MWSUG 2016 - Paper TT07

A SAS® Macro to Create a Data Dictionary with Ease

Amy Gravely, Center for Chronic Disease Outcomes Research, A VA HSR&D Center of Innovation
Barbara Clothier, Center for Chronic Disease Outcomes Research, A VA HSR&D Center of Innovation

ABSTRACT

Creating a data dictionary offers a huge added value over PROC CONTENTS output to communicate what is contained in a dataset. For example, a data dictionary shows the actual formatted values for each variable and organizes variables into chunks delineating them with sub-headings and even sub-sub headings. This creates a clear, complete, organized view of what is contained in a dataset. There is no need for the user to look up formatted values or to sift through variables labels to find what they are looking for. Not only can a data dictionary save time and prevent frustration for the data analyst, it can also provide a clear, self-explanatory view of what is contained in a dataset when handing off a dataset to someone else on your team or outside of your company. Data dictionaries can also help to keep your own projects organized when working with just one dataset or many. However, creating such a data dictionary by hand can be time consuming with cutting, pasting and typing. This paper walks through the creation and end product of a data dictionary macro. You can learn both the advanced macro techniques used to create the macro as well as to see the final product and request the macro to use for yourself. There are a couple of steps to take before running the macro and only a handful of parameters to fill in to create a data dictionary in a short amount of time. Additionally, changes to a dataset can be quickly accommodated and tracked.

INTRODUCTION

Much of the time when receiving a dataset or handing off a dataset to someone, the PROC CONTENTS output will be utilized to get a sense of what is inside the dataset. PROC CONTENTS can be a great tool, but without any further context, just looking at a PROC CONTENTS can be frustrating and inefficient. There are several ways to add and re-organize information to give context and clarity to what is in the dataset in an efficient way. For example, if the dataset contains survey data one might send a copy of the survey itself along with the PROC CONTENTS. However, even in the best case scenario where PROC CONTENTS is sorted in the exact same order as the survey questions, confusion and frustration can arise when analyzing this data. Additionally there are other situations where some or all of the data does not come from a survey. Areas of frustration might be looking up each format one by one to determine the formatted values, searching through labels to find the variables that you want to include, making sure that you have all of the demographic variables, trying to figure out which variables came from administrative data and which came from a survey and the list goes on. Something more than just a sorted PROC CONTENTS is needed such as a data dictionary. Small examples of output from both this data dictionary macro and PROC CONTENTS are shown in APPENDICEES A and B respectively.

A data dictionary can add context, ease and efficiency in knowing exactly what is contained in a dataset. This can save hours of time for an analyst and save needless time asking questions to the person who created the dataset. The data dictionary is intended to be a stand-alone, self-explanatory clear and complete view of what is contained in a dataset. Furthermore, by automating this process through advanced SAS macro techniques, one can further utilize data dictionaries as an organizational tool on projects. Having the ability to create a data dictionary easily quickly and repeatedly allows one to keep track of analysis datasets, keep track of versions of variables and helps to keep all projects organized in general with all project members on the same page about the decisions that were made. Lastly, having the most current analysis dataset documented with a data dictionary helps in those situations where one has to come back to a dataset after being away from it for a period of time.

WHAT DOES THE MACRO DO?

A small excerpt of what this macro produces can be seen below in Appendix A. Some of the features of this data dictionary macro happen automatically behind the scenes within the macro and other features one can specify through macro parameters. One feature that happens behind the scenes is SAS grabbing the formatted values (through SAS dictionary features not to be confused with this data dictionary). Through macro parameters one can specify the subheadings and sub-sub headings that one would like. There is also the option to specify areas where you might not want the formatted values shown such as an area where there is a large chunk of variables with the same formatted values. Showing the formatted values only once will save space. The outputted information includes: the dataset name, the date created, number of observations, number of variables, who the data dictionary was created by, the variable names, types, length, format name, formatted values (this is important) and label.

HOW TO USE IT

Steps to take before running the macro

The macro is very simple to use. First one must make all formats permanent in order for them to show up in the data dictionary. So if temporary formats are created simply put them into the format library to make them permanent. Generally speaking this is good practice anyway in most circumstances.

Additionally, one should run the following code and examine the formats to ensure that there are no duplicates or strange things in there (that in many cases will show up at the bottom of the file in an obvious way). Any duplicate formats must be removed. The macro can handle both character and numeric formats but they cannot have the same format name.

```
proc format lib=stat cntlout=sas_fmt; run;
```

Prepping your dataset

As a matter of good practice, it is best to have all the formats incorporated and labels assigned before running this macro. Additionally, before running the macro you must retain only those variables that you want to be in the data dictionary and put those variables in the order that you want them to appear in the data dictionary. The dataset itself can be temporary or permanent but especially if you are handing the dataset off to someone you might prefer it to be a permanent dataset.

