# Using RTF, EPS and PDFMARK to Automate the Creation of Your DEFINE.PDF Document for Electronic Submissions

Dirk Spruck, Aventis Behring, Marburg, Germany

## ABSTRACT

The Food and Drug Administration's (FDA) guideline 'Providing Regulatory Submissions in Electronic Format' outlines a data definition table, the so-called DEFINE document. This PDF document consists of three parts 1) a list of all datasets with labels and location, 2) a variable definition table grouped by datasets and 3) a list of all variables in alphabetical order.

This paper describes the technology used to create a DEFINE.PDF document complete with bookmarks, internal links and hyperlinks to external files. SAS is used to create an RTF file with all contents and structural document information, e.g. paper size and font color, including bookmarks and embedded Encapsulated Postscript (EPS) files. These EPS files contain PDFMARK language used by Acrobat to create external hyperlinks when the document is distilled to PDF.

## INTRODUCTION

For submitting clinical data as SAS transport files to the FDA the data definition table or DEFINE.PDF is probably the most important documentation of your datasets. Manual creation of this document can be very labor intensive and error prone. Since most information necessary is inherent in the datasets themselves and in the structure of the document an automation of this task seems reasonable.

This paper is based on the experiences of preparing datasets for the CBER (Center for Biologics Evaluation & Research) division of the FDA. The techniques described can easily be adapted to the requirements of CDER (Center for Drug Evaluation & Research).

### DEFINE.PDF DOCUMENT STRUCTURE

The DEFINE.PDF document is required by the FDA to document the submitted datasets. It consists of three parts:

1. List of datasets
2. List of all variables with descriptions by dataset
3. Alphabetical list of variables

The list of datasets typically has three columns: the dataset name, the dataset description and the location of the dataset in the submission. The entries in the dataset column are hyperlinks that open the respective section of the document listing the variables for this domain. The location column is also hyperlinked. Clicking on the location will open the SAS export file.

| Dataset | Description | Location |
|---|---|---|
| CONMED | Concomitant Medication | CRT\datasets\2002\conmed.xpt |
| DEMO | Demography | CRT\datasets\2002\demo.xpt |

The list of variables contains at least 5 columns: Variable name, Label, Type, Format/Code and a comment field. The comment field describes the derivation rule used to calculate the variable or the location of the variable on the annotated CRF. The column should be linked to the correct page of the annotated CRF or to additional documents (e.g. Statistical Analysis Plan).

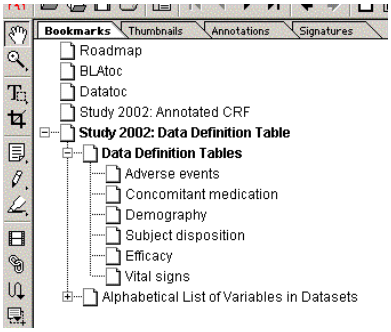
| Definition of Variables in Dataset **DEMO** (Demography) | | | | | |
|---|---|---|---|---|---|
| **Variable** | **Label** | **Type** | **For mat** | **Code** | **Comment** |
| SEX | Sex | Num | SEX | 1= male 2= female | Demo-graphy Page 10 |

The alphabetical list of variables has the following columns: Variable name, Variable label, Dataset name and Dataset label. The dataset name should be linked to open the respective dataset.

| **Alphabetical list of variables** | | | |
|---|---|---|---|
| **Variable** | **Label** | **Dataset** | **Dataset Label** |
| PT | Subject | AE | Adverse Events |
| PT | Subject | CONMED | Concomitant Medication |
| PT | Subject | DEMO | Demography |

The FDA requires hyperlinks and bookmarks for the ease of navigation within and between the documents. External bookmarks provide links to the different table of contents leading from the root of the submission to the DEFINE document.

Internal bookmarks are provided for the three parts of the document. The detailed variable listing should also have subsequent bookmarks to each variable domain.




In this example the bookmarks "Roadmap", "BLAtoc", "Datatoc" and "Study 2002: Annotated CRF" were added after the automatic creation of the DEFINE.PDF document.

For submissions to the CBER division of the FDA all links are required to be blue. Paper size and margins should follow the FDA guideline.

### PREPARING DATASET INFORMATION

Most content of the DEFINE document is already available through the datasets. Variable name, label, type and format can be retrieved from SASHELP.VCOLUMN, with the PROC CONTENTS procedure or with PROC SQL from the DICTIONARY.COLUMNS.

Dataset name and label are stored in the DICTIONARY.TABLES or SASHELP.VTABLES.

