

# Another Way to Create Data Definition Files

Han Zou, AstraZeneca, Wilmington, DE  
David Jemiolo, DataCeutics Inc., Pottstown, PA

## ABSTRACT

Reviewing clinical data can be a very complex and time-consuming process for both sponsor and the FDA reviewer. According to the Guidance for Industry on Providing Regulatory Submission in Electronic Format - NDAs item 11, Data Definition Files (data dictionary) should describe each SAS data set being submitted to the FDA by providing the following: Variable Name, Label, Type, Codes/Decodes, and Comments (ie, Variable Derivation). Producing this Data Definition File often becomes a complicated, error prone process when the pressure of ensuing time lines are approaching. To avoid these pitfalls, AstraZeneca has taken the approach of developing Data Definitions as analysis data sets are being created. Our tool, %XLSDFN, is helping SAS programmers at AstraZeneca create and edit Data Definition Files in a timely fashion. %XLSDFN implements SAS Macro, SAS Export/Import, SAS ODS, and Microsoft Excel and Word in order to create, update, and produce final Data Definition Files on an ongoing basis, alleviating the stress of "last minute" production.

## INTRODUCTION

Compliance to the FDA Guidance for Regulatory Submission in Electronic Format – Item 11– has mandated a more standard creation process of Data Definition Files in the statistical programming department at AstraZeneca. Data Definition Files need to describe each SAS data set being submitted to the FDA by providing the following: Variable Name, Label, Type, Codes/Decodes, and Comments (ie, Variable Derivation).

Creating Data Definition Files are very time consuming because of typing on the comments. Especially adding and removing the variables from analysis data set after the Data Definition Files have been created. To make sure the Data Definition Files are accurately present the analysis data sets before the submission, there are a lot of work load for the programmers.

Our goal was to develop a more efficient method to create, update, and finalize these Data Definition Files. The method of choice is using a SAS macro %XLSDFN to create new Data Definition Files into an EXCEL format, update these spreadsheets during the analysis data set developing period, and create a final version in RTF format.

In order to be a successful tool, %XLSDFN needed to be user friend. To accomplish this, only two macro parameters are absolutely necessary: dsname and version.

Finally, %XLSDFN needed to be versatile and flexible enough for programmers across clinical disciplines to implement for any analysis data set. To accomplish this, two macro parameters (protocol and study) were added. Additionally, although users run %XLSDFN in the Windows operating system environment, data stored on either the UNIX or Windows platform can be accessed.

The final result is as follows:

```
%xlsdefn (protocol=, /* protocol number/name */
          study=, /* study number/name */
          dsname=, /* analysis data set name */
          version=); /* new, update, or final */
```

## LOGIC OF PROCESSING

There are three options for %XLSDFN macro: "New", "Update", and "Final" When the user selects "New", the Data Definition File is created by importing the contents of the analysis data set (i.e., meta-data) and decoding any formatted variables formats via pre-existing format data set. When the user selects "Update", the Data Definition File already exists, but needs to be updated with any new variables or changes to existing variables. %XLSDFN macro will run a comparison process to add or remove the variables that have been updated in the current analysis data set. Additionally, a backup Data Definition File will be created for each updated file present (e.g., demog(backup).xls). When the user selects

"Final", the Data Definition File is deemed complete and the analysis data set is final and no changes are anticipated. The %XLSDFN macro will create RTF document based on existing EXCEL Define file.

Process flow within the %XLSDFN.sas :

- Version = New → [Get variable format from library.](#)  
[Get contents from SAS data set.](#)  
[Create EXCEL spreadsheet](#)  
[Execute EXCEL format macro](#)
- Or  
↓
- Version =Update → [Get variable format from library.](#)  
[Get contents from SAS data set.](#)  
[Compare and update current contents with existing contents](#)  
[Create EXCEL spreadsheet](#)  
[Execute EXCEL format macro](#)
- Or  
↓
- Version = Final → [Import existing EXCEL Define file into SAS data set.](#)  
[Create RTF document using ODS.](#)

## Example:

### Version = New (DEMO Data Set w/ 6 Variables)

- Get variable format from library...

| FORMAT NAME: RACEF |     |                |
|--------------------|-----|----------------|
| START              | END | LABEL          |
| 1                  | 1   | WHITE          |
| 2                  | 2   | BLACK          |
| 3                  | 3   | ASIAN/ORIENTAL |
| 4                  | 4   | HISPANIC       |
| 7                  | 7   | OTHER          |

| FORMAT NAME: \$SEXF |     |         |
|---------------------|-----|---------|
| START               | END | LABEL   |
|                     |     | MISSING |
| M                   | M   | MALE    |
| F                   | F   | FEMALE  |

- Get contents from SAS data set...

