Paper PO01

Retrieving SAS Programs Back From PDF Files

Eric Zhang, Centocor Inc., Malvern, PA

ABSTRACT

There are many tools available to convert text files, including SAS® programs, to files with the Portable Document Format ("PDF®"). PGM2PDF, a SAS macro developed by Zhang (2004), can convert SAS programs to PDF files with specific functions and features in order to store SAS programs with protection against unexpected modification. Unfortunately, there are few tools available to retrieve SAS programs back from PDF files once they have been converted to the PDF files. This paper illustrates a solution using SAS and not relying on third party software. In addition, this approach is straightforward and works well across platforms. The capabilities, features, and usage of the approach are described in the paper.

KEYWORDS: PDF2PGM, PGM2PDF, Macro, PDF Files, SAS Programs

INTRODUCTION

In order to retrieve a SAS program from a PDF file and save it as a SAS program as a text file, users usually open the PDF file, select contents in the PDF file, paste the selection into the SAS program window or a text editor, and then save it as a text file with the file extension SAS (SAS program). The PDF file, which is archived from a SAS program, may contain extra texts. These extra texts including titles, footnotes and/or anything else in the body of the PDF file are not the part of the SAS program. In this situation, the retrieved SAS program may contain additional texts that don't belong to the original SAS program. The retrieved SAS program is not the original SAS program and may get errors when executed in SAS. Even if the retrieved SAS program doesn't contain extra texts, its layout will be changed in most cases so that it is not easy to read the retrieved SAS program. Enlightened by those concerns, a SAS macro PDF2PGM is developed. This macro doesn't produce extra texts. It also keeps the same layout as the original SAS program. SAS programs produced by macro PDF2PGM are exactly the same as SAS programs before the conversion.

MACRO DEVELOPMENT

Since different tools convert SAS programs to PDF files in different ways, the converted PDF files may have different structures, for example, page objects and stream objects. It's hard to develop a SAS macro that is capable of retrieving SAS programs back from any PDF files of SAS programs. It's also hard to maintain the macro (PDF to SAS) after the tool (SAS to PDF) is updated or modified in the future, if we can't control the tool. Therefore, it is wise to select one tool that converts SAS programs to PDF files with the specific features we require and that produces PDF files with the structure we are familiar with. The macro PGM2PDF is a good selection in that it meets our requirements of the conversion and the structure of PDF files produced by macro PGM2PDF is fixed. In this sense, macro PDF2PGM is designed to work for the PDF files generated by macro PGM2PDF.

PDF files generated by macro PGM2PDF have many features. Titles, footnotes and watermark will affect the stream objects of PDF files, and are not the part of SAS programs. The macro PDF2PGM is able to identify these in PDF files and exclude them when retrieving SAS programs from PDF files.

When retrieving a SAS program back from a PDF file, users want the retrieved program to be exactly the same as the original SAS program. The macro PDF2PGM keeps the exact blank spaces before and within a statement and the exact blank lines between statements of a SAS program. As a result, it makes the layout of the SAS program the same as before the conversion to PDF. Samples of a PDF file converted from a SAS program and a SAS program retrieved from a PDF file are shown in the Appendix.

Macro PGM2PDF is portable across operating systems and Macro PDF2PGM works well in different platforms. A PDF file generated by macro PGM2PDF may have titles, footnotes and/or watermark in its pages, and the macro PDF2PGM has three macro parameters to handle them. Aside from these manual settings, the macro will take care of everything else. The macro

PDF2PGM is easy to use.

MACRO METHOD

A PDF file produced by macro PGM2PDF is a text file. The macro PDF2PGM reads in the PDF file as a text file and saves it as an initial SAS data. It then determines the text strings of the SAS program and saves them as a SAS dataset. Finally, the SAS dataset is put out as a SAS program. The following steps describe how to get text strings of a SAS program in a PDF file.

