PharmaSUG2010 - Paper PO07

Exploring SAS® PROC CDISC Model=ODM and Its Limitations

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. ODM is a platform-independent format that allows an exchange and archive of clinical trial electronic data from multiple sources, such as case report forms (CRFs), electronic patient diaries, or administrative data. 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);
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. One important application of the procedure PROC CDISC is the model ODM along with a combination of statement parameters that allows the user to read the ODM XML data into SAS® datasets.

In this paper we concentrate on the import of an external ODM XML file - exported from an Electronic Data Capture (EDC) system into SAS, by using the latest version of the SAS procedure PROC CDISC (SAS release 8.2 and later). The main objective is to list and explain parameters/options of the procedure and emphasize on some of the limitations of SAS PROC CDISC Model=ODM that may be crucial for the data completeness and accuracy. Furthermore, we include some specifications for the external XML file that need to be considered before using PROC CDISC. In section 5, we included helpful examples using SAS Macro language and PROC SQL for reading an external ODM XML data file into SAS datasets.

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

1 INTRODUCTION

XML is *the* industry standard for data transfer between dissimilar platforms. The SAS procedure PROC CDISC has been used extensively over the last few years in the clinical research industry for the purpose of validating SAS datasets in either the ODM or SDTM standard. The SAS Clinical Standards Toolkit (CST) extends the possibilities to work with the SDTM standard. For the ODM standard PROC CDISC is still the preferred way to validate ODM datasets and to import data from an ODM compliant XML file into SAS [1].

There are two ways to read an XML file into SAS either using the SAS XML libname engine or using PROC CDISC. We focus on using PROC CDISC since it 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. The user has the ability to specify different combinations of parameters in order to add more information from the ODM file into the resulting SAS datasets. This information may vary based on the analysis specifications for a clinical study. A list of these parameters is included in Section 2 as well as an example with commonly used options for the data import into SAS.

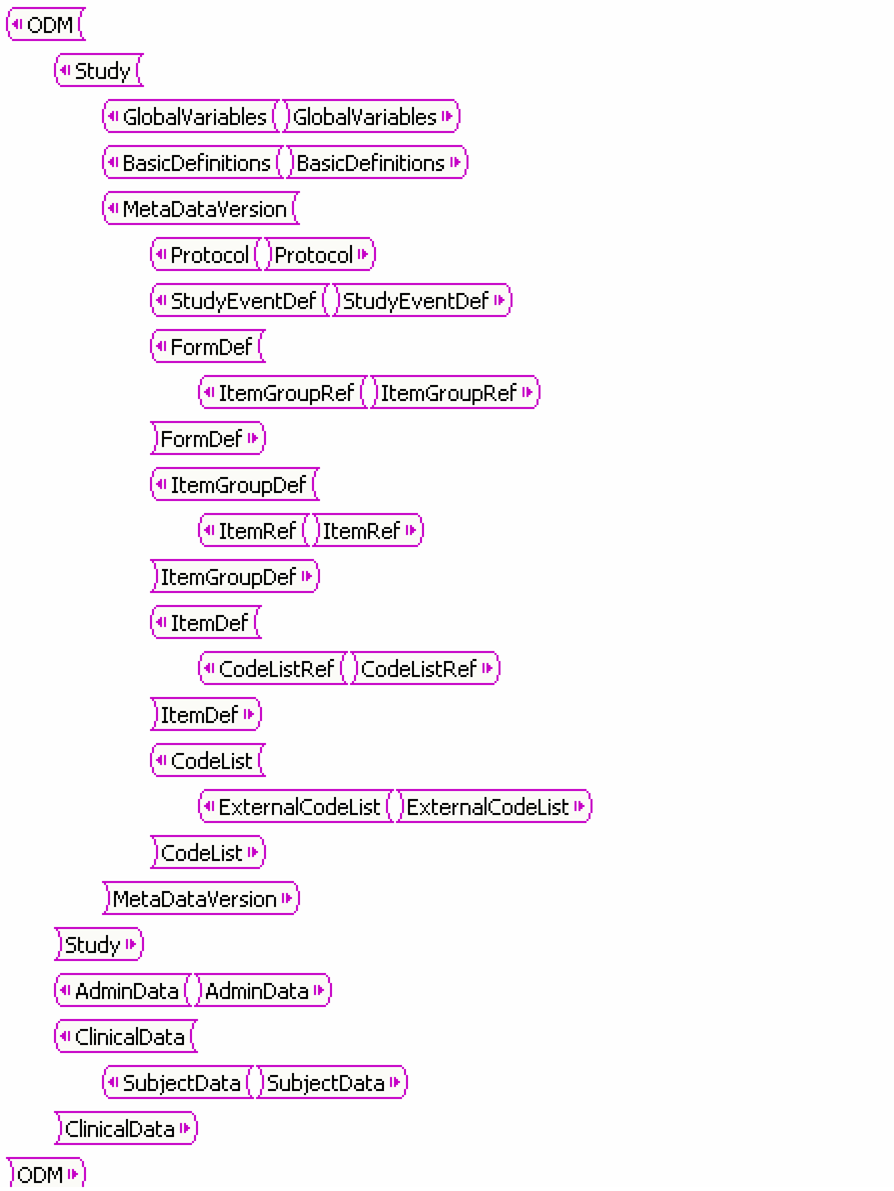
PROC CDISC beginners may come across different obstacles and need to be familiar with some common errors and warnings that are generated with the import of an ODM compliant XML file into SAS. Some of these challenges can be overcome with the suggestions included in our paper.
The paper is organized in four sections:

- ODM Model and SAS datasets
- PROC CDISC and statement parameters
- The external ODM XML file
- Import of the complete ODM data file into SAS

2 ODM MODEL AND SAS DATASETS

The ODM compliant XML file that will be imported into SAS with PROC CDISC contains the collected clinical subject data as well as the description of the study-level metadata. If amendments were applied during the data collection it is required that in the data export file different metadata 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 in the XML file (a group of items with related information) is mapped into one SAS dataset.

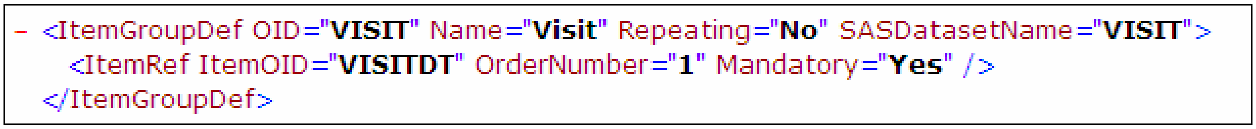
Figure 1



To retain control over the SAS Dataset, SAS Variable and SAS Format names it is important to include the *optional* attributes in the XML ODM metadata part:

- *SASDatasetName* on ItemGroup definition level;

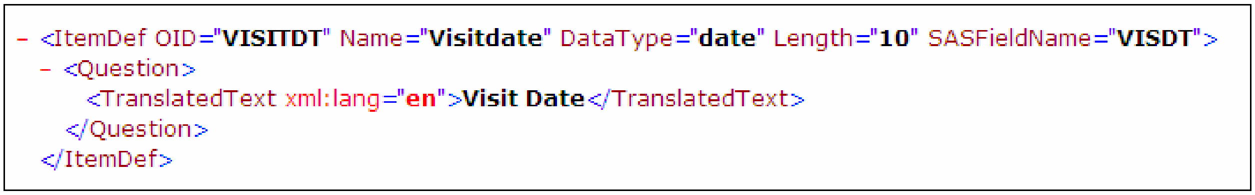
Figure 2



```
- <ItemGroupDef OID="VISIT" Name="Visit" Repeating="No" SASDatasetName="VISIT">
    <ItemRef ItemOID="VISITDT" OrderNumber="1" Mandatory="Yes" />
  </ItemGroupDef>
```

