

Bookmarking CRFs More Efficiently

Noory Kim, Statistics & Data Corporation

ABSTRACT

Preparing an annotated CRF (aCRF) entails adding PDF bookmarks. If done manually, this can be very time-consuming. This paper will show how to automate aspects of this process, including the specification of bookmark properties required by the FDA or CDISC (namely, nested levels and inherit zoom magnification). This paper assumes the reader knows how to edit the Windows PATH environment variable and use a Windows command line.

INTRODUCTION

The annotated case report form (aCRF) is a required component of the clinical data submission package sent to the FDA (Food and Drug Administration). It is a PDF (Portable Document Format) document bundled with the SDTM (Study Data Tabulation Model) data sets in the package. The aCRF requires two sets of bookmarks, by visit and by topic (CRF form).

Manually bookmarking an aCRF can be time-consuming within a PDF editor (such as Adobe Acrobat). Also, if the aCRF is updated, this may require having to add bookmarks again from scratch, as there may not be a clean and simple way to transfer bookmarks from one PDF file to another. This paper describes a semi-automated method which reduces the manual effort needed to bookmark an aCRF.

While bookmarking by SDTM domain has been unofficially proposed (Stoll, 2020), the CDISC documentation instead specifies bookmarking by topic. (CDISC, 2021) While bookmarking by topic has the disadvantage of making it harder for the SDTM reviewer to look up annotations for a specific SDTM domain, it has the advantage of being easier for the programmer to automate.

TRANSFERRING BOOKMARKING DATA INTO A PDF FILE

STARTING WITH AN EXCEL SPREADSHEET

For the sake of simplicity, my starting point will be an Excel spreadsheet. It will have one row per visit and topic (CRF form), and the following columns:

Column Name	VISITSEQ	VISIT	FORMNAME	PAGENUM
Explanation	To specify the chronological order of visits		CRF form name (a.k.a. 'topic')	Page Number
Sample Value	-21	Screening	Vital Signs	1

Table 1: Columns needed in the input spreadsheet for bookmarking by visit.

The spreadsheet needs to have a separate row for each VISIT-FORMNAME combination. The macros in the Appendix will generate a bookmark for each row in the spreadsheet.

For CRF pages not specific to any visit, CDISC specifies that these pages be presented last with the bookmark label "Running Records". Accordingly, we can set VISIT equal to "Running Records" and a VISITSEQ value higher than any other value.

A more sophisticated and complex process starts with the Study Data Specification (SDS), which is beyond the scope of this paper. The SDS approach relies on access to the EDC metadata to retrieve the page numbers of CRF forms, and so may not be a viable route for contracted studies where a CRO is only tasked with statistical programming and has no access to the EDC metadata.

Bookmarking by Visit

We can order by VISITSEQ to meet the following specification by CDISC:

“Bookmarks by chronology should be ordered according to the study schedule of activities (SOA).” (CDISC SDS MSG Team, 2021)

After sorting by VISITSEQ, we can sort CRF pages either alphabetically (FORMNAME) or by order of appearance in the CRF (PAGENUM).

Bookmarking by Topic (CRF Form)

We can sort topics either alphabetically or by order of appearance in the CRF. Either way, “Within each topic bookmarks should be ordered chronologically...” (CDISC SDS MSG Team, 2021)

If we choose to sort alphabetically, we can do so with FORMNAME and VISITSEQ. If instead we choose to sort by order of appearance, we can do so with PAGENUM and VISITSEQ.

