

Leveraging SAS for Data Science Applications

Sunil Gupta

**Director Statistical Programming, Cytel
Data Science using SAS Instructor
(UCSD, UCLA Extension Online Class)**



Data Science using SAS: Three Pillars

Computers (ETL, Speed)

- SAS Programming
- R, Python
- Big Data
- Data Cleaning

Statistics (MetaData Driven)

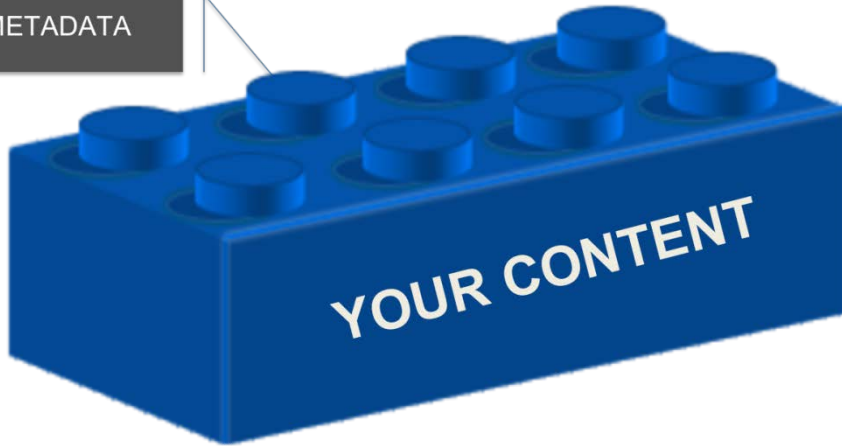
- Descriptive Stats
- Prediction
- Analytics
- Increase/
Decrease/ No
Difference

Mathematics (Applications)

- Complex equations
- Problem Solving
- Business Solutions

What is Metadata?

METADATA



- ✓ Metadata describes and documents the content
- ✓ Metadata is separate from content but within the same file
- ✓ Standard and custom metadata
- ✓ No hard coding of content
- ✓ Standardize process
- ✓ Automate production process

Metadata

The three main types work together.

Attributes/ Elements

Structural

Values

Descriptive or administrative

Metadata example

Structural

Title:

Description:

Creator:

Date:

Type:

Descriptive

'Mad Men' Season 5 plot predictions
The two-hour premiere quickly answered lingering questions from season four.

Administrative

Lucas Shaw

26/03/2012

Article

Metadata Is All Around Us – Excel, Word, etc.

The screenshot shows the MetadataTouch application window. The title bar reads "MetadataTouch - C:\\$Docs\The Importance of Using Metadata Removal Software.docx". The menu bar includes "File", "Batch Process", and "Help". Below the menu is a toolbar with icons for file operations. A file list on the left shows several documents, with "The Importance of Using Metadata Removal Software.docx" selected. The main area is titled "Document Properties" and contains a list of metadata attributes with checkboxes and their corresponding values. A blue rounded rectangle highlights the "Attributes (Variables)" column, and another blue rounded rectangle highlights the "Valid Values (Data Values)" column. Red text on the left side of the window states: "Word File", "> Info", "> Properties", "Invalid Metadata", "results in", "corrupt file".

Word File
> Info
> Properties

**Invalid Metadata
results in
corrupt file**

**Attributes
(Variables)**

**Valid Values
(Data Values)**

Attribute	Value
Title:	The Importance of Using Metadata Removal Software
Subject:	High-profile incidents of data leaks due to metadata
Author:	Yossi M.
Company:	Digital Confidence Ltd.
Manager:	
Keywords:	metadata removal software, data leakage prevention
Category:	Whitepapers
Status:	APPROVED FOR PUBLICATION
Comments:	the whitepaper demonstrates the importance of using metadata removal software to prevent sensitive data leakage due to hidden data & metadata.
Last modified by:	Shirley G.
Revision number:	10
Total editing time:	55 Minutes
Template:	Normal
Application:	Microsoft Office Word
Application version:	12.000
Created:	20/02/2009 17:55:00
Modified:	06/04/2009 15:37:00
Last printed:	15/04/2009 11:50:14
Custom properties:	
Website	http://www.digitalconfidence.com

SAS: Extract intelligence information from metadata and macro processing

INPUT:
Directory
of Files
(Datasets, SAS
Programs)

SAS Tools

- Libnames
- Data Step
- SAS Macro Programming
- SAS & Dataset Functions
- Proc SQL
- Proc Compare
- Proc Means