The format definitions can be accessed with PROC FORMAT and they need to be merged with the respective variables.

The data in the comment's column, which links the variables to the annotated CRF also needs to be merged with the dataset information. In our project EXCEL was chosen to enter this data. There are various ways to import the data into SAS, which is beyond the scope of this paper. If EXCEL is used together with PROC IMPORT it might be necessary to insert a dummy observation into the spreadsheet. This is because SAS determines the length of the variables looking at the lengths of the first observations in the file and the comment column can become lengthy.

The EXCEL files used in this project had three spreadsheets. The first spreadsheet contains annotated CRF page numbers for every dataset, e.g. all variables found in the DEMO dataset are on the Demography Page 10.

| | A | B |
|---|---|---|
| 1 | **MEMNAME** | **COMMENT** |
| 3 | AE | Adverse event Pages 74 and 89 |
| 4 | CONMED | Concomitant Medication / Prior Therapy Pages 69 and 85 |
| 5 | DEMO | Demography Page 10 |
| 6 | ECG | ECG Pages 34 and 63 |
| 7 | VITAL | Vital Signs Pages 13, 15 and 17 |

In the second spreadsheet there is information for general variables that were merged to all datasets like sex, race and treatment group, e.g. the date of birth variable BIRTHDI is on Demography Page 10.

| | A | B |
|---|---|---|
| 1 | **NAME** | **COMMENT** |
| 3 | SEX | Demography Page 10 |
| 4 | RACE | Demography Page 10 |
| 5 | BIRTHDI | Demography Page 10 |
| 6 | AGE | Derived: AGE = Subject's age at date of randomization (RNDI) |
| 7 | DOGR | Derived: DOGR= 60 |

The third spreadsheet contains derivations or CRF page information for individual variables, which are not originally in the dataset. In the example below the variable WT from the Demography page was added to the AE dataset.

| | A | B | C |
|---|---|---|---|
| 1 | **MEMNAME** | **NAME** | **COMMENT** |
| 3 | AE | STDYDAY | Derived: STDYDAY = AESTDI - DAY1 + 1 |
| 4 | AE | WT | Demography Page 10 |

First the information for individual variables from the third spreadsheet is merged to the dataset information. Then the data from the second spreadsheet for common variables is added. All missing information is then filled with the default assignments for datasets from the first table.

**CREATING RTF FILES**

Considering the complex and rather cryptic nature of RTF it would be a big challenge to design a macro for the creation of an RTF file. Fortunately this work has already been done by Jianmin Long, Merck, and is published on the SAS web site [1]. With the %RTF macro an RTF document is first initialized. Then table rows filled with information from SAS datasets are written to the file.

The macro was originally designed to tabulate results from statistical analysis. Therefore a few changes had to be implemented in order for the macro to be usable for creating 'normal' tables. The main changes were:

- Correct FCHARSET2 to FCHARSET3 for ARIAL font (typo in source code)
- Add stylesheet and color definitions
- Add blank as delimiter in scan function for column width parameter &M and change %EVAL to %SYSEVALF (necessary to use decimals for &M)
- Add \clbrdrl\brdrhair to vertical borders
- Change functionality of macro parameter for vertical borders &V to have full control over vertical borders for all table cells. E.g.:
  &V= 101 means two columns with outside vertical cell borders only,
  &V= 0 means no borders,
  &V= a means display of all vertical cell borders
- Macro variable NUM_OF_V removed

In this example two styles, Heading 1 and Heading 2, are defined. Style information can be used to create hierarchical bookmarks in the PDF document. The following code was inserted into the document header definition of the %RTF macro:

```
%* define heading styles;
put "{\stylesheet";
put "{\s1 \ql\widctlpar heading 1;}";
put "{\s2 \ql\widctlpar heading 2;}}";
```

The definition for color looks similar:

```
%* define colors;
put
"{\colortbl;\red0\green0\blue0;\red0\green0\b
lue255;...}";
```

When used in the text the color definition can simply be referenced by the number. In this example the first color is black and the second blue. When blue text is required for hyperlinks it can be used by specifying the RTF command \cf2.

| Other helpful RTF commands | |
|---|---|
| \b, \b0 | Turn bold on and off |
| \f2 | Use the font #2 |
| \trhdr | Use row as table header which appears at the top of every page |
| \trkeep | Keep table row together, no page break within a row |
| {\info{\title &doctitle} {\author &author} {\company &company}} | Fill document information fields Title, Author and Company. &doctitle, &author and &company are SAS macro variables |
| \header | Start of the document header definition |
| \footer | Start of the document footer definition |