USING SAS® TO GET FROM EXCEL TO PDFMARK

Using the macros in the Appendix, we can take the bookmark information in the spreadsheet to generate Postscript files that look something like this:

```
[ /PageMode /UseOutlines /Page 1 /View [/Fit] /DOCVIEW pdfmark
[ /Page 1 /View [/XYZ null null 0] /Title (Visit) /Count 9 /OUT pdfmark
[ /Page 5 /View [/XYZ null null 0] /Title (Screening ) /Count 45 /OUT
pdfmark
[ /Page 1 /View [/XYZ null null 0] /Title (Administration of Study Drug )
/OUT pdfmark
[ /Page 4 /View [/XYZ null null 0] /Title (Adverse Event ) /OUT pdfmark
[ /Page 8 /View [/XYZ null null 0] /Title (COVID-19 Testing and Diagnosis )
/OUT pdfmark
[ /Page 9 /View [/XYZ null null 0] /Title (COVID-19 Vaccination ) /OUT
pdfmark
[ /Page 10 /View [/XYZ null null 0] /Title (Demographics ) /OUT pdfmark
[ /Page 13 /View [/XYZ null null 0] /Title (ECG (12-Lead) Log ) /OUT pdfmark
[ /Page 14 /View [/XYZ null null 0] /Title (Electrocardiogram ) /OUT pdfmark
...
```

Articles by programmer Chris Kent (2017b, 2017c) are a very helpful guide to PDFMARK syntax, which includes the following features:

- Nesting: The number following /Count specifies the number of ‘child’ bookmarks that a bookmark has as a ‘parent’. The child bookmarks are listed directly after the parent bookmark. Multiple nesting levels are allowed.
- Inherit Zoom, which is required by the FDA (2016), is specified by the following syntax:

```
/View [/XYZ null null 0]
```

USING GHOSTSCRIPT TO GET FROM PDFMARK TO PDF

Using Ghostscript first requires that we download it and add it to the Windows environment PATH variable. For more details, see Chris Kent's article (2017a) or use a search engine to look up something like "add ghostscript to windows path". This only needs to be done one time on the computer you use.

After you have done this, you can use a Windows command line or Windows PowerShell to push the PDFMARK bookmark data into the aCRF (PDF file). Here are the steps using PowerShell.

1. [Windows Explorer] Save the PDFMARK bookmarks (PostScript file) in the same folder as the PDF file to which the bookmarks will be added.
2. [Windows PowerShell] Go to that folder using the cd (change directory) command:

Example:

```
>cd 'P:\Projects\[sponsor]\[protocol number]\Statistics\aCRF'
```

3. [Windows PowerShell] Run Ghostscript commands.

If we have separate Postscript files for bookmarking by visit and bookmarking by topic, then we can run a command for each sequentially.

Example: Say we have the Postscript files *bm_by_visit.ps* and *bm_by_topic.ps*. We can add these bookmarks one file at a time as follows.

```
>gswin64c -o tempfile.pdf -sDEVICE=pdfwrite -dPDFSETTINGS=/prepress  
acrf_without_bookmarks.pdf bm_by_visit.ps
```

```
>gswin64c -o acrf_with_bookmarks.pdf -sDEVICE=pdfwrite  
-dPDFSETTINGS=/prepress tempfile.pdf bm_by_topic.ps
```

Alternatively, you can also opt to concatenate the Postscript files in a text editor, then add all bookmarks with a single Ghostscript command.

```
>gswin64c -o acrf_with_bookmarks.pdf -sDEVICE=pdfwrite  
-dPDFSETTINGS=/prepress acrf_without_bookmarks.pdf  
bm_by_visit_and_by_topic.ps
```

The generalized Ghostscript syntax is as follows:

```
>gswin64c -o [output_PDF_filename] -sDEVICE=pdfwrite  
-dPDFSETTINGS=/prepress [original_PDF_filename] [pdfmark_filename]
```

CONCLUSION

This paper covered the following:

- Bookmarking by visits (chronologically) and by topic (CRF form), as required by CDISC
- When using the Study Data Specification (SDS) may be not feasible
- Using SAS to generate Postscript files with PDFMARK syntax needed to add bookmarks with nested levels and inherit zoom
- Using Ghostscript to add PDFMARK bookmarks to a PDF file

The process presented in this paper can reduce the effort it takes to bookmark an aCRF. Most of the work is done preparing the bookmarking data in a spreadsheet, where it is arguably easier to organize bookmarking data, rather than within the PDF file itself. The macros provided in the Appendix automatically specify nesting levels and inherit zoom, further reducing the manual effort and time required to complete this task.