| The CONTENTS Procedure                          |      |                  |               |
|---|------|------------------|---------------|
| Data Set Name: MDATA.DEMO                       |      | Observations: 82 |               |
| Member Type: DATA                               |      | Variables: 6     |               |
| --Alphabetic List of Variables and Attributes-- |      |                  |               |
| Variable  | Type | Format           | Label         |
| BIRTHDT   | Num  | DATE9.           | DATE OF BIRTH |
| RACE  | Num  | RACEF.           | RACE          |
| SEX   | Char | \$SEXF.          | SEX           |
| SITEID  | Char |                  | SITE ID       |
| STUDYID   | Char |                  | STUDY ID      |
| SUBJID  | Char |                  | SUBJECT ID    |

- Combine and create unformatted EXCEL spreadsheet...

| Variable | Label         | Type | Decodes   | Origin                              | Comments                  |
|----------|---------------|------|---|-------------------------------------|---------------------------|
| BIRTHDT  | DATE OF BIRTH | Num  | DATE  | CRF Page, Derived, Domain, Variable | Update 'Comments' Section |
| RACE     | RACE          | Num  | 1='WHITE'<br>2='BLACK'<br>3='ASIAN/ORIENTAL'<br>4='HISPANIC'<br>5='ASIAN/ORIENTAL'<br>7='OTHER' | CRF DDEMOG                          | ORIGIN                    |
| SEX      | SEX           | Char | '='='MISSING'<br>F='FEMALE'<br>M='MALE'   | CRF DDEMOG                          | SEX                       |
| SITEID   | SITE ID       | Char |   | CRF DDEMOG                          | CENTRE                    |
| STUDYID  | STUDY ID      | Char |   | CRF DDEMOG                          | TRIAL                     |
| SUBJID   | SUBJECT ID    | Char |   | CRF DDEMOG                          | PATIENT                   |

- Execute EXCEL format macro...

| Variable | Label         | Type | Decodes   | Origin                              | Comments                  |
|----------|---------------|------|---|-------------------------------------|---------------------------|
| BIRTHDT  | DATE OF BIRTH | Num  | DATE  | CRF Page, Derived, Domain, Variable | Update 'Comments' Section |
| RACE     | RACE          | Num  | 1='WHITE'<br>2='BLACK'<br>3='ASIAN/ORIENTAL'<br>4='HISPANIC'<br>5='ASIAN/ORIENTAL'<br>7='OTHER' | CRF DDEMOG                          | ORIGIN                    |
| SEX      | SEX           | Char | '='='MISSING'<br>F='FEMALE'<br>M='MALE'   | CRF DDEMOG                          | SEX                       |
| SITEID   | SITE ID       | Char |   | CRF DDEMOG                          | CENTRE                    |
| STUDYID  | STUDY ID      | Char |   | CRF DDEMOG                          | TRIAL                     |
| SUBJID   | SUBJECT ID    | Char |   | CRF DDEMOG                          | PATIENT                   |

### Example:

Version = Update (Add AGE variable/format to DEMO Data Set)

- Get variable format from library

| FORMAT NAME: RACEF |     |                |
|--------------------|-----|----------------|
| START              | END | LABEL          |
| 1                  | 1   | WHITE          |
| 2                  | 2   | BLACK          |
| 3                  | 3   | ASIAN/ORIENTAL |
| 4                  | 4   | HISPANIC       |
| 7                  | 7   | OTHER          |

| FORMAT NAME: \$SEXF |     |        |
|---------------------|-----|--------|
| START               | END | LABEL  |
| M                   | M   | MALE   |
| F                   | F   | FEMALE |

| FORMAT NAME: AGEGRPF |     |         |
|----------------------|-----|---------|
| START                | END | LABEL   |
| 1                    | 1   | 18 - 64 |
| 2                    | 2   | 65 - 74 |
| 3                    | 3   | >= 75   |

- Get contents from SAS data set...

```

The CONTENTS Procedure
Data Set Name: MDATA.DEMO  Observations: 82
Member Type:  DATA        Variables: 6

--Alphabetic List of Variables and Attributes--
Variable      Type      Format      Label
-----
AGE            Num       AGEGRPF.   AGE IN YEARS
AGEGRP        Num       AGEGRPF.   AGE GROUP IN YEARS
BIRTHDT       Num       DATE9.     DATE OF BIRTH
RACE          Num       RACEF.     RACE
SEX           Char      $SEXF.     SEX
SITEID        Char      SITEID     SITE ID
STUDYID       Char      STUDYID    STUDY ID
SUBJID        Char      SUBJID     SUBJECT ID

```

- Compare and update current contents with existing contents
- Create EXCEL spreadsheet
- Execute EXCEL format macro...

