Paper CC-018

Exploring SAS® PROC CDISC Model=ODM and Its Undocumented Parameters

Elena Valkanova, Biostat International, Inc, Tampa, FL
Irene Droll, XClinical GmbH, München, Germany

ABSTRACT

The CDISC Operational Data Model (ODM) is a well-established data standard amongst other known CDISC standards used within the clinical research industry. The format allows the exchange of electronic data from multiple sources, such as case report forms, electronic patient diaries, or administrative data. One use case to bring data into a different system is the mapping of ODM structured data into SAS® datasets. There are three major components that are integrated in the CDISC ODM standard: a) study metadata (i.e. CRF elements and database definition), b) snapshot data (i.e. clinical data), and c) transactional data (i.e. audit trail information). The ODM data structure is based on an XML (eXtensible Markup Language) schema to ensure consistency in how the data is represented and validated by XML applications.

In this paper we concentrate on the import of an external ODM XML file - coming from an Electronic Data Capture (EDC) system - into SAS, using the latest version of the SAS procedure PROC CDISC (SAS release 8.2 and later). The main objective is to explain undocumented parameters/options and limitations of SAS PROC CDISC Model=ODM. Additionally, we list specifications for the external XML file that need to be considered before using PROC CDISC. The listed checks can be crucial to the correct and complete mapping of data into SAS datasets. A SAS macro is presented to facilitate the ODM XML data import.

Note: Basic understanding of the CDISC ODM structure of Events, Forms, ItemGroups, Items, Codelists and MeasurementUnits is required.

1 INTRODUCTION

XML is used for data transfer between dissimilar platforms. The SAS procedure PROC CDISC has been used very extensively over the last years in the clinical research industry for the purpose of validating SAS datasets in either the ODM or SDTM standard. Additionally, PROC CDISC offers the interface to import data coming from an ODM compliant XML [1] file into SAS. This application of PROC CDISC is still new to many SAS users.

PROC CDISC allows more user-control on the metadata content than the SAS XML libname engine. The syntax of the procedure permits import of administrative and study dependent data via statement parameters. We include a list of these parameters and an example in Section 2 that will show more options for the data import in SAS.

SAS users who start to use PROC CDISC to import an ODM compliant XML file into SAS often come across the same obstacles. Most of these challenges can be solved by the following points that we will cover in this paper:

- ODM model and SAS datasets
- Undocumented parameters of PROC CDISC and applications
- Limitations of the procedure that can be taken into account before a file is imported in SAS

2 ODM MODEL AND SAS DATASETS

The ODM compliant XML file that will be imported with PROC CDISC contains the collected clinical subject data as well as the description of the study-level metadata (where all used Meta Data Versions are merged into one Metadata definition). By using PROC CDISC Model=ODM the hierarchical structure of the ODM data tree file is converted into the tabular structure of SAS datasets. Each *ItemGroup* definition (a group of items with related information) in the ODM is mapped into one SAS dataset.

To retain control over SAS Dataset and SAS Variable names it is important to fill in the attributes *SASDatasetName* on ItemGroup definition level, *SASFieldName* on Item definition level, and *SASFormatName* on Codelist definition level. PROC CDISC will output the generated SAS datasets, columns and formats according to the names defined in the ODM Metadata.

If the SAS name attributes are not defined in the ODM metadata definition, SAS PROC CDISC will use the first eight characters of the respective element's ODM Name. In case these are missing as well, SAS PROC CDISC will take the first eight characters of the respective element's OID (Object Instance Identifier) [2].

The attribute OID uniquely identifies elements and mainly is used for cross reference within an ODM or between multiple ODM files. Nevertheless, the truncation to the first eight characters can lead to non-unique SAS dataset, variable or format names.