REFERENCES

- Adobe Systems Inc. (2020) *Adobe pdfMark Reference*. 27 January 2021.
https://opensource.adobe.com/dc-acrobat-sdk-docs/acrobatsdk/pdfs/acrobatsdk_pdfmark.pdf
- CDISC SDS MSG Team. (2021) "Study Data Tabulation Model Metadata Submission Guidelines (SDTM-MSG): Human Clinical Trials." Version 2.0 (Final), 2021-03-30.
https://wiki.cdisc.org/download/attachments/129041841/SDTM-MSG_v2.0.pdf?version=1&modificationDate=1632255228956&api=v2
- FDA. (2016) "Portable Document Format (PDF) Specifications." v4.1 (September 2016)
<https://www.fda.gov/media/76797/download>
- Joy, Geo and Couturier, Andre. (2015) "SDTM Annotations: Automation by implementing a standard process." PharmaSUG 2015, paper AD07.
<https://www.pharmasug.org/proceedings/2015/AD/PharmaSUG-2015-AD07.pdf>
- Kent, Chris. (2017a) "Applying pdfmark To PDF Documents Using GhostScript."
<https://thechriskent.com/2017/02/13/applying-pdfmark-to-pdf-documents-using-ghostscript>
- Kent, Chris. (2017b) "Setting PDF View Options with pdfmark."
<https://thechriskent.com/2017/03/06/setting-pdf-view-options-with-pdfmark>
- Kent, Chris. (2017c) "Adding Bookmarks to PDF Documents with pdfmark."
<https://thechriskent.com/2017/04/12/adding-bookmarks-to-pdf-documents-with-pdfmark>
- Mead, Lynn. (2011) "Cooking up Enhanced PDF with pdfmark Recipes."
www.meadowmead.com/wp-content/uploads/2011/04/PDFMarkRecipes.pdf
- Stoll, Markus et. al. (2020) "SDTM aCRF Guideline - Guideline for SDTM annotations in Case Report Forms, Summary and Recommendations for Best Practice." V1.0, 2020-11-20.
https://wiki.cdisc.org/download/attachments/113589261/aCRF_Guideline_v1-0_20201120_publish.pdf?version=1&modificationDate=1606251671510&api=v2

ACKNOWLEDGMENTS

Thanks to Adam Hamm, Leslie Schneider, and Bharath Donthi for feedback on abstract drafts. Thanks to Chambers Volk for explaining aspects of the Study Data Specification (SDS) approach. All errors remain the responsibility of the author.

CONTACT INFORMATION

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

Noory Kim
nkim@sdclinical.com

Any brand and product names are trademarks of their respective companies.

