting of multiple panels. Captured data (conforming to internal data collection standards) was mapped to Lilly's operational ADS standards metadata based on study specific rules and derivations documented in an MS-Excel workbook. The mapping was instrumental in driving the creation of ADS metadata specific to the trial. A new SAS macro was developed to perform the mapping of the panel data to Lilly's operational ADS standards and generate metadata for the mapped data sets.

ADS METADATA STRUCTURE

Metadata of the operational standards describe the attributes of all the analysis data sets and variables used in clinical trials by the company. The metadata are themselves organized into distinct SAS data sets. For example, metadata of a horizontal standard analysis data set describing patient demographic information is organized into four SAS data sets:

1. DS metadata set: describes the attributes of the analysis data sets itself,
2. DS_Var metadata set: describes the various attributes of the variables such as the format,
3. DS_Var_Val metadata set: describes the valid set of values, associated with each variable in the data set
4. Cat_Desc: a SAS catalog giving detailed text description of the variable including its derivation in relation to the variables in the raw SAS data set of the CRF captured data (Figure 1).

Data Set: EVENTS Column: AEENDT

General Description

Event End Date

Detailed Description

AE1001 Panel (variables: AEENDTYY, AEENDTMO, AEENDTDD)

Figure 1: Example text from description catalog

TRIAL SPECIFIC OUTPUT

We have developed a macro that compares trial-specific information present in an MS-Excel Workbook to the company's operational ADS standards by searching through the descriptions catalog metadata (Cat_Desc). The SAS catalog metadata (especially the descriptive text) play a key role in generating trial-specific metadata. The macro repeats the process for every spreadsheet in the Workbook (and hence every ClinTrial panel) associated with the trial of interest and produces the following two distinct trial specific set of outputs.

1) Panel Map: An MS-Excel spreadsheet which maps the ClinTrial panels to the company's operational ADS standards and selects and displays only the analysis data sets needed for the specific clinical trial being analyzed.

2) Trial Specific Metadata: This includes the above-mentioned DS, DS_Var, DS_Var_Val and Cat_Desc metadata that is specific to the trial

Metadata for additional analysis variables that might be needed (based on the statistical analysis plan of the trial) and their derivations as well as additional trial specific descriptions of existing standard variables can further be incorporated into the above metadata set using simple SAS DATA and PROC steps if needed to create the final metadata for the trial-specific ADS.

Example:

We have used input ClinTrial panels (referred to as AE1001, SUM1001 etc.) containing data collected through CRFs. An MS-Excel work book containing trial-specific information of the captured data was used as the input for the SAS macro. The book contains a sheet for each ClinTrial Panel with variable lists and is easy to construct with the information required for the macro.

Code for reading multiple sheets of the MS-Excel workbook containing ClinTrial panel information

```
*  Accessing the MS-Excel WorkBook*;
LibName xls excel "C:\Documents and
Settings\Desktop\Today\Pharma\Study_ABC_Radical.xls" ;

* Accessing the ADS Standards library*;
libname catlib 'C:\Documents and
Settings\Desktop\Today\Pharma\ADS_STD';


%macro appds;
proc sql noprint;

      * Creating a dataset which contains the list of
panels*;
      create table pnl_ds as
      select memname from dictionary.tables
      where libname="XLS" and
substr(memname,length(memname),1)="$"
      and substr(memname,length(memname)-3,2)="00";

      *creating a sequence of macro variables with dataset
names*;
      select memname,count(*) into :table1-:table99,:totcnt
from pnl_ds;

quit;
```

Code for searching the catalog descriptions of ADS standard metadata

```
%macro cat_srch;

proc sql;
      select table,count(*) into :ads_lst1-
:ads_lst99,:tblcnt from ads_lst;
      select tbnm,count(*) into :pnl1-:pnl199,:pnlcnt from
pnl_lst;
quit;


%do i = 1 %to &tblcnt;
      filename _secat catalog
catlib.descriptions.&&ads_lst&i...source";
            filename _secat list;

            %do j = 1 %to &pnlcnt;

                  data catx&i&j;
                  infile _secat length=len;
                  input record $varying200. len;
                  length adsnm pnlnm $50;
                    adsnm = "&&ads_lst&i";
                    pnlnm = "&&pnl&j";
                  if
index(upcase(record),upcase("&&pnl&j")) > 0  then output;
```

```
                        run;

                    %if &i = 1 and &j =1 %then %do;
                       data catbase;
                                set catx&i&j;
                     run;
                    %end;
                    %else %do;
                          proc append base=catbase data=catx&i&j;
                          run;

                          proc sql;
                                drop table catx&i&j;
                          quit;
                    %end;
              %end;
        %end;

        %mend cat_srch;
```