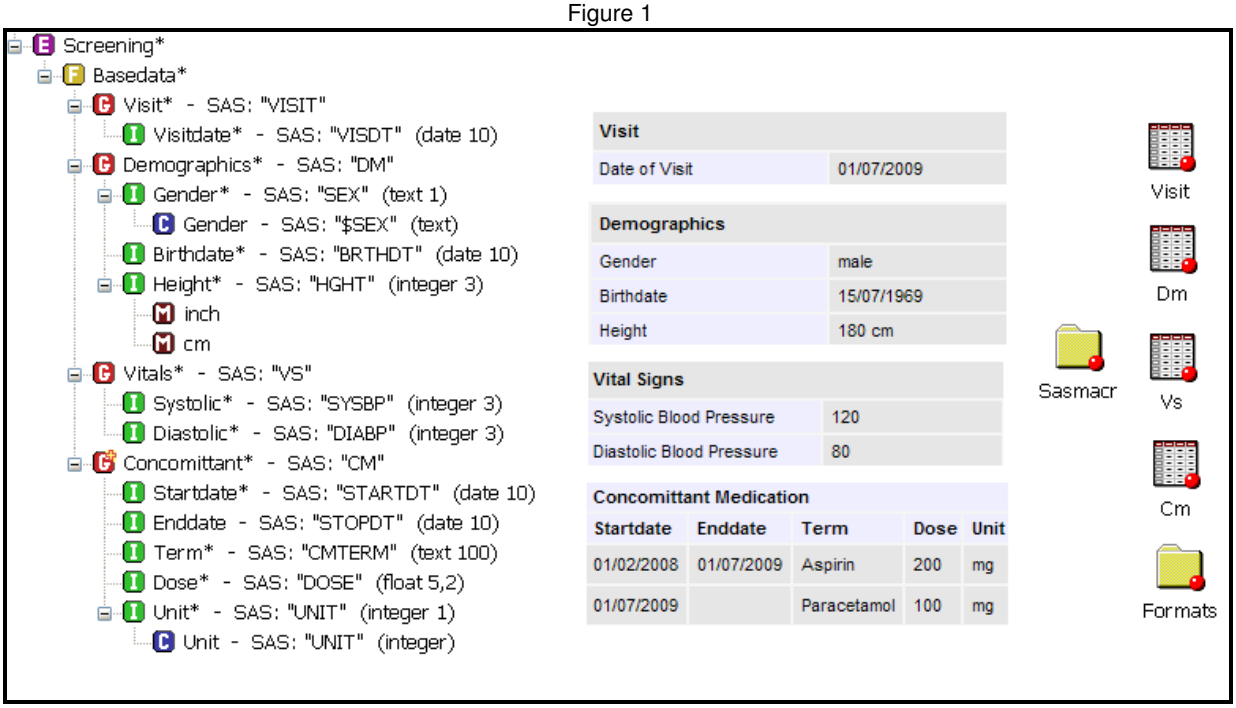
For each ItemGroup defined in the ODM file a SAS dataset with a unique SASDatasetName is generated. To reduce the number of SAS datasets and thus the number of extra merge steps needed after the import the following tips are suggested:

- Using repeating ItemGroups if possible (e.g. for CM)
- Re-using ItemGroups in several visits if possible (e.g. for VS)
- More Items within one ItemGroup if possible and reasonable (e.g. AE)

The goal is to keep all repeating data over visits (Events) and all data Items of related information in one SAS dataset.

Figure 1 is a graphical representation of an example ODM tree, the corresponding EDC screenshot display for data entry, and the corresponding list of SAS datasets. The SAS datasets are created by importing an ODM compliant XML file into SAS by using PROC CDISC.

Figure 1



3 PROC CDISC

3.1 TECHNICAL SPECIFICATIONS

- Operating environments: Windows, UNIX and z\OS.
- SAS/Base –8.2 or later.
- Installation of latest PROC CDISC version 2.15.62. (Check for the version installed with:
  `PROC CDISC version; RUN;`

3.2 PARAMETERS FOR IMPORTING ODM XML DATA WITH PROC CDISC

Each ODM element contains an OID attribute that uniquely identifies the element within the XML structure. OIDs that result in SAS dataset columns/fields are also called keyset fields. Most of the keyset fields used to identify the exact data line of the SAS dataset in the ODM data file are: StudyOID, MetaDataVersionOID, SubjectKey, StudyEventOID, StudyEventRepeatKey, FormOID, FormRepeatKey, ItemGroupOID, ItemGroupRepeatKey and TransactionType (can only have one value: "insert").