- *SASFieldName* on Item definition level;

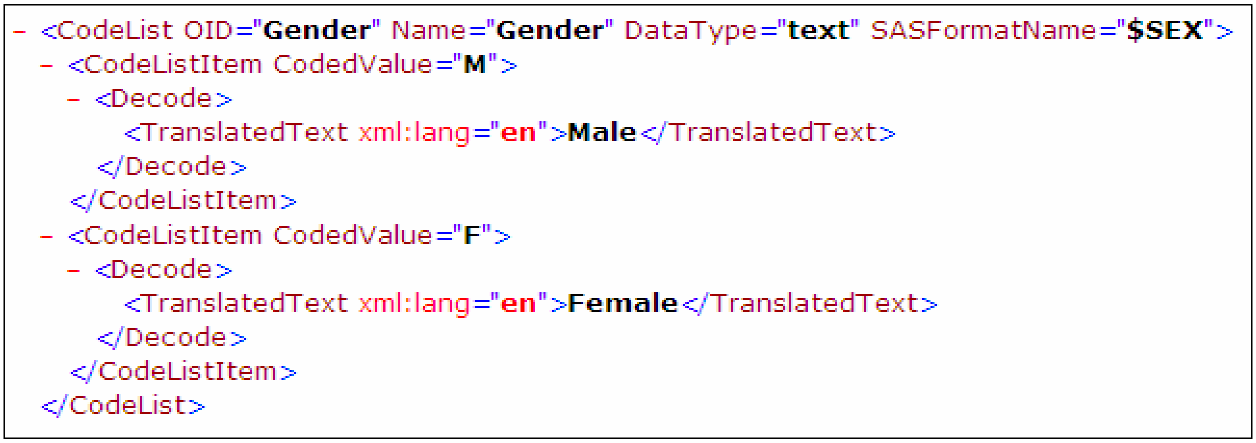
Figure 3



```
- <ItemDef OID="VISITDT" Name="Visitdate" DataType="date" Length="10" SASFieldName="VISDT">
  - <Question>
      <TranslatedText xml:lang="en">Visit Date</TranslatedText>
    </Question>
  </ItemDef>
```

- *SASFormatName* on Codelist definition level.

Figure 4



```
- <CodeList OID="Gender" Name="Gender" DataType="text" SASFormatName="$SEX">
  - <CodeListItem CodedValue="M">
    - <Decode>
        <TranslatedText xml:lang="en">Male</TranslatedText>
      </Decode>
    </CodeListItem>
  - <CodeListItem CodedValue="F">
    - <Decode>
        <TranslatedText xml:lang="en">Female</TranslatedText>
      </Decode>
    </CodeListItem>
  </CodeList>
```

PROC CDISC will output the generated SAS datasets, columns and formats according to the attributes defined in the ODM metadata. We recommend including the optional parameters in the ODM file to facilitate the SDTM standard preparation.

If the optional attributes are not defined in the ODM metadata definition, SAS PROC CDISC will use the first eight characters of the respective element's ODM attribute *Name*. If the attribute *Name* is missing as well, SAS PROC CDISC will take the first eight letters of the respective element's *OID* (Object Instance Identifier) [2]. The attribute OID uniquely identifies elements and is mainly used for cross reference within an ODM file or between multiple ODM files. Because only the first eight characters are taken, this can lead to a non-unique SAS dataset, variable or format names, and will result in ERROR messages or overwriting of datasets, variables and formats when reading the file.

