aCRF Preparation through Automated R-Driven Annotation

Priyadharshini Rajasekaran, Zifo, Chennai, Tamil Nadu, India
Vaishali Govindarajan, Zifo, Chennai, Tamil Nadu, India

BENGALURU, INDIA PharmaSUG SDE APRIL 11, 2026

Abstract

Annotated Case Report Form (aCRF) is one of the most essential document for regulatory submission and clinical data management. Manually annotating and bookmarking them in accordance with criteria is difficult, time-consuming and prone to errors.

To alleviate the tedious manual tasks associated with aCRF annotation and aesthetics handling, we propose an automation using an R tool. By outlining the subsequent steps for completing aCRF annotation, this tool has streamlined the process, impacting productivity and accuracy, which makes programmers focus on essential tasks.

Introduction

In the world of clinical trials adherence to Metadata Submission Guidelines (MSG) is paramount before submitting data to regulatory bodies.

Annotated Case Report Form (aCRF) is vital in any clinical trial to map the protocol required participant data collected in the Case Report Form (CRF) to the corresponding variables or discrete variable values contained within the SDTM datasets.

Manually annotating all the unique CRF pages by checking the different variables to be mapped into each domains by keeping consistency in font, appearance, giving colour codes for variables and bookmarking them in accordance with criteria is quite time consuming and can lead to human made errors.

The aCRF automation tool was configured using R software to tackle the difficulties encountered while manually annotating the CRF. The tool develops an editable PDF with annotations and bookmarking, reducing the manual work leaving only minimal drag-and-drop adjustments to finalize annotations quickly and efficiently.

Overview

The aCRF automation tool is designed in such a way that it complies with the regulatory standards utilizing various visual cues such as color-coded boxes, font sizes to denote different domains and variables. The tool is operated through a simple interface.

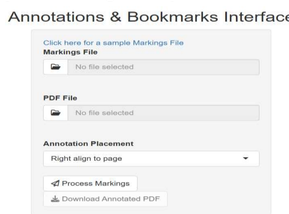


Figure 1. Interface of aCRF automation tool

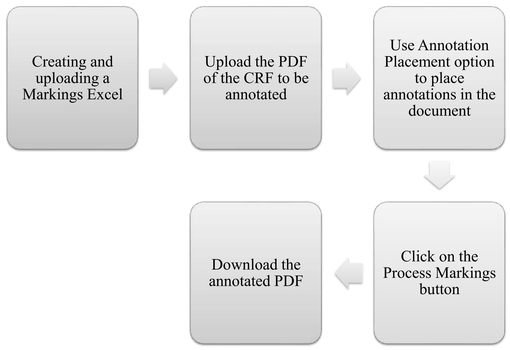
The fluidPage function in the library R shiny is used to create the web page layout (interface). Inbuilt functions like observeEvent, downloadHandler and certain user defined functions are used for the development of the tool. The libraries used include shiny, shinyjs, dplyr, pdftools, rlang, openxlsx and stringi.

```
server <- function(input, output) {
  annotations <- data.frame()
  scheduleOrig <- data.frame()
  pageOrig <- data.frame()
  pagesCount <- 0
  pdfdata <- list()
  pdfpagesize <- data.frame()
  COLORS <- c('0.75 1.0 1.0', '1.0 1.0 0.66', '1.0 0.75 0.66', '1.0 0.66 0.75', '0.66 0
  FONT_SIZE <- 12
  rv <- reactiveValues(mrkpath = "")

  getTextDimensions <- function(text) {
    pdf(NULL)
    par(ps = 12, family = "serif")
    width <- strwidth(text, units = "inches")
    height <- strheight(text, units = "inches")
    dev.off()
    return(list(width = ceiling(width * 72) + (nchar(text)*.5) + 10, height = 9 + 6))
  }
```

Figure 2. Snip of a user defined function that initializes server-side variables and defines a helper function to calculate text dimensions for annotations.