Table 1 explains all statements and parameters that need to be or can be used for the import of an ODM XML file into SAS:

Table 1

PROC CDISC	Parameters	Values	Description
	MODEL*	ODM/ SDTM	Name of supported CDISC model.
	READ*	FILENAME	The location of the source file.
	FORMATACTIVE	YES/NO	If Yes, ODM CodeList elements are converted to SAS formats (Creation of SAS format catalogue) and assigned to the respective SAS variables. The default value is NO.
	FORMATNOREPLACE	YES/NO	If Yes, formats with the same name are not replaced. The default value is NO. This option can only be used if FORMATACTIVE = YES.
	FORMATLIBRARY	LIBRARY	To store a permanent Format library. This option can only be used if FORMATACTIVE = YES.
	LANGUAGE+	EN, DE, etc.	To specify the language of the labels, if there is more than one language in the study. The default is EN.
ODM*	ODMVERSION*	1.2	Version of the ODM model.
ODM	ODMMINIMUMKEYSET	YES/NO	If Yes, only SubjectKey is included in the SAS dataset; If No, all keyset fields (as described above) are included in the SAS dataset when importing ODM. As soon as an ItemGroup is re-used or a repeating ItemGroup is defined this parameter needs to be set to NO. The default value is NO.
ODM	ODMMAXIMUMOIDLENGTH	NUMBER	If set, the OID value will be truncated according to the specified number. If not set, the OID value can have up to 100 characters.
ODM	USENAMEASLABEL+	YES/NO	If Yes, The ODM Name attribute of the item is used as a label for the SAS column/ field. The default value is NO.
ODM	LONGNAMES +	YES/NO	If Yes, the ODM Name attributes are used rather than the ODM SAS attributes. The restriction is 32 characters; blanks will be replaced with an underscore (_). The default value is NO.

PROC CDISC	Parameters	Values	Description
CLINICALDATA*	OUT	TEXT	Name of the resulting SAS dataset.
CLINICALDATA	SASDATASETNAME	TEXT	This name matches the ODM SASDatasetName.
	IN	LIBRARY	Specifies the library and member name for the resulting SAS dataset.
CLINICALDATA	SITEREF +	YES/NO	If Yes, a new field “LocationOID” is added to the corresponding SAS dataset containing the site info.
CLINICALDATA	INVESTIGATORREF +	YES/NO	If Yes, a new field “UserOID” is added to the corresponding SAS dataset containing the investigator info.

Note: + undocumented parameter, * required parameter

3.3 EXAMPLE

Below is an example using PROC CDISC with some of the parameters listed in Table 1 to read in an ODM file.

Figure 2

```
1     LIBNAME OUT 'C:\MYPROJECT\DATA\SAS';
2     FILENAME XMLIMP 'C:\MYPROJECT\DATA\XML\MYFILE.XML';

3       PROC CDISC MODEL=ODM
                   READ=XMLIMP
                   FORMATACTIVE=YES
                   FORMATNOREPLACE=NO
                   LANGUAGE="EN";

4       ODM        ODMVERSION="1.2"
                   ODMMAXIMUMOIDLENGTH=20
                   ODMMINIMUMKEYSET=NO

                   USENAMEASLABEL=YES;

5       CLINICALDATA OUT=OUT.DSET

                     SASDATASETNAME = "DSET"
                     SITEREF=YES

                     INVESTIGATORREF=YES;
        RUN;
FILENAME XMLIMP;
```

To use the English decode as the variable value.

To use the item attribute Name instead of the attribute SASFieldName as SAS variable name

Add an extra variable “LocationOID” to the output SAS data set.