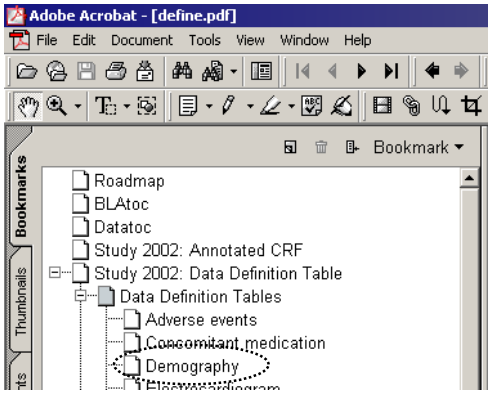
Document information, table size and column width, paper size, margins, header- and footer information and font information can easily be standardized using the %RTF macro within a SAS program.

**PDF BOOKMARKS**

PDF bookmarks are created using the font style information. All text that is supposed to be a bookmark in the PDF file is formatted as a header in the RTF document. Using hierarchical header styles can create hierarchies of bookmarks. Acrobat Distiller recognizes the style information and converts the text to bookmarks.

In the following example 'Demography', the dataset label, is formatted as 'Heading 3'. It will appear in the PDF file as a bookmark to the start of the section listing the demography dataset variables.

| Definition of Variables in Dataset **DEMO** (Demography) | | | | | |
|---|---|---|---|---|---|
| **Variable** | **Label** | **Type** | **For mat** | **Code** | **Comment** |
| PT | Subject | Num | | | Demo-graphy Page 10 |




**INTERNAL DOCUMENT LINKS**

The internal document links are used to link the list of datasets to the specific section of the document describing the respective dataset. For this type of link the word processing bookmark functionality was used.

In the following example a click on the link 'DEMO' should open the section listing all variables residing in the demography dataset.

| **Dataset** | **Description** | **Location** |
|---|---|---|
| AE | Adverse Events | CRT\datasets\2002\ae.xpt |
| CONMED | Concomitant Medication | CRT\datasets\2002\conmed.xpt |
| DEMO | Demography | CRT\datasets\2002\demo.xpt |

| Definition of Variables in Dataset **DEMO** (Demography) | | | | | |
|---|---|---|---|---|---|
| **Variable** | **Label** | **Type** | **For mat** | **Code** | **Comment** |
| PT | Subject | Num | | | Demograph y Page 10 |

The dataset names serve as hyperlinks to the section of the DEFINE document describing the variables of the respective dataset. With the field codes turned on in Microsoft Word the contents of the cell in the dataset column looks like this:

{HYPERLINK \l "DEMO"}

with the RTF definition of the link 'DEMO':

{\field {*\fldinst {HYPERLINK \\l "DEMO" }}{\fldrslt {\cf2 DEMO }}}

Please note that double '\' is necessary because '\' indicates a new RTF command. But in this case '\l' is not an RTF command but part of the word processing hyperlink syntax.

The start of the section describing the demography dataset is a bookmark. The unique dataset name ist used as the name for the bookmark.

RTF definition of the bookmark "DEMO":

{*\bkmkstart DEMO}{Definition of Variables in Dataset }
{*\bkmkend DEMO}

This type of link is very easy to implement because it uses the word processing functionality. Unfortunately it can only be used for links within the document. For links referencing other PDF files or SAS datasets external hyperlinks must be used.

**EXTERNAL HYPERLINKS**

There are two types of external hyperlinks. The first type opens a file by launching the program associated with the file extension. This is used to open SAS export files, e.g. in the SAS Viewer.

| **Dataset** | **Description** | **Location** |
|---|---|---|
| AE | Adverse Events | CRT\datasets\2002\ae.xpt |
| CONMED | Concomitant Medication | CRT\datasets\2002\conmed.xpt |
| DEMO | Demography | CRT\datasets\2002\demo.xpt |

The hyperlinks functionality in Word, which was used for the internal document links, could not be used here. After conversion to PDF the relative pathnames used were converted to absolute path names.