OUTPUT: Metadata Attributes

- File pathname and names
- # of Files, Datetime stamps
- # and Type of variables
- Macro loop through all files
- Required datasets, variables, etc.
- Codelist of values
- Maximum variable lengths
- Data cleaning & monitoring of valid variables and special characters
- Compare and contrast previous file
- Descriptive Statistics on categorical and continuous variables
- Search for ERRORS, WARNINGS or Notes in SAS Logs

SAS's Four Key Dictionary Tables

Category	Description	PROC SQL or DATA Step: SASHelp View
1. Dictionaries	Category of files: Catalogs, etc.	sashelp.vdctnry
2. Members	List of SAS libraries and data sets	sashelp.vmember
3. Tables	List of data sets and columns	sashelp.vtable
4. Columns	List of columns and attributes	sashelp.vcolumn

 dm	8/25/2016 7:40 PM	SAS Data Set	256 KB
 dosing	1/13/2013 1:11 PM	SAS Data Set	17 KB
 formatdata	1/13/2013 1:11 PM	SAS Data Set	17 KB
 labs	1/13/2013 1:11 PM	SAS Data Set	33 KB
 pain	1/13/2013 1:11 PM	SAS Data Set	9 KB
 responders	8/25/2016 7:40 PM	SAS Data Set	256 KB
 sdtm_metadata	8/25/2016 7:08 PM	SAS Data Set	0 KB
 sdtm_metadata1	8/25/2016 7:09 PM	SAS Data Set	0 KB
 sdtm_metadata2	8/25/2016 7:10 PM	SAS Data Set	256 KB
 suppdm	8/25/2016 7:40 PM	SAS Data Set	256 KB
 xp	8/25/2016 7:40 PM	SAS Data Set	256 KB
 adae	9/18/2013 6:51 AM	SAS System Program	2 KB
 adamtrace	1/13/2013 1:11 PM	SAS System Program	13 KB
 adef	9/18/2013 6:50 AM	SAS System Program	2 KB
 adsl	9/18/2013 6:51 AM	SAS System Program	3 KB
 adtte	9/18/2013 6:51 AM	SAS System Program	4 KB

Datasets

**Apply
DO
LOOP
process**

**SAS
Programs**

**Metadata File:
Pathname, Type,
Name, DateTime
Stamp, Size,
Vars, # OBS, etc.**



Metadata Application SAS Examples

Access/Process Metadata like a SAS dataset

**Check
DateTime
Sequence of
Source/QC
SAS/Log Files**

**Check
DateTime
Sequence of
Datasets
Across
Multiple
Libraries**

**Automatically
Proc Compare
Two SDTM
Libraries**

Check DateTime Sequence of Source/QC SAS/Log Files

Source SAS < QC SAS < SAS Log < QC Log

file_name	s_sas_dt	q_sas_dt	s_log_dt	q_log_dt	issue
f_dor_crprc3	01JUN18:10:54:10	01JUN18:11:02:04	10JUL18:12:26:49	10JUL18:12:47:47	Correct Sequence
f_dorc3	13JUL18:17:22:04	01JUN18:11:01:13	13JUL18:17:22:52	13JUL18:17:23:09	
f_forest_cr_addcov	20JUN18:11:03:27	18JUN18:10:50:28	23JUL18:09:13:54	23JUL18:09:19:45	
f_forest_cr_keyc3	06JUL18:07:10:55	05JUL18:14:13:55	10JUL18:12:02:53	10JUL18:12:49:44	
f_forest_crc3	06JUL18:07:01:26	05JUL18:14:08:26	10JUL18:12:02:48	10JUL18:12:49:42	

SAS Path Names

- Source & QC SAS Programs and Logs
- **Metadata Information – No hard coding**

Datasets

- Dataset functions to access metadata
- Four datasets with datetime stamps

Proc SQL

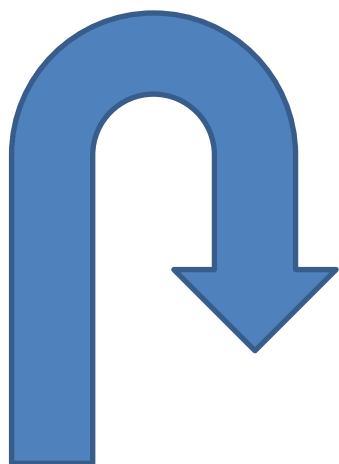
- Full Join by root SAS Program Name
- Issue variable to identify SAS Programs