PAGE OBJECT

A PDF file has a page object that defines object references, which contain text strings in pages. As an example, the following page object defines a PDF file with nine pages and its nine stream object references.

```
4 0 obj
<<
/Type /Pages
/Count 9
/MediaBox [ 0 0 792 612 ]
/Kids [
11 0 R
14 0 R
17 0 R
20 0 R
23 0 R
26 0 R
29 0 R
32 0 R
35 0 R
]
>>
endobj
```

STREAM OBJECTS

Stream objects are referenced by entries of kids in the page object. The contents in a page of a PDF file must be contained between the statements STREAM and ENDSTREAM in a stream object. Each line in a stream object starts with the beginning “T* (“ and ends with the ending “) Tj”. The following example is a stream object.

```
11 0 obj
stream
…
…
…
endstream
endobj
```

After determining a page object and stream objects in a PDF file, the macro PDF2PGM is able to select text strings of the SAS program and saves them as a SAS dataset.

MACRO STRUCTURE

Macro PDF2PGM has 7 parameters. The parameters handle file names and locations of PDF files and SAS programs, page headers and footnotes, and watermarks in PDF files. All of these parameters along with the structure of macro PDF2PGM are described below in detail.

```
%PDF2PGM(Dir=,
         Member=,
         MemType=PDF,
         PgmDir=&dir,
         PgHeader=N,
         PgFooter=N,
         Wmark=N)
```

Dir = Filename associated with a directory where PDF files reside in
Member = File names of PDF files which will be converted back to SAS programs. File names must be separated by vertical bars
MemType = File extension of PDF files which will be converted back to SAS programs. The default is PDF
PgmDir = Filename associated with a directory where SAS programs will reside in. The default directory is the same as the directory that contains PDF files
PgHeader = Whether PDF files have headers at pages. It could be Y (Yes) or N (No). The default is N
PgFooter = Whether PDF files have footers at pages. It could be Y (Yes) or N (No). The default is N
Wmark = Whether PDF files have watermark. It could be Y (Yes) or N (No). The default is N

MACRO USAGE

Users may retrieve SAS programs back from the selected PDF files or from all PDF files in a directory. Users also have choices to save the retrieved SAS programs in the same directory as the PDF files or in a different directory. The retrieved SAS programs keep the same file names as the PDF files but with a different file extension SAS. The following examples demonstrate specifically how the macro PDF2PGM is used to retrieve SAS programs back from PDF files.

EXAMPLE 1

Users want to retrieve SAS programs back from two PDF files “Counts Table.PDF” and “Adverse Events.PDF” and save them in the same directory c:\pdf2pgm\pdfiles. The retrieved SAS programs will be “Counts Table.SAS” and “Adverse Events.SAS”. Before calling the macro, users should define a FILENAME associated with the directory where the PDF files are stored. The PDF files, generated by macro PGM2PDF, have titles, footnotes, and no watermark.

```
FILENAME DIRFN 'c:\pdf2pgm\pdfiles';
%PDF2PGM(Dir=dirfn,
         Member=counts table | adverse events,
         PgHeader=Y,
         PgFooter=Y)
```

EXAMPLE 2

Users may select all PDF files in a directory c:\pdf2pgm\pdffiles and retrieve SAS programs from them and save them in the different directory c:\pdf2pgm\sasprograms. Before calling the macro, users should define two FILENAMEs associated with two directories. One is the directory where PDF files are stored and the other is the directory where SAS programs will be saved. The PDF files have watermark and don’t have titles and footnotes.

```
FILENAME PDFDIRFN 'c:\pdf2pgm\pdffiles';
FILENAME PGMDIRFN 'c:\pdf2pgm\sasprograms';
%PDF2PGM(Dir=pdfdirfn,
         Member=_ALL_,
         PgmDir=pgmdirfn,
         Wmark=Y)
```

CONCLUSION

Macro PDF2PGM enables users to retrieve SAS programs back from PDF files, which were converted from the SAS programs by macro PGM2PDF, without any third-party software support. Macro PDF2PGM is portable across operating systems. The retrieved SAS programs are exactly the same as the original SAS programs. If users would like to plan to submit SAS programs with PDF in a programming submission package, macros PGM2PDF and PDF2PGM are good tools. SAS users could submit the PDF files of SAS programs generated by macro PGM2PDF as well as macro PDF2PGM. Reviewers can retrieve SAS programs back from PDF files by utilizing macro PDF2PGM at their platforms.