Table 1. The macro parameters

dataset	the dataset you want to create a data dictionary for
lib	the library where the permanent formats are stored
numvar	the number of variables in your dataset
sstart	the number of the variable where you want a subheading to start above
sstop	the number of the variable where you want the subheading to end
headt	in order with spaces between write out the words you want in your sub headings
numhead	the number of sub headings that you want
substart	where you want your sub-sub headings to start
subsstop	where you want your sub-sub headings to end
subheadt	in order with spaces between write out the words you want in your sub-sub headings
subnumhead	the number of sub-sub headings you want
skipformat	list the variable number for any variable that you don't want the formatted values listed
title	the title of the data dictionary
Filepathname	where you want the final outputted dataset to reside with .rtf at the end

Here is an example of the macro parameters that produce the data dictionary in Appendix A, although only the first page is shown.

```
%amydatadict(
dataset=anls28,
lib=stat,
numvar=65,
sstart=1 8 14 19 30 32 34 38 40 49 60,
sstop=7 13 18 29 31 33 37 39 48 59 65,
headt=Demographics Admin_Dx Survey_Drug_Alc Survey_Scales Survey_Symptoms
PMAQ Cold_Pressor Chair_Test Gait Hand_Grip Other,
numhead=11,
subsstart=1 5,
subsstop=4 65,
subheadt=Admin_Demographics Survey_Demographics,
subnumhead=2,
skipformat=9 10 11 12 63,
title=Data Dictionary SPACE Analysis Dataset,
filepathname=G:\Project_Analysis\Krebs_SPACE\Programs\SAS\playupdatemacrodd\o
utput\spacedraft223ssNEXT.rtf); run;
```

LIMITATIONS

At this time the maximum number of variables that this data dictionary macro can intake and output is 234. However, if there are more variables than that one can run the macro more than once. Additionally each variable name can only be up to 32 characters or the macro might run into issues.

MACRO TECHNIQUES UTILIZED TO CREATE THE MACRO

This macro was really easy to create.

Generally the following steps were taken:

1. Created a style template with style=statistical as the base
2. Read in macro parameters; some had to be read in as lists
3. Put the lists into datasets
4. Made macro variables and macro lists from the datasets
5. ODS output PROC CONTENTS; order=varnum
6. Use SAS dictionary.format features to get the actual formatted values
7. Do some fancy merging to make sure that the PROC CONTENTS output and formatted value output merge correctly
8. Use PROC REPORT with style created in step 1 (one can play with changing this style to a pre-existing or newly created style to achieve a different look)
9. Use PROC REPORT compute block with line statement features to achieve sub headings and sub-sub headings

CONCLUSION

To save time, frustration and confusion, consider using this data dictionary macro to create data dictionaries. This one small change may save you time and energy in ways that you haven’t even yet considered. One example might be saving time when running large models in that you can be sure that you are including all of the variables that you want to or in knowing quickly that you have identified the variables that you want (only baseline measures or only imputed measures as two examples) without sifting relentlessly though PROC CONTENTS. Another example might be when you are recoding variables to make them better suited for fitting a large model. Having all of the formatted values in front of you rather than looking each one up one by one could save you loads of time and headache. Yet another example is the time it can save in answering questions after handing an analysis dataset off to someone else. Using this data dictionary macro can improve project communication, project organization, project documentation and save you time and frustration.

REFERENCES

Carpenter A (1997), Resolving and Using &&var&i Macro Variables, Proceedings of the Twenty-Two Annual SAS Users Group International Conference.

Carpenter A (2005), Storing and Using List of Values in a Macro Variable, *SUGI 30*, Paper 028-30.

Carpenter A (2007), Advanced PROC REPORT: Doing more with the compute block, *SAS Global Forum*, Paper 242-2007.

Eslinger J (2015), The Report Procedure: A Primer for the Compute Block, *SAS Paper*, 1642-2015.

Hamilton J (2005), Using the COMPUTE Block in PROC REPORT, *WUSS,* 2005.

Haworth L (2004), SAS® with Style: Creating your own ODS Style Template for RTF Output. *SUGI 29*, Paper 125-29.

Lafler K (2005), Exploring DICTIONARY Tables and Views, *SUGI 30*, Paper 070-30.

Thorton P (2011), SAS® Dictionary Step by Step, *SAS Global Forum*, Paper 264-2011.

Zender C (2008), Creating Complex Reports, *SAS Global Forum*, Paper 173-2008.

ACKNOWLEDGMENTS

This material is the result of work supported with resources and the use of facilities at the Minneapolis VA Health Care System.

This work was supported by Merit Review Award #101-HX-000671 from VA HSR&D.

Small portion of VA data shown with permission.

RECOMMENDED READING

SAS® Macro Language 1: Essentials.