| Variable | Label                          | Type | Decodes   | Origin                              | Comments                  |
|----------|--------------------------------|------|---|-------------------------------------|---------------------------|
| AGE      | AGE IN YEARS AT BASELINE       | Num  | 1='18 - 64'<br>2='65 - 74'<br>3='>= 75'   | CRF Page, Derived, Domain, Variable | Update 'Comments' Section |
| AGEGRP   | AGE GROUP IN YEARS AT BASELINE | Num  | 1='18 - 64'<br>2='65 - 74'<br>3='>= 75'   | CRF Page, Derived, Domain, Variable | Update 'Comments' Section |
| BIRTHDT  | DATE OF BIRTH                  | Num  | DATE  | CRF DDEMOG                          | DATBIRTH                  |
| RACE     | RACE                           | Num  | 1='WHITE'<br>2='BLACK'<br>3='ASIAN/ORIENTAL'<br>4='HISPANIC'<br>5='ASIAN/ORIENTAL'<br>7='OTHER' | CRF DDEMOG                          | ORIGIN                    |
| SEX      | SEX                            | Char | '='='MISSING'<br>F='FEMALE'<br>M='MALE'   | CRF DDEMOG                          | SEX                       |
| SITEID   | SITE ID                        | Char |   | CRF DDEMOG                          | CENTRE                    |
| STUDYID  | STUDY ID                       | Char |   | CRF DDEMOG                          | TRIAL                     |
| SUBJID   | SUBJECT ID                     | Char |   | CRF DDEMOG                          | PATIENT                   |

### Example:

Version = Final (Update define file and create RTF)

- Import existing define file...

| Variable | Label                          | Type | Decodes   | Origin     | Comments  |
|----------|--------------------------------|------|---|------------|---|
| AGE      | AGE IN YEARS AT BASELINE       | Num  | Derived   | Derived    | FLOOR((DATECONS-BIRTHDT)/365.25)  |
| AGEGRP   | AGE GROUP IN YEARS AT BASELINE | Num  | 1='18 - 64'<br>2='65 - 74'<br>3='>= 75'   | Derived    | IF 18 <= AGE <= 64 THEN AGEGRP = 1; ELSE IF 65 <= AGE <= 74 THEN AGEGRP = 2; ELSE IF AGE >= 75 THEN AGEGRP = 3; |
| BIRTHDT  | DATE OF BIRTH                  | Num  | DATE  | CRF DDEMOG | DATBIRTH  |
| RACE     | RACE                           | Num  | 1='WHITE'<br>2='BLACK'<br>3='ASIAN/ORIENTAL'<br>4='HISPANIC'<br>5='ASIAN/ORIENTAL'<br>7='OTHER' | CRF DDEMOG | ORIGIN  |
| SEX      | SEX                            | Char | '='='MISSING'<br>F='FEMALE'<br>M='MALE'   | CRF DDEMOG | SEX   |
| SITEID   | SITE ID                        | Char |   | CRF DDEMOG | CENTRE  |
| STUDYID  | STUDY ID                       | Char |   | CRF DDEMOG | TRIAL   |
| SUBJID   | SUBJECT ID                     | Char |   | CRF DDEMOG | PATIENT   |

- ... into SAS data set
- Create RTF document using ODS

| Variable | Label                          | Type | Decodes   | Origin     | Comments  |
|----------|--------------------------------|------|---|------------|---|
| AGE      | AGE IN YEARS AT BASELINE       | Num  | Derived   | Derived    | FLOOR((DATECONS-BIRTHDT)/365.25)  |
| AGEGRP   | AGE GROUP IN YEARS AT BASELINE | Num  | 1='18 - 64'<br>2='65 - 74'<br>3='>= 75'   | Derived    | IF 18 <= AGE <= 64 THEN AGEGRP = 1; ELSEIF 65 <= AGE <= 74 THEN AGEGRP = 2; ELSEIF AGE >= 75 THEN AGEGRP = 3; |
| BIRTHDT  | DATE OF BIRTH                  | Num  | DATE  | CRF DDEMOG | DATBIRTH  |
| RACE     | RACE                           | Num  | 1='WHITE'<br>2='BLACK'<br>3='ASIAN/ORIENTAL'<br>4='HISPANIC'<br>5='ASIAN/ORIENTAL'<br>7='OTHER' | CRF DDEMOG | ORIGIN  |
| SEX      | SEX                            | Char | '='='MISSING'<br>F='FEMALE'<br>M='MALE'   | CRF DDEMOG | SEX   |
| SITEID   | SITE ID                        | Char |   | CRF DDEMOG | CENTRE  |
| STUDYID  | STUDY ID                       | Char |   | CRF DDEMOG | TRIAL   |
| SUBJID   | SUBJECT ID                     | Char |   | CRF DDEMOG | PATIENT   |

### CONCLUSION

There are several approaches to the creation of Data Definition Files. Through experience, the most efficient and pragmatic way is to create them concurrently as the Analysis data sets are created. This approach reduces the incidence of error as well as provides a valuable validation guideline. However, this approach can at times become redundant. To eliminate redundancy, the macro %XLSDFN allows us to create, update, and finalize Data Definition Files in an easy, versatile manner across operating system platforms. No longer are the pressing timelines of Study completion associated with creation of Data Definition Files.

### CONTACT INFORMATION

If you have any comments or questions, please contact us at:

Han Zou  
AstraZeneca  
1800 Concord Pike  
PO Box 15437  
Wilmington, DE 19850  
Work Phone: (302)-886-1514  
Email: han.zou@astrazeneca.com

David Jemiolo  
DataCeutics, Inc.  
1610 Medical Drive  
Pottstown, PA 19464  
Work Phone: (610)-970-2333  
Email: jemiolod@dataceutics.com