Add an extra variable “UserOID” to the output SAS data set.

1. Specify the libname for the SAS dataset.
2. The FILENAME statement assigns the ODM file reference to the physical location of the ODM XML file.
3. The READ parameter in PROC CDISC MODEL=ODM is pointing to the defined filename reference.

4. ODM parameters are included such that all keyset fields will be imported in the SAS dataset DSET, the maximum length for each of the key fields is specified to 20. If the parameter `ODMMAXIMUMOIDLENGTH` is not specified, the key fields are output with its default length of 100.
5. CLINICALDATA parameters are included in this example so that a SAS dataset called DSET is created as the output SAS Dataset. This SAS Dataset contains two additional fields for recruitment sites and investigator info. The two parameters `SITEREF` and `INVESTIGATORREF` are optional.

4 REQUIREMENTS FOR THE EXTERNAL ODM FILE

The conversion of an XML ODM data file into SAS datasets by using PROC CDISC requires a few checks in order to ensure completeness and correctness of the mapping of ODM ItemGroups into corresponding SAS datasets. The checks may be classified into three groups:

- ODM code snippets that produce error messages in the SAS log.
- ODM code snippets that produce warning messages in the SAS log.
- Other code considerations that may lead to incomplete mapping into the SAS dataset.

If an Error occurred, the resulting SAS dataset may be incomplete. Examples of common errors or warnings with useful descriptions are given in Table 2 and Table 3 below.

Table 2

Error message in SAS Log	Reason
`ERROR: Invalid member name for file WORK.CON.DATA`	SAS attribute names that are used by the operation system like AUX, CON, NUL, PRN, LPT1 - LPT9, and COM1 - COM9 cannot be used.
`ERROR: ItemGroupDef OID=ig.PSA has no matching ItemGroupRef.`	Unused elements, i.e. element definitions that are not referenced, need to be removed from the ODM metadata definition file.
`ERROR: ItemData ItemOID = "i.dummy". Corresponding ItemDef not found.`	SAS PROC CDISC expects explicit closing tags as opposed to implicit closing tags in the ODM XML file. For example, <ItemGroupName OID=”ig.test”/> will produce an error message and incomplete dataset. The correct syntax for PROC CDISC is <ItemGroupName OID=”ig.test”></ItemGroupName>. This may be achieved with a general XSLT conversion of the ODM data file.
`ERROR: ItemDef "Birthdate" has an invalid Length attribute value.` `ERROR: 1 ItemDef elements had incorrect Length or SignificantDigits attributes`	Date Items need to be defined with Length = 10 in the ODM File. Text Items need to be defined with Length = 200 in the ODM File. Please be aware that a text Item can have max. 200 characters for import into SAS. Time Items need to be defined with Length = 5 in the ODM File.
`ERROR: Some code points did not transcode.`	Characters typed in from a foreign key board can lead to errors, e.g. Cyrillic C.

Table 3

Warning message in SAS Log	Reason
`WARNING: Variable DATE already exists on file WORK.CM`	SASFieldNames need to be unique within one ItemGroup/ SASDataSet. If not, the column will be replaced so that one of the two Items with the same name will not appear in the dataset.
`WARNING: SAS character format names must begin with a $. SEX changed to SAS compliant format name $SEX`	If using text values instead of integer values for codes the SASFormatName defined in the ODM file should start with $ (e.g. $SEX). Otherwise SAS will add $ and give a warning message.
`WARNING: ItemRef OID="Height" OrderNumber="4" outside range. Ignoring OrderNumber.`	If an order number attribute of an Item within an ItemGroup exceeds the maximum number of Items within this ItemGroup SAS will produce a warning message.

Other errors that may lead to an incomplete mapping into the SAS dataset are related to:

- All SAS attribute names can have maximum eight characters and need to start with a character or with an underscore (no numbers). If the SAS name starts with a number, PROC CDISC will add an underscore (_) in front of the name and cut 1 character at the end. (e.g. 8DATE becomes _8DAT).