SAS® Macro Language 2: Advanced Techniques

CONTACT INFORMATION

Your comments, questions or requests to have the macro code are valued and encouraged. Contact the author at:

Amy Gravely, Statistician
Center for Chronic Disease Outcomes Research, A VA HSR&D Center of Innovation
612-467-5208
Amy.gravely@va.gov

APPENDIX A

Data Dictionary SPACE Analysis Dataset
Date Created: 30AUG2016
Datasetname: anls28
Number of Observations: 264
Number of Variables: 65
This document was created by: VHAMINGraveA

	Variable	Type	Length	Format	User Defined Format?	Label
Demographics						
Admin_Demographics						
1	**age_a**	Num	4			Age
2	**gender_a**	Num	8	GE	yes	**Gender**
						1=female
						2=male
3	**primarypainlocation**	Num	8	FPL	yes	**PrimaryPainLocation**
						1=Back Pain
						2=Knee or Hip Pain
4	**treatmentgroup**	Num	8	FTX	yes	**TreatmentGroup**
						0=Opioid - Intensive
						1=Opioid - Avoidant
Survey_Demographics						
5	**race**	Num	8	RACE	yes	**Race**
						1=White
						2=Black/AfAmer
						3=Asian
						4=Native Amer/Alaska
						5=Native Hawaii/Pacific
						6=Hispanic/Latino
						7=Multiple
6	**education**	Num	8	ED	yes	**Education level**
						1=< 4 yr degree
						2=4 yr degree +
7	**employment**	Num	8	EMP	yes	**Employment Status**
						1=employed for wages
						2=self employed
						3=retired
						4=other

APPENDIX B

```
                                The CONTENTS Procedure

           Data Set Name        WORK.ANLS28                      Observations          264
           Member Type          DATA                             Variables             65
           Engine               V9                               Indexes               0
           Created              Thu, Sep 08, 2016 11:01:55 AM    Observation Length    608
           Last Modified        Thu, Sep 08, 2016 11:01:55 AM    Deleted Observations  0
           Protection                                            Compressed            NO
           Data Set Type                                         Sorted                NO
           Label
           Data Representation  WINDOWS_32
           Encoding             wlatin1  Western (Windows)


                             Engine/Host Dependent Information

                        Data Set Page Size          16384
                        Number of Data Set Pages    11
                        First Data Page             1
                        Max Obs per Page            26
                        Obs in First Data Page      10
                        Number of Data Set Repairs  0
                        Filename                    E:\_TD5540\anls28.sas7bdat
                        Release Created             9.0202M3
                        Host Created                W32_ESRV08



 # Variable            Type Len Format    Informat Label

 1 age_a               Num    4                    Age
 2 gender_a            Num    8 GE.                Gender
 3 primarypainlocatio  Num    8 FPL.      11.      PrimaryPainLocation
   n
 4 treatmentgroup      Num    8 FTX.      11.      TreatmentGroup
 5 race                Num    8 RACE.              Race
 6 education           Num    8 ED.                Education level
 7 employment          Num    8 EMP.               Employment Status
 8 charlsonscore_a     Num    8                    Charlson comorbidity score
 9 depression_a        Num    8 NOYES.             Depression disorder
10 anxiety_a           Num    8 NOYES.             Anxiety disorder
11 ptsd_a              Num    8 NOYES.             Post traumatic stress disorder
12 alcoholuse_a        Num    8 NOYES.             Alcohol use disorder
13 druguse_a           Num    8 NOYES.             Drug use disorder
14 V0_smoking_now      Num    8 SMOKENOW. 6.       time0  How often do you smoke NOW
15 V0_alcohol_         Num    8 NOYES.    6.       time0  Had at least 1 alcoholic
   12months                                        drink in past 12 months
16 V0_audit_score      Num    8                    Audit Total Score on People who had at
                                                   least 1 drink in the past 12 months
17 V0_audit_cut        Num    8 NOYES.             Audit Total score of 8 or higher
18 V0_drug_use         Num    8 NOYES.             Illicit drug use in the past 12 months
                                                   based on V0_drug_times question
19 V0_bpi_severity     Num    8                    BPI Severity Average Score
20 V0_bpi_             Num    8                    BPI Interference Average Score
   interference
21 V0_roland_          Num    8                    Total Roland Disability Score
   disability
22 V0_mcs12            Num    8                    MCS Score from VR12
23 V0_pcs12            Num    8                    PCS Score from VR12
24 V0_phq_score        Num    8                    PHQ Score
25 V0_phqcut           Num    8 CU.                PHQ Cut Score
26 V0_gad_score        Num    8                    Generalized Anxity Disorder total score
27 V0_gadcut           Num    8 CU.                Generalized Anxiety Discorder Total Score Cut
```