

Comparing dataset metadata

**Jim Groeneveld,
OCS Consulting,
's Hertogenbosch, Netherlands.**

PhUSE 2011

AGENDA / CONTENTS

A. Comparing dataset data and metadata

1. PROC COMPARE
2. macro %CrossRef

B. Dataset and variable attributes

C. Example results (in dataset)

1. Dataset attributes
2. Variable attributes

D. Application of macro %CrossRef

E. Some technical information

F. Future features

A. Comparing dataset data and metadata

1. PROC COMPARE

- a. data oriented (attributes: NOVALUES option)
- b. only 2 datasets (or variables in one) at a time
- c. cumbersome output (summary: OUT= dataset)
- d. may be tuned as desired, yet limited to pairs

2. SAS macro %CrossRef

- a. structure oriented: dataset & variable attributes
- b. any number of specified datasets (from 1)
- c. tabular summarisation (in result dataset only)
- d. columns: dataset names; rows: attributes
- e. user specification of desired attributes

B. Dataset and variable attributes

1. Dataset attributes

- a. MemName, MemLabel and LibName
- b. Creation and Modification date and time
- c. Number of variables and physical observations

2. Variable attributes

- a. Name (common name in first attribute column)
- b. Label: as value in above Name attribute record
if no label then text: "-no label-"
if no corresponding variable: empty
- c. optional variable's Type and Length (combined)
- d. optional variable's Informat and Format

C. Example results (in dataset) 1/2

Dataset attributes

attribute
column

dataset dataset dataset
1 2 3

columns=datasets, rows=variables and attributes (Structure)	cumulative_ac (cumulative_ac)	recon_20100128 (recon_20100128)	Intrial_Images (Intrial_Images)
*LibName	Datalib	Datalib	Datalib
*Dataset	cumulative_ac	recon_20100128	Intrial_Images
*DsLabel	-no label-	-no label-	-no label-
*NofVars	48	43	18
*NofObs	1706	1978	46
*Created	11JAN2010:13:41:59.78	28JAN2010:10:53:05.08	22SEP2008:14:15:54.36
*Changed	11JAN2010:13:41:59.78	28JAN2010:10:53:05.08	22SEP2008:14:15:54.36

C. Example results (in dataset) 2/2

Variable attributes

attribute
column

dataset dataset dataset
1 2 3

columns=datasets, rows=variables and attributes	cumulative_ac	recon_20100128	Intrial_Images
SCAN_DATE	scan_date	scan_date	scan_date
>Type/Length	Numeric/8	Numeric/8	Numeric/8
>Informat	DATE9.0	DATE9.0	DATE9.0
>Format	DATE9.0	DATE9.0	DATE9.0
SCAN_TIME	scan_time	scan_time	scan_time
>Type/Length	Char/5	Char/5	Numeric/8
>Informat	\$5.0	\$5.0	TIME8.0
>Format	\$5.0	\$5.0	TIME8.0
SCAN_TYPE	-no label-	-no label-	
>Type/Length	Char/11	Char/11	
>Informat			
>Format			

D. Application of macro %CrossRef

1. not with entirely different datasets but with a (limited) number of rather similar datasets to view differences
 - a. master datasets and subsets of them
 - b. different versions of datasets
 - c. same datasets with different names
 - d. similar datasets with different data
2. Goal: to see whether more datasets could be combined into one dataset (or ignored if the data are identical)

E. Some technical information

1. all fields are type character of length \$256, first, attribute field has \$36
2. internally SAS name literal variable names are applied
 - a. OPTIONS VALIDVARNAME=ANY is set, and reset to the original state at the end of the macro
 - b. variable names starting with an asterisk (*) or ending with an exclamation mark (!) and one digit. Avoid such names in your datasets and limit your variable name length to maximally 30
3. **WORK** dataset names start with

F. Future features 1/2

- 1. comparing all datasets in one or more libraries using a wildcard (LibName.*)**
- 2. optional aggregated data for both numerical and character variables**
 - a. (non-deleted) logical number of observations
 - b. number of non-missing values
 - c. number of missing values
 - d. frequency distribution of a limited number of distinct (formatted) values (categories)
 - e. minimum and maximum (formatted) value (first and last non-missing character value)

F. Future features 2/2

3. optional aggregated, univariate data for (mainly) numerical variables

- a. mean value
- b. median value (also approximate middle, non-missing, sorted, character value)
- c. (formatted) mode value (also most occurring non-missing character value)
- d. standard deviation
- e. various percentiles
- f. and more, e.g. distribution information and the statistics that PROC COMPARE can generate

QUESTIONS & ANSWERS

SASquestions@ocs-consulting.com

Jim.Groeneveld@OCS-Consulting.com

<http://jim.groeneveld.eu.tf>

SAS name literal

A name expressed as a string within quotes, followed by the letter N.

Applicable to variable names, statement labels and imported variable and table names from DBMS tables (e.g. Excel).

Advantage: more compatibility. Example:
`'This @#$name'n = 'a SAS name literal';`

More information in:

SAS Language Reference: Concepts.

Straightforward inventory of metadata

1. save results of PROC CONTENTS (or of the CONTENTS statement of PROC DATASETS for one or more libraries) to datasets,
2. if desired keep the most important variables LibName, MemName, Name, Label, Type, Length, Format, FormatL, FormatD, Informat, InformL and InformD;
3. concatenate all metadata datasets (SET);
4. if desired sort by variable NAME.

This generates all dataset and variable information in subsequent records.