REFERENCES

Eric Zhang (2004). “Creating PDF Files for SAS Programs”. Presented at PharmaSUG 04, San Diego, California, May 23 – 26, 2004.

ACKNOWLEDGMENTS

The author would like to thank Leah Bednarek for reviewing the paper, comments and the support of publication of the paper.

TRADEMARKS

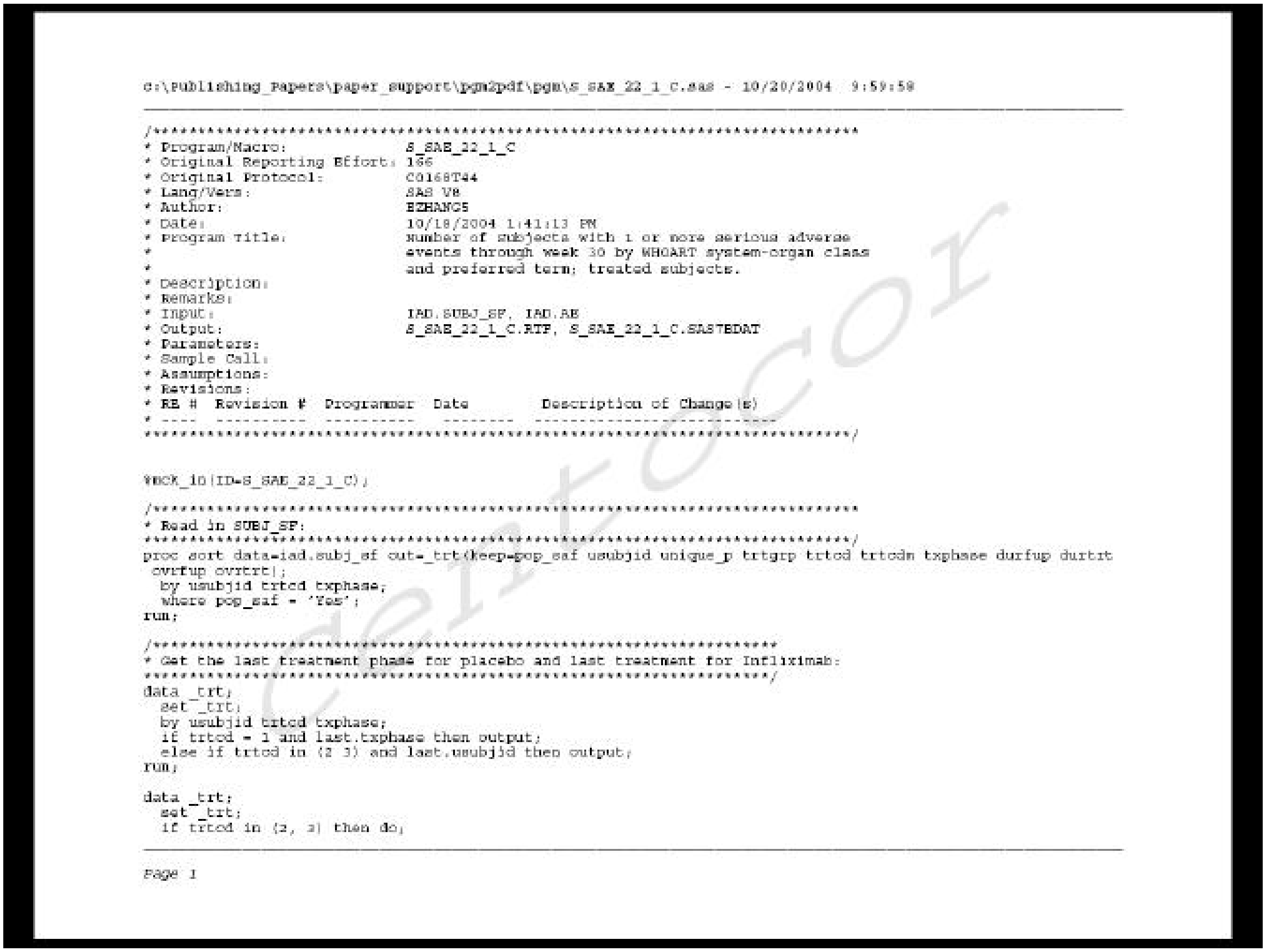
CONTACT INFORMATION

Your comments and questions are valued and encouraged. You may contact the author by mails or e-mails at the following addresses:

Eric Zhang, ezhang5@cntus.jnj.com

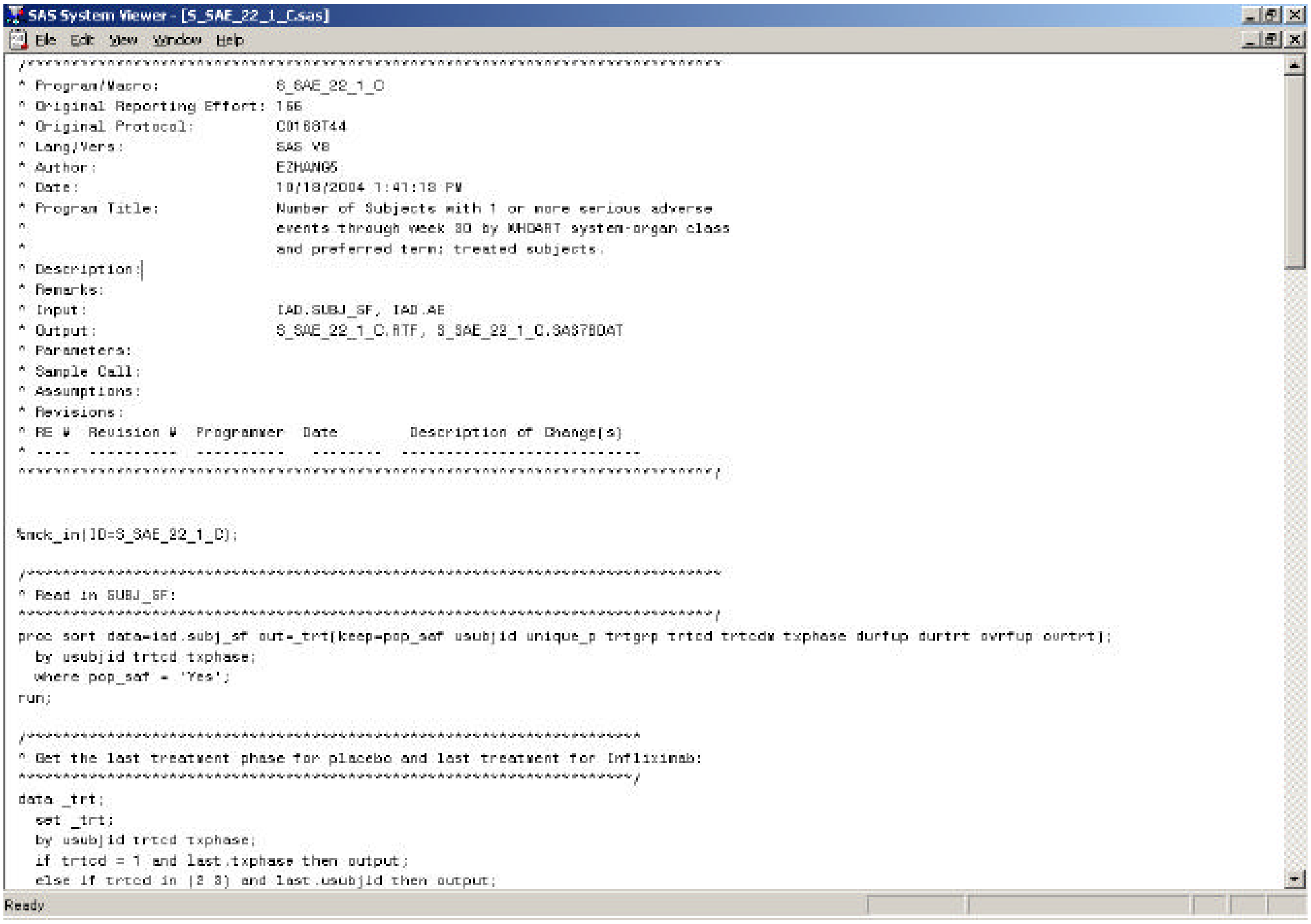
Centocor Inc., C4-1
200 Great Valley Parkway
Malvern, PA 19355

