DBCompare: A powerful tool to compare data definitions of datasets across multiple databases
Balakrishna Dandamudi, Clinical Pharmacology Services a division of PharmaNet, Kennett Square, PA

ABSTRACT:
SAS® programmers often end up working on data from different databases, referring to SAS® data only. In this scenario a programmer’s primary task would be analyzing the inconsistencies in defining the datasets and variables within the database and across databases.

DBCompare generates an extensive analysis report about the databases that are being compared:

- What datasets are in common across databases
- What datasets are in one database and not in the other
- Whether matching datasets have different labels, different variables
- Whether matching variables have different formats, labels, or types

DBCompare is a tool written in Base SAS® for the comparison of data definitions of SAS® datasets across databases.

INTRODUCTION:
We are all familiar with PROC COMPARE. The COMPARE procedure does exactly what it says: it compares two data sets. What if you need to compare more than two data sets?

DBCompare is the program written to address this problem. However, this program is for comparing data definitions of multiple data sets and reporting them in a matrix form. One other interesting aspect is, in the entire program you find four macro variables, not even a single macro. This is an attempt made to give an alternate answer to the same old problem.

STEP 1:

options nofmterr;
%let db1path = C:\NESUG2006\DBCompare\DB1;
%let db2path = C:\NESUG2006\DBCompare\DB2;
%let db3path = C:\NESUG2006\DBCompare\DB3;
%let outpath = C:\NESUG2006\DBCompare\Output;

libname DBCMP1 "&db1path";
libname DBCMP2 "&db2path";
libname DBCMP3 "&db3path";

Assign the DB folder locations to the first three macro variables and report destination to the OUTPATH macro variable. User modifies this part before executing the program.

STEP 2:

proc sql noprint;
create table ndbs as
    select distinct 'DB' as DB, libname
    from sashelp.vtable
    where COMPRESS(UPCASE(libname)) contains 'DBCMP'
    order by db;
quit;
data ndbs;
set ndbs;
libnum = _N_
run;

Collect library names into a data set NDBS with derived variable LIBNUM which is holding the record number. This is the key data set we use in data _null_ steps to navigate through multiple data sets.

**STEP 3:**
data _null_; set ndbs;
call execute("proc contents data = "||trim(left(libname))||"._all_ out=work.m"||trim(left(libname)))||" noprint;run;"); call execute("proc sort data = "||trim(left(libname))||"; by memname name; run;");
run;

Collect meta data of each database into MDBCMPx data set. "x" represents the number of the DB, for example MDBCMP1, MDBCMP2, and MDBCMP3. Order the data sets by MEMNAME and NAME.

**STEP 4:**
data _null_; set ndbs;
by db;
if first.DB then do;
call execute("proc sql noprint; ");
call execute(" create table mdata as ");
call execute(" select distinct memname, name ");
call execute(" from m"||trim(left(libname)))||" ");
end;
if not first.DB and not last.db then do;
call execute(" union ");
call execute(" select distinct memname, name ");
call execute(" from m"||trim(left(libname)))||" ");
end;
if last.DB then do;
call execute(" union ");
call execute(" select distinct memname, name ");
call execute(" from m"||trim(left(libname)))||" ");
call execute(";quit;");
end;
run;

Collect unique MEMNAME and NAME combinations into the MDATA1 data set.