The example shows the PDFMARK language used to open the SAS export file containing the adverse event data:

```
%!PS-Adobe-3.0 EPSF-3.0
%%BoundingBox: 0 0 1000 1000
%%Creator: Automatic
%%Title: PDFMark
%%BeginProlog
/pdfmark where
{pop} {userdict /pdfmark /cleartomark load
put} ifelse
%%EndProlog
[ /Rect [ 0 0 1000 1000 ]
/Border [ 0 0 1 ]
/Color [ 0 1 1 ]
/Action /Launch /File (AE.xpt) /Subtype /Link
/ANN
pdfmark
```

The second type of external hyperlinks opens a PDF document on a specific page.

| Definition of Variables in Dataset **DEMO** (Demography) | | | | | |
|---|---|---|---|---|---|
| **Variable** | **Label** | **Type** | **For mat** | **Code** | **Comment** |
| PT | Subject | Num | | | Demography Page 10 |

The PDFMARK syntax used is slightly different from simply opening a file. The example shows the PDFMARK language used to open the annotated CRF, called BLANKCRF.PDF, on page 15:

```
%!PS-Adobe-3.0 EPSF-3.0
%%BoundingBox: 0 0 1000 1000
%%Creator: Automatic
%%Title: PDFMark
%%BeginProlog
/pdfmark where
{pop} {userdict /pdfmark /cleartomark load
put} ifelse
%%EndProlog
[ /Rect [ 0 0 1000 1000 ]
/Border [ 0 0 1 ]
/Color [ 0 1 1 ]
/Action /GoToR /File (blankcrf.pdf) /Subtype
/Link /Page 15
/ANN
pdfmark
```

**EMBEDDING POSTSCRIPT FILES**

The PDFMARK language is used to define Adobe Acrobat navigation features to Acrobat Distiller. PDFMARK language can be added to documents as an encapsulated postscript (EPS) file. In this case the EPS files are nothing but a text file with PDFMARK code and the file extension EPS. It can be inserted into the document like a picture file. When the document is distilled this PDFMARK commands will become part of the PDF document.

A program automatically creates one EPS file for every page of the annotated CRF and one EPS file for every dataset in the data library. As a naming convention it was helpful to add the page number to the name of the EPS files for the annotated CRF so that our file names were PAGE1.EPS, PAGE2.EPS... If links to additional documents are needed the EPS files can easily be modified manually, e.g. SAP.EPS. The EPS files for the datasets were named AE.EPS, CONMED.EPS, DEMO.EPS…

This following RTF syntax is used to insert the EPS files as a picture into the cell of the table:

```
{\shp{\*\shpinst\shptop&top.\shpbottom&bottom
.\shpleft&startbox.\shpright&endbox.\shpfhdr0
\shpbxcolumn\shpbxignore\shpbypara\shpbyignor
e\shpwr3\shpwrk0\shpfblwtxt1\shpz0{\sp{\sn
shapeType}{\sv 75}}{\sp{\sn pib}{\sv
{\pict\wmetafile8}}}{\sp{\sn pibName}{\sv
&defpath.&fname..eps}}{\sp{\sn pibFlags}{\sv
10}}{\sp{\sn pibPrintFlags}{\sv 10}}{\sp{\sn
fLine}{\sv 1}}{\sp{\sn fEditedWrap}{\sv
0}}{\sp{\sn fBehindDocument}{\sv 0}}{\sp{\sn
fLayoutInCell}{\sv 1}}}{\shprslt\ql{\pict}}}
```

The inserted file appears as a box on top of the text in the document. The dimensions of the box represent the area of the link later in the PDF document. The RTF commands \shptop, \shpbottom, \shpleft and \shpright define the size and location of the box within the table cell.

Explanation of the macro variables used:

| SAS macro variable | Description |
|---|---|
| TOP | Vertical start relative to top border of cell |
| BOTTOM | Vertical end relative to top border of cell |
| STARTBOX | Start relative to left border of cell |
| ENDBOX | End relative to left border of cell |
| DEFPATH | Path location of EPS file |
| FNAME | Name of EPS file |

The difference between top and bottom depends on the size of the font and the space between lines if the box covers multiple lines.

The values of macro variables STARTBOX and ENDBOX have to be manually determined once for every unique link. The DEFINE document is printed and the length of the text, which is supposed to be converted into hyperlinks, is measured. Afterwards macro calls are added to the EXCEL spreadsheet with the dimensions of the boxes representing the EPS files. In the following example the macro call %ZLINK is added for the concomitant medication variables:

| | A | B |
|---|---|---|
| 1 | **MEMNAME** | **COMMENT** |
| 3 | AE | Adverse event Pages 74 and 89 |
| 4 | CONMED | {\cf2 Concomitant Medication / Prior Therapy Pages 69 and 85}%zlink(page81,0,ENDBOX1,1)%zlink(page81,0,ENDBOX2,2) %zlink(page101,STARTBOX,ENDBOX3,2) |
| 5 | DEMO | Demography Page 10 |

The text ‘Concomitant Medication / Prior Therapy Pages 69 and 85’ is enclosed in parenthesis and \cf2 is added. This ensures that the text will appear in blue.