APPENDIX

BOOKMARKING BY VISIT

The following SAS macro outputs a Postscript file with PDFMARK syntax for bookmarks by visit.

```
%macro xlsx_to_bookmarks_by_visit(input_xlsx=, output_ps=bm_by_visit, formsortvar=formname);
  * author      : Noory Kim;
  * version     : 0.1 (04APR2022);
  * parameters;
  * input_xlsx  : name of input file (Excel XLSX) with bookmark data;
  * output_ps   : name of output file (PostScript) with PDFMARK bookmarks;
  * default value: bm_by_visit;
  * formsortvar : sort topics alphabetically (=formname) or in order of appearance (=pagenum)
  * default value: formname;

  * import source data;
  proc import file="&input_xlsx..xlsx" out=_010 dbms=xlsx replace;
  run;

  * get counts for nesting bookmarks ;
  proc sql;
    create table _020 as
    select *
      , count(*) as count
      , put(count(*), 8.) as countc
    from _010
    group by visitseq
    order by visitseq, &FORMSORTVAR.
  ;
  quit;

  * get number of visits;
  proc sql;
    create table _030 as
    select *
      , count(unique visitseq) as n_visit
    from _020
  ;
  quit;

  * sort source data;
  proc sort data=_030 out=_077;
    by visitseq &FORMSORTVAR.;
  run;

  * generate and output postscript file with PDFMARK syntax;
  data _null_;
    file "&output_ps..ps" ;
    set _077;
    by visitseq;

    if _n_ = 1 then do;
      put '[ /PageMode /UseOutlines /Page 1 /View [/Fit] /DOCVIEW pdfmark';
      put '[ /Page 1 /View [/XYZ null null 0] /Title (Visit) /Count ' n_visit ' /OUT pdfmark';
    end;

    if first.visitseq = 1 then do;
      put '[ /Page ' pagenum ' /View [/XYZ null null 0] /Title (' visit ') /Count ' countc ' /OUT
pdfmark';
    end;
    put '[ /Page ' pagenum ' /View [/XYZ null null 0] /Title (' formname ') /OUT pdfmark';
  run;
%mend;

  * sample macro calls;
  %xlsx_to_bookmarks_by_visit(input_xlsx=bookmark_data, output_ps=bm_by_visit_formname,
formsortvar=formname);
  %xlsx_to_bookmarks_by_visit(input_xlsx=bookmark_data, output_ps=bm_by_visit_pagenum,
formsortvar=pagenum);
```

BOOKMARKING BY TOPIC (CRF FORM)

This macro outputs a Postscript file with PDFMARK syntax for bookmarks by topic.

```
%macro xlsx_to_bookmarks_by_topic(input_xlsx=, output_ps=bm_by_topic, formsortvar=formname);
* author      : Noory Kim;
* version     : 0.1 (04APR2022);
* parameters;
* input_xlsx  : name of input file (Excel XLSX) with bookmark data;
* output_ps   : name of output file (PostScript) with PDFMARK bookmarks;
*             : default value: bm_by_topic;
* formsortvar : sort topics alphabetically (=formname) or in order of appearance (=pagenum)
*             : default value: formname;

* import source data;
proc import file="&input_xlsx..xlsx" out=_010 dbms=xlsx replace;
run;

* get counts for nesting bookmarks ;
proc sql;
  create table _020 as
  select *
  , count(*) as count
  , put(count(*), 8.) as countc
  from _010
  group by formname
  order by formname, visitseq
  ;
quit;

* get number of visits;
proc sql;
  create table _030 as
  select *
  , count(unique formname) as n_formname
  from _020
  ;
quit;

* sort source data;
proc sort data=_030 out=_077;
  by &FORMSORTVAR. visitseq;
run;

* generate and output postscript file with PDFMARK syntax;
data _null_;
  file "&output_ps..ps" ;
  set _077;
  by &FORMSORTVAR.;

  if _n_ = 1 then do;
    put '[ /PageMode /UseOutlines /Page 1 /View [/Fit] /DOCVIEW pdfmark';
    put '[ /Page 1 /View [/XYZ null null 0] /Title (Forms) /Count ' n_formname ' /OUT pdfmark';
  end;

  if first.&FORMSORTVAR. = 1 then do;
    put '[ /Page ' pagenum ' /View [/XYZ null null 0] /Title (' formname ') /Count ' countc '
/OUT pdfmark';
  end;

  put '[ /Page ' pagenum ' /View [/XYZ null null 0] /Title (' visit ') /OUT pdfmark';
run;
%mend;

* sample macro calls;
%xlsx_to_bookmarks_by_topic(input_xlsx=bookmark_data, output_ps=bm_by_topic_formname,
formsortvar=formname);
%xlsx_to_bookmarks_by_topic(input_xlsx=bookmark_data, output_ps=bm_by_topic_pagenum,
formsortvar=pagenum);
```