



S-CUBED

AD07

Post-Processing PDFs with Ghostscript

Jonathan North

PhUSE 2018: Frankfurt

Solutions . . . Knowledge . . . Expertise ₁

Agenda

- History of the paper
- Introduction to Ghostscript
- Basics of using Ghostscript
 - Combining PDF files
 - Splitting PDF files

Agenda

- Using pdftools with Ghostscript
 - What are they, and basic syntax
 - Creating annotations
 - Creating bookmarks
 - Setting PDF document properties
- More complex use of Ghostscript

History of the paper

- Previously presented at PhUSE 2015
 - Coders Corner CC04: Automating Production of the blankcrf.pdf in a CRO Environment
 - Extending functionality of Adobe Acrobat® using JavaScript™
 - Bookmarks were created by adding JavaScript into the aCRF PDF
 - FDA does not allow JavaScript in submission PDFs
 - Portable Document Format (PDF) Specifications (September 2016, Version 4.1)

Introduction to Ghostscript

- An interpreter for the PostScript language and for PDF
- Open source
 - Owned and fully controlled by Artifex
 - Commercial support and licences are available
 - Developed for 20+ years
 - Cross platform
- Command line

Basics of using Ghostscript

The command line

```
gs [options] {filename} ...
```

Most common options

<code>-dBATCH</code>	exit once finished processing
<code>-dNOPAUSE</code>	process all pages without pausing
<code>-dQUIET</code>	suppress routine log messages

Basics of using Ghostscript

- Combining files

```
gs -sDEVICE=pdfwrite  
-sOutputFile=combined.pdf  
file1.pdf file2.pdf file3.pdf
```

Basics of using Ghostscript

- Splitting a PDF into multiple files
Extract a continuous block of pages

```
gs -sDEVICE=pdfwrite  
-dFirstPage=2 -dLastPage=3  
-sOutputFile=subset.pdf  
original.pdf
```

Basics of using Ghostscript

- Splitting a PDF into multiple files
Extract a non-continuous block of pages

```
gs -sDEVICE=pdfwrite  
  -sPageList=odd  
  -sOutputFile=subset.pdf  
  original.pdf
```

Basics of using Ghostscript

- Splitting a PDF into multiple files
Extract a non-continuous block of pages

```
gs -sDEVICE=pdfwrite  
  -sPageList=even  
  -sOutputFile=subset.pdf  
  original.pdf
```

Basics of using Ghostscript

- Splitting a PDF into multiple files
Extract a non-continuous block of pages

```
gs -sDEVICE=pdfwrite  
-sPageList=1, 3, 5-8  
-sOutputFile=subset.pdf  
original.pdf
```

Basics of using Ghostscript

- Splitting a PDF into multiple files
Extract selected pages into multiple files, one page per file

```
gs -sDEVICE=pdfwrite  
-dFirstPage=2 -dLastPage=3  
-sOutputFile=subset%d.pdf  
original.pdf
```

Using pdftools with Ghostscript

- What are pdfmarks?
 - Extension of PostScript code used to represent PDF features
 - PDF features are generated when the PostScript code is converted into PDF
 - They are written as simple text files

```
gs -sDEVICE=pdfwrite  
-sOutputFile=processed.pdf  
-dPDFSETTINGS=/prepress pdfmarks.txt  
original.pdf
```

Using pdfmarks with Ghostscript

pdfmark syntax

[	The [is used to mark the start of a new pdfmark
Arguments	describe the features for the current pdfmark
Name	the kind of pdfmark being described

An example – to define a bookmark

```
[  
/Page 1  
/Title (Bookmark Title)  
/OUT pdfmark
```

Using pdfmarks with Ghostscript

pdfmark syntax

[	The [is used to mark the start of a new pdfmark
Arguments	describe the features for the current pdfmark
Name	the kind of pdfmark being described

An example – to define a bookmark

```
[/Page 1 /Title (Bookmark Title)/OUT pdfmark
```

Using pdfmarks with Ghostscript

Creating annotations

DM=Demographics		Screening
CDISC Study: CDISC01	Assessment Date	
STUDYID	SCDTC	___/___/___
DEMOGRAPHY		
Date of Birth: ___/___/___ BIRTHDTC		
SEX		
Gender: ☐ Male ☐ Female		
ETHNIC		
Ethnicity: ☐ Hispanic or Latino ☐ Not Hispanic or Latino		
RACE		
Race: Check all that apply		
☐ White		
☐ American Indian or Alaska Native		
☐ Black or African American		
☐ Native Hawaiian or Other Pacific Islander		
☐ Asian		
☐ Other:___ RACEOTH in SUPPDM		
RACE, when more than one selected, RACE=MULTIPLE and individual responses are RACE1, RACE2, etc. in SUPPDM		
SC=Subject Characteristics		
FamilyStatus: ☐ Never Married ☐ Domestic Partner		
☐ Married ☐ Divorced		
☐ Legally Separated ☐ Widowed		
SCORRES when SCTESTCD = MARISTAT		