The macro call %ZLINK consists of four parameters:

- Name of EPS file
- Left border of box coordinate
- Right border of box coordinate
- Line

Because the width of lines is constant the top and bottom coordinates for the hyperlink can be calculated within the %ZLINK macro using the line information.

In the example three boxes are created, which will be hyperlinks in the PDF document. The first part of the link ‘Concomitant Medication / Prior Therapy Pages 69’ will point to the physical page 81 of the annotated CRF and is in spread over two lines. The second part ‘and 85’ pointes to page 101. Values ENDBOX1 – ENDBOX3 are the coordinates of the right borders of the boxes depending on the font size used. Here STARTBOX is equal to ENDBOX2 plus an offset.

The macro %ZLINK adds the RTF syntax to the text documented in the EXCEL spreadsheet. After importing the EXCEL

spreadsheet into SAS datasets the macro calls are resolved using the RESOLVE function in SAS.

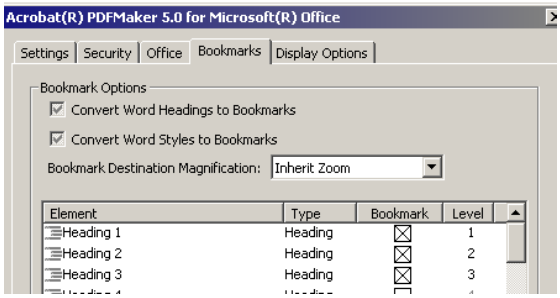
### CONVERSION OF RTF TO PDF

Once the RTF file is created Microsoft Word is used to open the document. The embedded Postscript files are visible as boxes sitting on top of the text in the specified size.

| Study 2002: Definition of Variables in Datase | | | | Adverse events |
|---|---|---|---|---|
| **Variable** | **Label** | **Type** | **Format** | **Codes** |
| PT | Subject | Num | | |
| VISIT | Visit | Num | | |

With the full version of Adobe Acrobat installed the document can be distilled to a PDF file. The PDFMARK information that is stored in the EPS files is converted to hyperlinks.

Modifying the Acrobat Distiller options panel in Microsoft Word determines which Word Headings and Styles are converted to PDF bookmarks.




In this example Heading1 to Heading3 are used for creating hierarchically structured bookmarks.

### HARD- AND SOFTWARE REQUIREMENTS

The following programs were used under Windows 2000:

- SAS 8.2 (server)
- Adobe Acrobat 5.0
- Microsoft Word 2000
- Microsoft EXCEL 2000

In the Aventis Behring hardware and software environment Adobe Acrobat had to be installed locally. When installed on the server the process of distilling the document did not function correctly. The boxes for the hyperlinks were misplaced.

Except for SAS, which runs on a server, a PC workstation with 1.6 GHz and 252 MB RAM was used. It took about 15 minutes to convert an RTF file with over 300 pages and thousands of embedded EPS files and no problems were observed. But hardware performance could become an issue with older PCs.

## CONCLUSION

When we started exploring the automatization of creating the DEFINE document it was not clear what would be possible. This challenge remained throughout the process. Finding the bits and pieces of appropriate technology was the biggest obstacle. But once this technology was developed it was very easy to adapt.

The advantages:

- No manual editing of the document necessary
- Changes can be tracked
- Full control over hyperlinks
- Consistent formatting
- Easy to implement structural changes (e.g. change of column width)
- Consistency between the datasets and the document
- Fast re-publication of the document in case of last minute changes

The disadvantages:

- The somewhat anachronistic measuring of the size of the hyperlinks
- Complex structure of PDFMARK and RTF language
- Difficult to find technical documentation

## REFERENCES


Automate the Creation and Manipulation of Word Processor Ready SAS Output
Izabella Peszek and Robert Peszek
Including Jianmin Long's %RTF macro
http://support.sas.com/documentation/periodicals/obs/obswww13/index.html

An RTF dictionary can be found in Microsoft's MSDN Library under Office Solutions Development
http://msdn.microsoft.com/library

SAS OnlineDoc® documentation

The key to creating links with PDFMARK language was found on the following site: http://www.rdpslides.com/psfaq/FAQ00007.htm

pdfmark Reference Manual
Technical Note #5150
Adobe Systems Incorporated


## ACKNOWLEDGMENTS


The author would like to thank Monika Kawohl from Aventis Behring for her valuable help and input.


## CONTACT INFORMATION

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


Dirk Spruck
Aventis Behring
P.O. Box 1230
35002 Marburg
Germany
Dirk.Spruck@aventis.com