**STEP 5:**
data _null_; set ndbs;
by db;
call execute("proc sql;");
call execute(" create table mdata "
||compress(put(input(substr(trim(left(libname)),6),best.)+1,best.))
||" as ");
call execute(" select a.*, b.libname as libnam
||substr(trim(left(libname)),6)||" label =', b.memlabel as memlab"
||substr(trim(left(libname)),6)||" label =', "");
call execute(" b.TYPE as TYPE"||substr(trim(left(libname)),6)
||" label =', b.LENGTH as LENGTH"||substr(trim(left(libname)),6)
||" label =', "");
call execute (" b.VARNUM as VARNUM" || substr(trim(left(libname)),6) ||" label ='", b.LABEL as LABEL" || substr(trim(left(libname)),6) ||" label ='");
call execute (" b.FORMAT as FORMAT" || substr(trim(left(libname)),6) ||" label ='", b.FORMATL as FMTL" || substr(trim(left(libname)),6) ||" label ='");
call execute (" b.JUST as JUST" || substr(trim(left(libname)),6) ||" label ='", b.NPOS as NPOS" || substr(trim(left(libname)),6) ||" label ='");
call execute (" b.NOBS as NOBS" || substr(trim(left(libname)),6) ||" label ='", b.ENGINE as ENGINE" || substr(trim(left(libname)),6) ||" label ='");
call execute (" b.GENNEXT as GENNXT" || substr(trim(left(libname)),6) ||" label =''");
call execute (" from mdata" || substr(trim(left(libname)),6)||" a left join mdbcmp"|| substr(trim(left(libname)),6)||" b ");
call execute (" on a.memname = b.memname");
call execute (" and a.name = b.name");
call execute ("quit;");
run;

Collect all meta data variables needed for comparison into the MDATAx data sets - one data set per DB. This time for DB1 the data set is MDATA2 and for DB2 the data set is MDATA3 and for DB3 the data set is MDATA4.

The final MDATA4 data set contains all meta data information of all data bases. The meta data variables are differentiated within the data bases by suffixed numbers.

**STEP 6:**

data _null_
set ndbs;
by db;
if first.DB then do;
call execute ("proc sql noprint;");
call execute (" create table misslbl as");
call execute (" select distinct memname, memlab
"||compress(put(libnum,best.))|| ",");
end;
if not first.DB and not last.db then do;
call execute (" memlab"||compress(put(libnum,best.))||",");
end;
if last.DB then do;
call execute (" memlab"||compress(put(libnum,best.)))

call execute (" from mdata"||compress(put(libnum+1,best.)))||";");
call execute ("quit;");
end;
run;
data _null_
set ndbs;
by db;
if first.DB then do;
call execute ("data misslbl;")
call execute (" set misslbl;"");
call execute (" if compress(MEMLAB"
||compress(put(libnum,best.))
||")= '0' then LBLFLAG = 0; ");
call execute (" else LBLFLAG = (compress(MEMLAB"
||compress(put(libnum,best.)))||") = ");
end;
if not first.DB and not last.db then do;
call execute (" compress(MEMLAB"|| compress(put(libnum,best.))
||")= ");
end;
if last.DB then do;
call execute ("compress(MEMLAB"||compress(put(libnum,best.))
||"));
call execute ("run;");
end;
run;

Now we have all the meta data information in MDATA4, if we have two data bases to compare then this data set will be MDATA3. Analyze the data and create two data se ts – MISSLBL and DDCOMP. MISSLBL holds the analysis results on dataset labels. The LBLFLAG is the flag used to indicate the label differences. Flag result 0 indicates no exact match or missing label and 1 indicates an exact match.

data _null_; set ndbs; by db;
if first.DB then do;
call execute ("proc sql noprint;");
call execute ("create table ddcomp as");
call execute ("select distinct memname length = 10,name length = 10,");
call execute ("case ");
call execute ("when type"||compress(put(libnum,best.))
||" = 1 then 'CHAR'");
call execute ("when type"||compress(put(libnum,best.))
||" = 2 then 'NUM'");
call execute ("else put(type"||compress(put(libnum,best.))
||",BEST.)");
call execute ("end as TYPE"||compress(put(libnum,best.))
||"C length = 4,");
call execute ("length"||compress(put(libnum,best.))||",");
call execute ("format"||compress(put(libnum,best.))||"",");
end;
if not first.DB and not last.db then do;
call execute ("case ");
call execute ("when type"||compress(put(libnum,best.))
||" = 1 then 'CHAR'");
call execute ("when type"||compress(put(libnum,best.))
||" = 2 then 'NUM'");
call execute ("else put(type"||compress(put(libnum,best.))
||",BEST.)");
call execute ("end as TYPE"||compress(put(libnum,best.))
||"C length = 4,");
call execute ("length"||compress(put(libnum,best.))||",");
call execute ("format"||compress(put(libnum,best.))||"",");
end;
if last.DB then do;
call execute ("case ");
call execute ("when type"||compress(put(libnum,best.))
||" = 1 then 'CHAR'");
call execute ("when type"||compress(put(libnum,best.))
||" = 2 then 'NUM'");
call execute ("else put(type"||compress(put(libnum,best.))
||",BEST.)");
call execute ("end as TYPE"||compress(put(libnum,best.))
||"C length = 4,");
call execute ("length"||compress(put(libnum,best.))||",");
call execute ("format"||compress(put(libnum,best.))||"",");
end;
run;
Data set DDCOMP holds analysis results for the remaining data definitions. These include Type, Length and Format. The TFLAG, LFLAG, and FFLAG are the flags used to indicate the differences. Flag result 0 indicates no exact match and 1 indicates an exact match.
STEP 7:

`proc template;`  
`define style styles.Arial;`  
`parent=styles.printer;`  
`replace fonts /`  
'BatchFixedFont' = ("Arial",7pt)  
'TitleFont2' = ("Arial",7pt)  
'TitleFont' = ("Arial",7pt)  
'StrongFont' = ("Arial",7pt)  
'EmphasisFont' = ("Arial",7pt)  
'FixedEmphasisFont' = ("Arial",7pt)  
'FixedStrongFont' = ("Arial",7pt)  
'FixedHeadingFont' = ("Arial",7pt)  
'FixedFont' = ("Arial",7pt)  
'headingEmphasisFont' = ("Arial",7pt)  
'headingFont' = ("Arial",7pt)  
'docFont' = ("Arial",7pt);`  
`end;`  
`run;`  
`ods listing close;`  
`ods html body = "&outpath\body.html"`  
`contents = "&outpath\contents.html"`  
`frame = "&outpath\frame.html";`  
`ods pdf file = "&outpath\DBCompare.pdf" style=styles.Arial;`  
`ods proclabel "DB Compare";`  
`title 'Data Definitions Compare';`  
`data _null_;`  
`file print ods;`  
`set ddcomp ;`  
`put _ods_;`  
`run;`  
`ods proclabel "DB Compare";`  
`title 'Data set labels compare - with missing label flag';`  
`data _null_;`  
`file print ods;`  
`set misslbl ;`  
`put _ods_;`  
`run;`  
`ods _all_ close;`  
`ods listing;`  

Used ODS to generate both HTML and PDF reports.

CONCLUSION

There are different ways to address the task of comparing the data definitions of data sets across multiple data bases. DBCompare is one of those many answers.

ACKNOWLEDGMENTS

The author would like to express his deepest gratitude to Dr. Mike P. Adams, Clinical Pharmacology Services a division of PharmaNet and Mrs. Jennifer Horan, Clinical Pharmacology Services a division of PharmaNet. Without their encouragement, this paper might not have come alive.
TRADEMARKS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at

Balakrishna Dandamudi
Senior SAS Programmer
SFBC Clinical Pharmacology Services
Longwood Corporate Center South
415 McFarlan Road, Suite 201
Kennett Square, PA 19348-2412
Work Phone: 610.444.4722
Fax: 610.444.4663
Email: kdandamudi@sfbci.com