Using pdfmarks with Ghostscript

Creating annotations

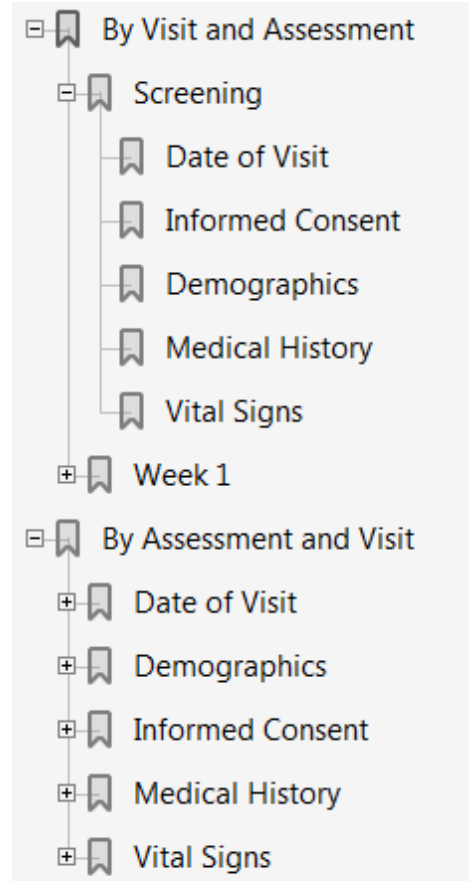
DM=Demographics		Screening
CDISC Study: CDISC01	Assessment Date	
STUDYID	SCDTC	____/____/____
DEMOGRAPHY		
Date of Birth: ____/____/____ BRTHDTC		
SEX		
Gender:	☐ Male	☐ Female
ETHNIC		
Ethnicity:	☐ Hispanic or Latino	☐ Not Hispanic or Latino
RACE		
Page:	Check all that apply	RACE when more than one selected

```
[  
/SrcPg 1  
/Rect [100 400 300 430]  
/Subtype /FreeText  
/Color [1 1 0.75]  
/DA ([0 0 0] rg /Cour 12 Tf)  
/Contents (Dataset Annotation)  
/ANN pdfmark
```

page
bounding box
type of annotation
background colour
text colour and font
annotation text

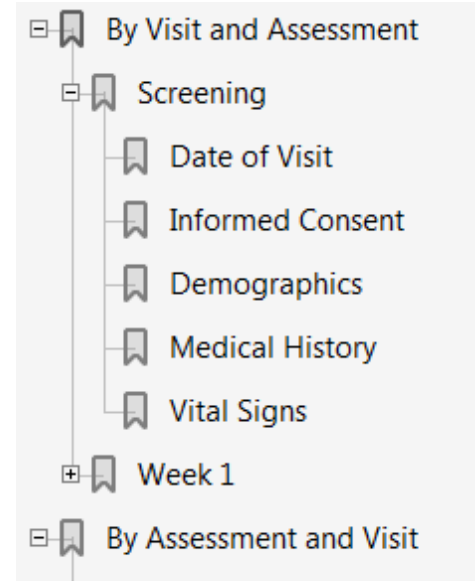
Using pdfmarks with Ghostscript

Creating bookmarks



Using pdfmarks with Ghostscript

Creating bookmarks



```
[  
/Title (By Visit and Assessment)  bookmark text  
/Count 2                          number of child bookmarks  
/Page 1                           target page  
/OUT pdfmark
```

Using pdfmarks with Ghostscript

Setting PDF document properties

- Set the view mode when the PDF is opened, e.g. single-page or continuous scrolling
- Set the initial page to show, and zoom factor
- Set document properties, e.g. title, author, key words

More complex use of Ghostscript

- Ghostscript is an interpreter of the PostScript language
- PostScript is a language originally created by Adobe in 1982, and is designed to allow control of printers
- Ghostscript can combine PostScript code with PDFs for additional processing
 - Adding a watermark to all pages in an existing PDF

Concluding thoughts

- Ghostscript is a powerful command line tool for manipulating PDFs
- It can be integrated into a workflow process to automate and ease combining files, and allow for setting document properties
- It can be used to help create aCRF.pdf

Comments / Questions / Suggestions?