- If the SAS name has more than eight characters it will be truncated to eight characters. This holds the risk that names are cut and double SAS names occur on dataset level. If the value of the attribute SASDatasetName(s) is the same for different ItemGroup(s) in the ODM file the existing dataset will be overwritten and only one SAS dataset will be created.

- PROC CDSIC does not read the measurement units from the ODM file. There are no errors/messages in the SAS log file indicating that the measurement units have been ignored. By using an XSLT transformation on the ODM XML file a new mapping can be defined so that the measurement unit (e.g cm., ft.) is used as an additional ODM Item that can be imported into the SAS dataset.

5 IMPORT THE COMPLETE ODM DATA FILE INTO SAS

One SAS dataset is created after one invocation of PROC CDISC for one ItemGroup defined in the ODM file. Often the task is to provide a method for reading in an ODM file containing many ItemGroup Definitions resulting into multiple SAS datasets. To achieve this, we used the resource DICTIONARY.TABLES available in PROC SQL, PROC CDISC and SAS Macro in order to get at once all the data from the external ODM file.

Figure 3

```
LIBNAME OUT 'C:\TEST';
FILENAME XMLIMP 'C:\EXAMPLES\TEST.XML';
LIBNAME XMLIMP XML XMLTYPE=CDISCODM;

%MACRO ALLSETS (LIB);
PROC SQL NOPRINT;
SELECT UNIQUE MEMNAME INTO :DSETS SEPARATED BY '|'
  FROM DICTIONARY.TABLES
    WHERE UPCASE(LIBNAME)="&LIB" ;
       %PUT DATASETS: &DSETS;
QUIT;
%LET NUM=1;
%LET DSET=%SCAN(%QUOTE(&DSETS),&NUM,|);
   %DO %UNTIL (%QUOTE(&DSET)=%STR());
```

```
      PROC CDISC MODEL=ODM
                 READ=XMLIMP
                  FORMATACTIVE=YES
                 FORMATNOREPLACE=NO
                  LANGUAGE="EN";
       ODM        ODMVERSION="1.2"
                  ODMMAXIMUMOIDLENGTH=20
                  ODMMINIMUMKEYSET=NO
                   USENAMEASLABEL=YES;
       CLINICALDATA OUT=OUT.&DSET
                    SASDATASETNAME = "&DSET"
                     SITEREF=YES
                     INVESTIGATORREF=YES;
        RUN;
       %LET NUM=%EVAL(&NUM+1);
       %LET DSET=%SCAN(%QUOTE(&DSETS),&NUM,|);
   %END;
%MEND ALLSETS;
% ALLSETS(XMLIMP);
```

CONCLUSION

SAS PROC CDISC is a very useful tool that allows the data transfer from XML to SAS datasets. Many parameters for PROC CDISC are undocumented. For new users of PROC CDISC it can be time consuming to get familiar with these parameters and limitations of the procedure. In this paper we share our experience with PROC CDISC's import application. We proposed some solutions and preliminary checking on the external ODM XML file that will facilitate the data import into SAS.

REFERENCES

[1] Specification for the Operational Data Model (ODM), CDISC
http://www.cdisc.org/models/odm/v1.2/ODM1-2-0.html.

[2] "The CDISC Procedure for SAS® Software, Release 8.2 and Later", SAS Institute, Inc.
http://support.sas.com/rnd/base/xmlengine/proccdisc/TW8774.pdf

[3] http://www.w3.org/TR/REC-xml/#wf-entities.

CONTACT INFORMATION

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

Elena Valkanova
Biostat International, Inc.
14506A University Point Place
Tampa, FL, 33613
E-mail: evalkanova@biostatinc.com
Web: http://www.biostatinc.com

Irene Droll
XClinical GmbH
Siegfriedstrasse 8
80333 München, Germany
E-mail: id@xclinical.com
Web: http://www.xclinical.com