<table>
<thead>
<tr>
<th>Library Member Name</th>
<th>Variable Name</th>
<th>TYPE1C</th>
<th>LENGTH1</th>
<th>FORMAT1</th>
<th>TYPE2C</th>
<th>LENGTH2</th>
<th>FORMAT2</th>
<th>TYPE3C</th>
<th>LENGTH3</th>
<th>FORMAT3</th>
<th>TFLAG</th>
<th>LFLAG</th>
<th>FFLAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENDSTUDY</td>
<td>DSTDIM</td>
<td>CHAR</td>
<td>6</td>
<td>TIME</td>
<td>CHAR</td>
<td>6</td>
<td>TIME</td>
<td>CHAR</td>
<td>6</td>
<td>TIME</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ENDSTUDY</td>
<td>DSSSTMC</td>
<td>NUM</td>
<td>4</td>
<td>NUM</td>
<td>NUM</td>
<td>4</td>
<td>NUM</td>
<td>NUM</td>
<td>4</td>
<td>NUM</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ENDSTUDY</td>
<td>ENDTST</td>
<td>CHAR</td>
<td>8</td>
<td>DATE</td>
<td>CHAR</td>
<td>8</td>
<td>DATE</td>
<td>CHAR</td>
<td>8</td>
<td>DATE</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LAB</td>
<td>COMPANY</td>
<td>NUM</td>
<td>20</td>
<td>NUM</td>
<td>20</td>
<td>NUM</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAB</td>
<td>DOB</td>
<td>CHAR</td>
<td>8</td>
<td>MMDDYY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>GREPGNO</td>
<td>CHAR</td>
<td>6</td>
<td></td>
<td>CHAR</td>
<td>6</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>IECAT</td>
<td>CHAR</td>
<td>6</td>
<td>IECAT</td>
<td>CHAR</td>
<td>6</td>
<td>IECAT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>IECODE</td>
<td>CHAR</td>
<td>8</td>
<td></td>
<td>CHAR</td>
<td>8</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>IEDIT</td>
<td>CHAR</td>
<td>8</td>
<td>DATE</td>
<td>CHAR</td>
<td>8</td>
<td>DATE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>IEDTC</td>
<td>NUM</td>
<td>0</td>
<td></td>
<td>NUM</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE3</td>
<td>IETEST</td>
<td>NUM</td>
<td>60</td>
<td></td>
<td>NUM</td>
<td>60</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>IETESTCD</td>
<td>CHAR</td>
<td>8</td>
<td></td>
<td>CHAR</td>
<td>8</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>LINEDUM</td>
<td>CHAR</td>
<td>8</td>
<td></td>
<td>CHAR</td>
<td>8</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>PROTOCOL</td>
<td>NUM</td>
<td>20</td>
<td></td>
<td>NUM</td>
<td>20</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>SITEID</td>
<td>NUM</td>
<td>3</td>
<td></td>
<td>NUM</td>
<td>3</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>SUBJID</td>
<td>NUM</td>
<td>4</td>
<td></td>
<td>NUM</td>
<td>4</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>SUBJUNIT</td>
<td>NUM</td>
<td>3</td>
<td></td>
<td>NUM</td>
<td>3</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>VISID</td>
<td>CHAR</td>
<td>8</td>
<td>VSN</td>
<td>CHAR</td>
<td>8</td>
<td>VSN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Library Member Name</td>
<td>memlab1</td>
<td>memlab2</td>
<td>memlab3</td>
<td>LBLFLAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>Additional</td>
<td>Additional</td>
<td>Additional</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AES</td>
<td>Adverse Events Summary</td>
<td>Adverse Events Summary</td>
<td>Adverse Events Summary</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AES.CODED</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONMEDS</td>
<td>Concomitant Medications Summary</td>
<td>Concomitant Medications Summary</td>
<td>Concomitant Medications Summary</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEMO0</td>
<td>Demographics</td>
<td>Demographics</td>
<td>Demographics</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ea</td>
<td>Erythema Assessment of Application Site</td>
<td>Erythema Assessment of Application Site</td>
<td>Erythema Assessment of Application Site</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECG</td>
<td>Electrocardiogram (12-Lead)</td>
<td>Electrocardiogram (12-Lead)</td>
<td>Electrocardiogram (12-Lead)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENDRudy</td>
<td>Study Exit</td>
<td>Study Exit</td>
<td>Study Exit</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENDCRT</td>
<td>Eligibility Criteria - Inclusion/Exclusion</td>
<td>Eligibility Criteria - Inclusion/Exclusion</td>
<td>Eligibility Criteria - Inclusion/Exclusion</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FH</td>
<td>Family History</td>
<td>Family History</td>
<td>Family History</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>Waiver</td>
<td>Waiver</td>
<td>Waiver</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>Investigator Signature Form</td>
<td>Investigator Signature Form</td>
<td>Investigator Signature Form</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAB</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>