APPENDIX



```
c:\Publishing_Papers\paper_support\pgm2pdf\pgm\S_SAE_22_1_C.sas - 10/20/2004  9:59:58
_____________________________________________________________________________________________
/****************************************************************************************
* Program/Macro:              S_SAE_22_1_C
* Original Reporting Effort:  166
* Original Protocol:          C0168T44
* Lang/Vers:                  SAS V8
* Author:                     EZHANG5
* Date:                       10/18/2004 1:41:13 PM
* Program Title:              Number of subjects with 1 or more serious adverse
*                             events through week 30 by WHOART system-organ class
*                             and preferred term; treated subjects.
* Description:
* Remarks:
* Input:                      IAD.SUBJ_SF, IAD.AE
* Output:                     S_SAE_22_1_C.RTF, S_SAE_22_1_C.SAS7BDAT
* Parameters:
* Sample Call:
* Assumptions:
* Revisions:
* RE #  Revision #  Programmer  Date        Description of Change(s)
* ----  ----------  ----------  --------  ---------------------------
****************************************************************************************/


%mck_in(ID=S_SAE_22_1_C);

/****************************************************************************************
* Read in SUBJ_SF:
****************************************************************************************/
proc sort data=iad.subj_sf out=_trt(keep=pop_saf usubjid unique_p trtgrp trtcd trtcdm txphase durfup durtrt
 ovrfup ovrtrt);
  by usubjid trtcd txphase;
  where pop_saf = 'Yes';
run;

/**********************************************************************************
* Get the last treatment phase for placebo and last treatment for Infliximab:
**********************************************************************************/
data _trt;
  set _trt;
  by usubjid trtcd txphase;
  if trtcd = 1 and last.txphase then output;
  else if trtcd in (2 3) and last.usubjid then output;
run;

data _trt;
  set _trt;
  if trtcd in (2, 3) then do;
_____________________________________________________________________________________________

Page 1
```

Figure 1: A PDF file converted from a SAS program by macro %PGM2PDF – S_SAE_22_1_C.PDF


SAS System Viewer - [S_SAE_22_1_C.sas]
File Edit View Window Help
Ready

```
/*******************************************************************************
* Program/Macro:              S_SAE_22_1_C
* Original Reporting Effort: 166
* Original Protocol:          C0168T44
* Lang/Vers:                  SAS V8
* Author:                     EZHANG5
* Date:                       10/18/2004 1:41:13 PM
* Program Title:              Number of Subjects with 1 or more serious adverse
*                             events through week 30 by WHOART system-organ class
*                             and preferred term; treated subjects.
* Description:
* Remarks:
* Input:                      IAD.SUBJ_SF, IAD.AE
* Output:                     S_SAE_22_1_C.RTF, S_SAE_22_1_C.SAS7BDAT
* Parameters:
* Sample Call:
* Assumptions:
* Revisions:
* RE #  Revision #  Programmer  Date        Description of Change(s)
* ----  ----------  ----------   --------  ---------------------------
*******************************************************************************/


%mck_in(ID=S_SAE_22_1_C);

/*******************************************************************************
* Read in SUBJ_SF:
*******************************************************************************/
proc sort data=iad.subj_sf out=_trt(keep=pop_saf usubjid unique_p trtgrp trtcd trtcdw txphase durfup durtrt ovrfup ovrtrt);
  by usubjid trtcd txphase;
  where pop_saf = 'Yes';
run;

/**********************************************************************
* Get the last treatment phase for placebo and last treatment for Infliximab:
**********************************************************************/
data _trt;
  set _trt;
  by usubjid trtcd txphase;
  if trtcd = 1 and last.txphase then output;
  else if trtcd in (2 3) and last.usubjid then output;
```

Figure 2: A SAS program retrieved from a PDF file by macro %PDF2PGM – S_SAE_22_1_C.SAS