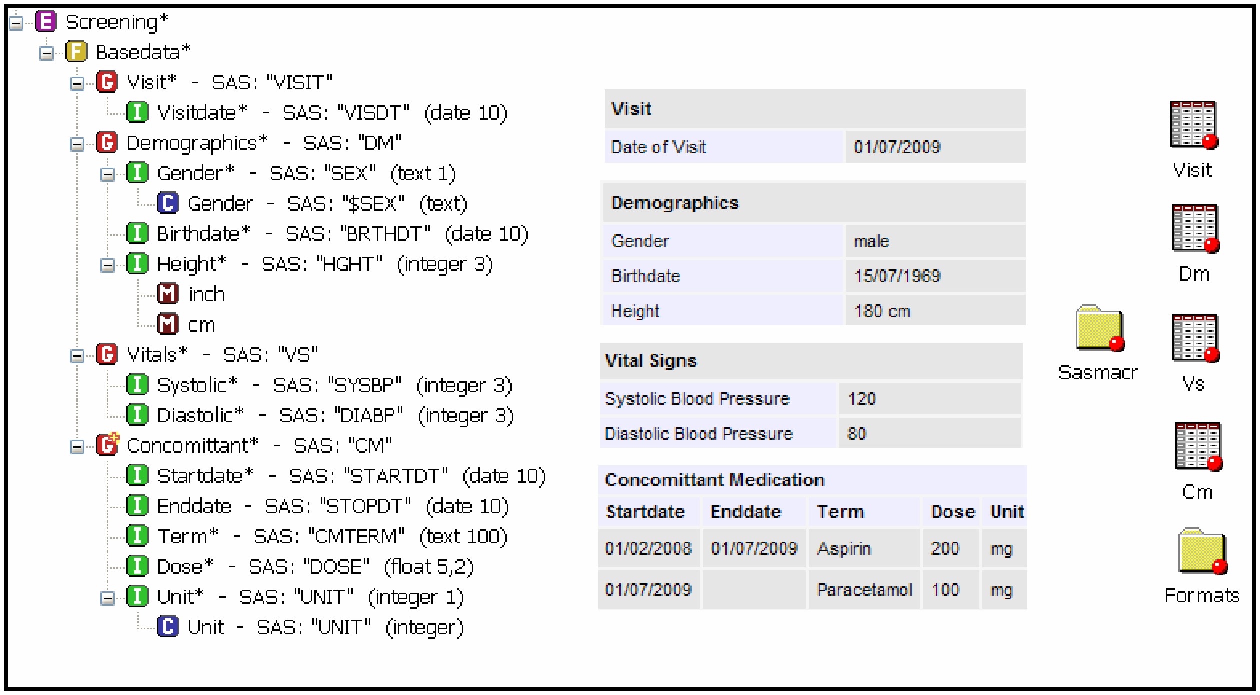
For each *ItemGroup* defined in the ODM file a SAS dataset with a unique *SASDatasetName* is generated. To reduce the number of SAS datasets and thus the number of extra merge steps after the import the following tips are suggested:

- Use repeating *ItemGroups,* if possible – (e.g. CM)
- Re-use *ItemGroups* in different visits – consistency in the name convention (e.g. VS)
- More Items within one *ItemGroup* is preferable, if possible and reasonable (e.g. AE)

The goal is to reuse *ItemGroups* throughout visits (Events) and data *Items*, of related information in one SAS dataset. These are preliminary steps related to the design of the ODM XML file and they are not required but it will reduce the amount of effort when the datasets need to be prepared for submission.

Figure 5 is a graphical representation consisting of three parts. The left part is an ODM tree showing the metadata design of a study, the middle part shows the corresponding EDC screenshot from the WUI (Web User Interface), and the right part shows the corresponding SAS datasets.