Check DateTime Sequence of Files: Code Snippet

**SAS
Files**

file_name	s_sas_dt	q_sas_dt	s_log_dt	q_log_dt	issue
f_dor_crprc3	01JUN18:10:54:10	01JUN18:11:02:04	10JUL18:12:26:49	10JUL18:12:47:47	Correct Sequence
f_dorc3	13JUL18:17:22:04	01JUN18:11:01:13	13JUL18:17:22:52	13JUL18:17:23:09	
f_forest_cr_addcov	20JUN18:11:03:27	18JUN18:10:50:28	23JUL18:09:13:54	23JUL18:09:19:45	
f_forest_cr_keyc3	06JUL18:07:10:55	05JUL18:14:13:55	10JUL18:12:02:53	10JUL18:12:49:44	
f_forest_crc3	06JUL18:07:01:26	05JUL18:14:08:26	10JUL18:12:02:48	10JUL18:12:49:42	

**Loop through
each record in
metadata
dataset**



```

do i=1 to num_of_files;
  file_name=dread(dirID, i);
  rc1=filename("fRef",catx("/", "&pathdir", file_name));
  fid=fopen("fRef");

  if fid > 0 then do;
    full_name=finfo(fid, "FileName");
    bytes=finfo(fid, "File Size (bytes)");
    &dsnme._dt=input(finfo(fid, "Last Modified"), anydtdtm.);
    owner=finfo(fid, "Owner Name");
    rc2=fclose(fid);
    * remove extentions;
    file_name2 = tranwrd(strip(file_name), '.sas', '');
    file_name2 = tranwrd(strip(file_name2), '.log', '');
    output;
  end;
end;
fld = fClose( fld );
run;
  
```

Automatically Proc Compare Two SDTM Libraries

Source Library				QC Library		
 dm	8/25/2016 7:40 PM	SAS Data Set		 dm	8/25/2016 7:40 PM	SAS Data Set
 dosing	1/13/2013 1:11 PM	SAS Data Set		 dosing	1/13/2013 1:11 PM	SAS Data Set
 formatdata	1/13/2013 1:11 PM	SAS Data Set		 formatdata	1/13/2013 1:11 PM	SAS Data Set
 labs	1/13/2013 1:11 PM	SAS Data Set		 labs	1/13/2013 1:11 PM	SAS Data Set
 pain	1/13/2013 1:11 PM	SAS Data Set		 pain	1/13/2013 1:11 PM	SAS Data Set

Summary Level

- **SASHELP.VTABLE** from two Libnames
- **Metadata: CRDATE, NOBS, NVAR, FILESIZE**

Metadata

- Same SDTMs in both Libraries
- **SASHELP.VCOLUMN** to create Lookup Table of Dataset Names

Detail Level

- **CALL EXECUTE** Proc Compare each Dataset Name
- Extract NOBS from OUTNOEQUAL datasets

Dataset Exists Metadata Application Example

Example Application 7: Testing for dataset existence and number of observations; using this metadata to control program flow

General principles: %SYSFUNC is used within a macro to call SCL functions which access dataset information. This information is used to control the actions of, in this example, a reporting program. This permits the program to be run without errors when the expected dataset is empty, or even if it does not exist at all.

Syntax:

```
%macro genrpt(dsn=);  
  %if %sysfunc(exist(&dsn)) = 1 %then %do;  
    %let dsid = %sysfunc(open(&dsn));  
    %let nobs = %sysfunc(attrn(&dsid,nobs));  
    %if &nobs GT 0 %then %do;  
      %put NOTE: &nobs obs will be reported from &dsn..;  
      *<Usual reporting syntax here>;  
    %end;  
    %else %do;  
      %put NOTE: 0 obs in &dsn - Producing placeholder report. ;  
      *<Alternative reporting syntax here>;  
    %end;  
    %let rc = %sysfunc(close(&dsid));  
  %end;  
  %else %do;  
    %put WARNING: Dataset &dsn does not exist.;  
  %end;  
%mend genrpt;  
%genrpt(dsn=sae);
```

Examples of Standard Metadata from SASHELP

Standard Metadata for all Domains and Variables

1. vtable: Confirm # of datasets, observations and datetime stamps
2. vtable: Confirm required datasets and variables
3. vcolumn: Confirm # and type of variables and attributes
4. vcolumn: Confirm order of variables

Examples of Customized Metadata

Proc SQL Custom Metadata on Key Domains and Variables

1. Confirm # of patients
2. Confirm required non-missing variables
3. Confirm not less than Maximum variable lengths
4. Confirm descriptive statistics on key categorical and continuous variables
5. Confirm compare and contrast previous datasets, attributes (more/less datasets/records/variables)
6. Confirm no duplicate records in each dataset
7. Include codelist dictionary from each dataset, ex. lab units

Leveraging SAS for Data Science Applications

Sunil Gupta

**Director Statistical Programming, Cytel
Data Science using SAS Instructor
(UCSD, UCLA Extension Online Class)**