Panel Map Output Sample

ADS Name	ClinTrial Panel Name	ADS Variable	Panel Variable List
DISPOSIT	SUM1001	DS	RSDCFIN
DISPOSIT	SUM1001	DSDT	DISCDT,DISCDTDD,DISCDTMO,DISCDTYY
DISPOSIT	SUM1001	DSDTC	DISCDT,DISCDTDD,DISCDTMO,DISCDTYY
DISPOSIT	SUM1001	DSDTTM	DISCDT,DISCDTDD,DISCDTMO,DISCDTYY
DISPOSIT	SUM1001	DSDTTMC	DISCDT,DISCDTDD,DISCDTMO,DISCDTYY,DTHDT
DISPOSIT	SUM1001	DSLNM	RSDCFIN
DISPOSIT	SUM1001	DSSNM	RSDCFIN
DISPOSIT	SUM1001	DTHDT	DTHDT,DTHDTDD,DTHDTMO,DTHDTYY
DISPOSIT	SUM1001	DTHDTC	DTHDT,DTHDTDD,DTHDTMO,DTHDTYY
DISPOSIT	SUM1001	DTHDTTM	DTHDT,DTHDTDD,DTHDTMO,DTHDTYY
DISPOSIT	SUM1001	DTHDTTMC	DTHDT,DTHDTDD,DTHDTMO,DTHDTYY
DISPOSIT	SUM1001	PRICOD	PRICOD
DISPOSIT	SUM1001	PRICODLNM	PRICOD
DISPOSIT	SUM1001	PRICODSNM	PRICOD
EVENTS	AE1001	AEDISCFLG	AEID
EVENTS	AE1001	AEDUR	AEENDT,AESTDT
EVENTS	AE1001	AEENDT	AEENDT,AEENDTDD,AEENDTMO,AEENDTYY
EVENTS	AE1001	AEENDTC	AEENDT,AEENDTDD,AEENDTMO,AEENDTYY
EVENTS	AE1001	AEENDTTM	AEENDT,AEENDTDD,AEENDTHR,AEENDTMI,AEENDTMO,AEENDTYY
EVENTS	AE1001	AEENDTTMC	AEENDT,AEENDTDD,AEENDTHR,AEENDTMI,AEENDTMO,AEENDTYY
EVENTS	AE1001	AEFLG	AEFLAG

Figure 2: Panel Map generated by the SAS macro discussed above.

Code for generating trial-specific metadata from the standards metadata

```
proc sql;

create table out_table as
select * from catlib.tables
where table in ( select distinct adsnm  from colmnbase4);
```

```
create table out_columns as
select * from catlib.columns
where column in (select distinct adscolnm from colmnbase4)
and table in (select distinct adsnm from colmnbase4);

create table out_values as
select * from catlib.values
where format in ( select distinct cformat from out_columns
where cformatflag > 1);

quit;
```

Sample Metadata generated by the macro

VIEWTABLE: Work.Out_table

	table	tshort	tlabel	torder	type	tdescription
1	DISPOSIT		Disposition	.	TABLE	DISPOSIT
2	EVENTS		Events	.	TABLE	EVENTS

VIEWTABLE: Work.Out_columns

	table	column	cshort	cpkey	corder	clabel
1	DISPOSIT	DS		.	.	Subject Disposition
2	DISPOSIT	DSDT		.	.	Subject Disposition Date
3	DISPOSIT	DSDTC		.	.	Subject Disposition Date Character

VIEWTABLE: Work.Out_values

	format	start	end	flabel	flab
1	DICTTPE	Event			
2	DICTTWO	CTCAE			
3	MAESVC				
4	MAESVC	1		Mld	Mld
5	MAESVC	2		MOD	Moderate
6	MAESVC	3		SEV	Severe
7	MAESVC	4		MORSEV	More Severe than baseline
8	MASRA				
9	MASRA	1		Self	Self (e.g., Patient)
10	MASRA	10		Social Worker	Social Worker
11	MASRA	11		AttendMD	Attending Physician
12	MASRA	12		Nurse	Nurse
13	MASRA	13		SponPharmMD	Sponsor Pharmacovigilance MD
14	MASRA	14		Investigator	Investigator
15	MASRA	15		HCProf	Health Care Professional
16	MASRA	16		Interpreter	Interpreter
17	MASRA	17		SponCRP	Sponsor Clinical Research Physician
18	MASRA	18		ClinPharm	Clinical Pharmacologist
19	MASRA	19		Rater	Rater
20	MASRA	2		Parent	Parent

Figure 3: Trial Specific metadtasets generated from standards

Figure 3 (superimposed on one another for ease of display in this paper) shows the three metadatasets, DS (WorkOut_table), DS_Var (WorkOut_columns), DS_Var_Val(WorkOut_values) discussed above.

CONCLUSION

The above macro and the strategy provides several advantages including considerably reducing the time involved in generating the metadata for ADS from captured (through CRFs) data by eliminating several intermediate processes and SAS programs that are typically involved in comparing the captured data to existing standards.

If the operational standards are already modeled on CDISC standards (as is the current norm in the industry) the above approach would ensure that the trial specific metadata and ADS generated also conform to the industry standards. The DS, DS_Var and DS_Var_Val metadata sets can be further used to drive the ADS creation based on the descriptions present in the Cat_Desc metadata. These metadata sets collectively contain considerable information about the analysis variables and hence can also be used to develop Define Document and or other regulatory submission materials

REFERENCES

SAS Institute (2002-2003) Software: Reference, Version 9.1.3
Copyright 2002-2003 by SAS Institute Inc., Cary, NC, USA

ACKNOWLEDGMENTS

The authors would like to thank Mr. Maruful Chowdhury, Team Lead, Neurosciences, Statistics Modernization & Reporting Team, Eli Lilly & Company for his technical insights and review.

Contact Information

Sathish Sundaram
Tata Consultancy Services
Lilly Corporate Center
Indianapolis, In 46285
sundaramsa@lilly.com

Ganesh Munuswamy
Tata Consultancy Services
Lilly Corporate Center
Indianapolis, In 46285
munuswamyga@lilly.com

Mominul Islam
Lilly Corporate Center
Indianapolis, In 46285
IslamMo@lilly.com