Figure 5



E Screening*
F Basedata*
G Visit* - SAS: "VISIT"
I Visitdate* - SAS: "VISDT" (date 10)
G Demographics* - SAS: "DM"
I Gender* - SAS: "SEX" (text 1)
C Gender - SAS: "$SEX" (text)
I Birthdate* - SAS: "BRTHDT" (date 10)
I Height* - SAS: "HGHT" (integer 3)
M inch
M cm
G Vitals* - SAS: "VS"
I Systolic* - SAS: "SYSBP" (integer 3)
I Diastolic* - SAS: "DIABP" (integer 3)
G Concomittant* - SAS: "CM"
I Startdate* - SAS: "STARTDT" (date 10)
I Enddate - SAS: "STOPDT" (date 10)
I Term* - SAS: "CMTERM" (text 100)
I Dose* - SAS: "DOSE" (float 5,2)
I Unit* - SAS: "UNIT" (integer 1)
C Unit - SAS: "UNIT" (integer)

Visit	
Date of Visit	01/07/2009

Demographics	
Gender	male
Birthdate	15/07/1969
Height	180 cm

Vital Signs	
Systolic Blood Pressure	120
Diastolic Blood Pressure	80

Concomittant Medication

Startdate	Enddate	Term	Dose	Unit
01/02/2008	01/07/2009	Aspirin	200	mg
01/07/2009		Paracetamol	100	mg

3 PROC CDISC AND STATEMENT PARAMETERS

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;`

Note: If an older version of PROC CDISC is installed some of the parameters may not be available and their usage will generate an error in the log.

3.2 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 that are used to identify the extact 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”).
In Table 1 are listed all statements/parameters that could be used for importing an ODM XML file into SAS [3].

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. Only ODM Version 1.2 is supported.
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. If 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 cut off 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.
ODM	ORDERNUMBER+	YES(USE)/ NO(IGNORE)	If Yes, the OrderNumber attributes of ItemRefs in your ODM file are validated for missing and out-of-range integers. The default value is YES.

PROC CDISC	Parameters	Values	Description
CLINICALDATA*	OUT	TEXT	Name of the resulting SAS dataset.
CLINICALDATA	SASDATASETNAME	TEXT	This parameter value should match the ODM SASDatasetName.
CLINICALDATA	NAME +	TEXT	If the parameter LONGNAMES=YES is included in ODM part then the value should match the value of the attribute “Name” on Itemgroup definition level.
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: + optional parameter, * required parameter

3.3 EXAMPLE

This is an example of PROC CDISC with some of the common used parameters from Table 1.

Figure 6

```
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;
```

1. Specify the location of the SAS datasets that will be output with PROC CDISC.
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 key fields will be imported into 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 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 clinical sites and investigators. The two parameters `SITEREF` and `INVESTIGATORREF` are optional.

4 THE EXTERNAL ODM XML FILE

The conversion of an XML ODM data file into SAS datasets by using PROC CDISC require checks in order to ensure completeness and correctness of the mapping of ODM ItemGroups into corresponding SAS datasets. The checks are 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

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 transformation 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. If a text item is longer than 200 characters then the length will be truncated when imported with PROC CDISC.
	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 the 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 SAS will produce a WARNING. To avoid this you can use ORDERNUMBER=NO.

Additional 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 they 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 off at the end. (e.g. 8DATE becomes _8DAT).

- If the SAS name has more than eight characters it will be cut to 8 characters. This holds the risk that names are shortened 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 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.

- If an *Item* in the ODM file references an external code list (CDISC ODM) the external code list either needs to be inserted into the metadata part of the ODM XML file or control this with the statement parameters (Table 1) in the PROC CDISC.

5 IMPORT OF 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. The following macro (Figure 7) provides a method for reading in an ODM file containing many *ItemGroup* definitions resulting in multiple SAS datasets. We used the resource DICTIONARY.TABLES available in PROC SQL, PROC CDISC and SAS Macro in order to get at all the data from the external file.

Figure 7

```
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

PROC CDISC is the only tool provided by SAS® Institute currently that allows data to be transferred from ODM XML to SAS datasets. The primary goal is accuracy of the mapping of ODM *ItemGroups* into SAS datasets and this can be achieved by getting familiar with the limitations, possible errors and warnings in advance. The examples in this paper will help the user to better understand and apply the parameters when reading an ODM XML data file 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.sas.com/industry/pharma/cdisc/index.html

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
E-mail: id@xclinical.com
Web: www.xclinical.com