Workflow



Markings Excel File

The markings excel file contains the information about the CRF forms, and the SDTM domains that we need each variables to be mapped into, the type of annotation boxes that need to be developed, the number of pages of the CRF etc. The sample template of the markings excel that needs to be uploaded to the tool can be obtained from the interface. The sample from the tool has self-explanatory comments explaining the requirements for each sheet in the excel.

The input mapping excel has three sheets:

SCHEDULE

- The form names in the CRF and their corresponding visits should be mapped into this sheet based on the schedule of assessment section in the protocol

FORM	Screening/Baseline	Treatment	Day 2	Day 3	Day 4	Day 5	Day
Subject Information	X						
Inclusion Criteria	X						
Exclusion Criteria	X						
Demography	X						
Medical History (Excluding Study Indication)	X						
Pregnancy	X						
Local Laboratory-Hematology	X				X	X	X
Local Laboratory-Chemistry	X				X	X	X
Vital Signs	X		X	X	X	X	X

DELETE THIS TEXT BOX BEFORE UPLOADING
Sheet name should be 'SCHEDULE'.
Visits should be in chronological order.
Add the non-chronological visits at the end.
All the forms listed here should have an entry in the 'PAGENO' sheet.

Figure 3. Schedule sheet in markings excel with help text

PAGENO

- Page number of the form names given in the schedule sheet of the markings excel should be filled here.

FORM	PAGENO
Subject Information	2
Inclusion Criteria	5
Exclusion Criteria	8
Demography	10
Medical History (Excluding Study Indication)	16
Pregnancy	18
Local Laboratory-Hematology	20
Local Laboratory-Chemistry	25
Vital Signs	43

DELETE THIS TEXT BOX BEFORE UPLOADING
Sheet name should be 'PAGENO'.
Both columns should have values.
Common forms also should be listed in this sheet.

Figure 4. Pageno sheet in markings excel with help text

ANNOTATIONS

- There are seven columns in the ANNOTATIONS sheet: DOMAIN, TEXT, TYPE, PAGE, FIELD, OCCURRENCE and PLACEMENT.

DOMAIN	TEXT	TYPE	PAGE	FIELD	OCCURE	PLACEMEN
DM	DM = Demo	1	2			
DS	DS = Dispos	1	2			
DM	SITEID	2	2			
DM	SUBJID	2	2			
DM	[NOT SUBM	2	2			
DM	RFICDTC	2	2			
DM	TIVERS in SI	2	2			
DM	PROTDTC in	2	2			
DS	DSTERM = 'I	2	2			
DS	DSCAT = 'PR	2	2			
DS	DSSTDTC	2	2			
DM	RFENDTC	2	2			
DM	RFPENDTC	2	2			
IE	IE = Inclusio	1	5			
IE	IECAT = 'INC	2	5			
IE	IEORRES	2	5			
IE	IETESTCD =	2	5			
IE	IETEST = Ho	2	5			
IE	IETESTCD =	2	5			
IE	IETEST = Sul	2	5			
IE	IETESTCD =	2	5			

DELETE THIS TEXT BOX BEFORE UPLOADING
TYPE = 1 for Header annotations
TYPE = 2 for Item annotations
TYPE = 3 for extra (dashed-border)

- First four columns should have value

FIELD column
- FIELD column is optional. If FIELD column is filled, then the annotation will be as per the placement chosen.

OCCURENCE column
- If FIELD column is filled, then OCCURENCE column can be used to specify the occurence of the field in the CRF page incase the same field appears more than once in the same CRF page. If OCCURENCE is left blank, then by default the first occurence will be considered
NOTE: Filling OCCURENCE without filling FIELD will be ignored.

PLACEMENT column
- If FIELD column is filled, then PLACEMENT column can be used to specify the position of the annotation. This field level placement will override the document level placement (chosen in the UI)
NOTE: Filling PLACEMENT without filling FIELD will be ignored.

Figure 5. Annotations sheet in markings excel with help text

Columns in Annotations sheet explained

COLUMNS	DETAILS
DOMAIN	The SDTM Domain name for which the CRF page should be annotated will be mentioned here.
TEXT	The text will be the annotations, that must be displayed in the corresponding boxes of aCRF. Length of the text should not exceed 200 characters. Eg: "PROTDTC in SUPPDM"
TYPE	Type is to mention the appearance of the annotations. TYPE = 1 for Header annotations TYPE = 2 for Item annotations TYPE = 3 for extra (dashed-border)
PAGE	After running the macro, the annotations will be appearing on the respective pages of the CRF mentioned in the csv. The page number mentioned in the PAGENO sheet and this sheet should be the same.
FIELD	Mention the text of the field that is designed to collect the data in the CRF to be annotated.
OCCURENCE	Specify the occurrence of the field in the CRF page incase the same field appears more than once in the same CRF page. If OCCURENCE is left blank, then by default the first occurrence will be considered.
PLACEMENT	The position of the annotation will be based on the field column and the placement mentioned in this column. Eg., Override field – Annotation will be placed on top of the field text

ADVANTAGES
Improves accuracy and consistency
Reduces time for annotation and the programmer can focus on other tasks
Reduces manual errors
Bookmarking is done inside the tool itself. Hence it does not call for the need for manual bookmarking
Can easily annotate CRFs with huge number of pages
Can easily be used again for different version of CRF and reused across various studies

PITFALLS
Reduces efforts but does not eliminate review and quality checks
PDF of CRF should be text based, image based CRF cannot be uploaded
Cannot annotate documents with blank pages
The fields filled in the excel should be consistent with the pages in the CRF, else the tool would throw the following error as "argument is of length zero"

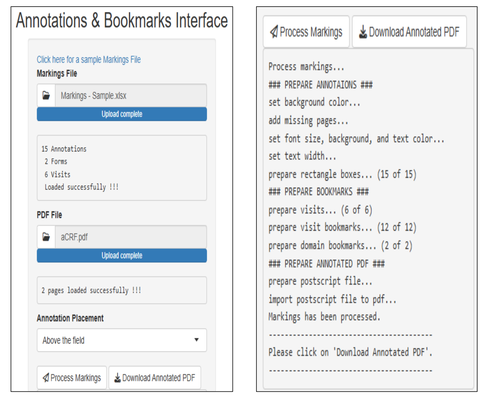


Figure 6. Interface of aCRF automation tool after we upload the documents and upload is completed

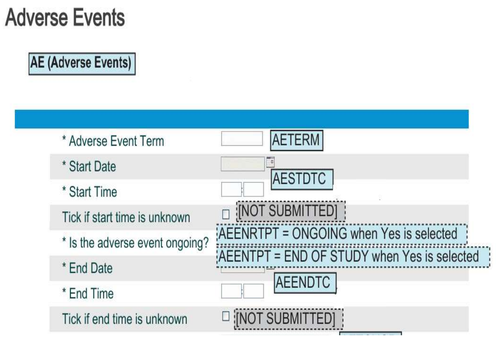


Figure7. Sample of aCRF after annotation

Conclusion

The aCRF automation tool complies with regulatory standards allowing programmers only to focus on identifying SDTM domains and variables, providing this information via simple Excel input to the tool. It generates an aCRF with 85% completeness which significantly enhances efficiency in clinical data workflows. By providing a clearly defined sequence of steps for completing CRF annotation, the tool standardizes the workflow, reduces process variability, and improves overall efficiency and accuracy, allowing programmers to concentrate on higher-value development tasks.

References

•shiny.posit.co/r
•dplyr.tidyverse.org
•stringi.gagolewski.com
•rlang.r-lib.org

Contact Information

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

Name: Priyadharshini Rajasekaran
Company: Zifo
E-mail: Priyadharshini.R@zifornd.com
Web: LinkedIn