PharmaSUG 2025 - Paper AP-213

Programmatic Annotation of Case Report Forms

Matthew Finnemeyer, Vertex Pharmaceuticals, Inc.;

ABSTRACT

Annotation of Case Report Forms (CRFs) is an essential yet time-consuming task; manual creation or import of annotations as Comments to a PDF can be tedious and error prone. This paper outlines a robust macro that: imports pre-existing PDF annotations into SAS, parses key information for user review and modification, "up-versions" annotations to conform with the latest standards, aligns annotations with their "target page" in the Destination CRF, and exports the finalized annotations to a Forms Data Format (.FDF) file for direct import by the user. This macro has reduced internal programmer-hours for CRF annotation by more than 50% when creating new or modified aCRFs.

INTRODUCTION

For any human-subject trial involving a therapy, it is expected that at least some subject data will be collected by a site/investigator; these data are typically collected on paper or electronic forms that collectively serve as the CRF. For CDISC-compliance, all mapping of raw data to SDTM datasets must be identified through annotations on the CRF, typically using the Comment tool on PDF editing software, for every single page of the CRF.

Depending on the scope of the study, a CRF may range in size from a few dozen pages to several hundred, necessitating many hours of precise, repetitive, and tedious human effort to generate properly formatted (size, spacing, font) annotations. Performing this process entirely by hand risks human error (e.g. typographical, omissions), so an alternative option is to extract and import annotations from a previously completed annotated CRF (aCRF). This process can also prove challenging, as imported annotations are mapped to the "destination" PDF based on the "source" PDF's page numbers. As a result, any deviation in the page order or page count between the "source" and "destination" PDFs can cause annotations to appear on the incorrect page (or be lost altogether). This inflexibility can result in additional time spent moving annotations to their proper pages and ensuring no annotations have been lost (or, in rare cases, duplicated) by the import process.

OUTLINING A SOLUTION

Previously (Hufford) proposed a process for streamlining/automating the annotation of CRFs, by (1) generating a SAS database of annotations through the import of the annotation file (FDF), followed by (2) manual indexing of these annotations within a permanent dataset to allow for (3) their selective export to an FDF file for import into the "destination" CRF. While this proposal is inventive, it creates the burden of ongoing maintenance of a SAS-based annotation "repository," especially the manual categorization of annotations to their target CRF Form (e.g. Adverse Events, Informed Consent, Drug Administration, etc.). Additionally, when a user wishes to extract groups of annotations from this repository, they are required to manually map the target pages, for each CRF Form, to ensure proper import into the "destination" CRF.

Building on the recommendations of Hufford, this paper will present an alternative to the permanent annotation repository, instead using an open-source utility, XDP, to facilitate the direct mapping of annotations from one or more "source" aCRFs to a "destination" aCRF, through the import, modification, and export of FDF files (Figure 1).

XPDF FOR PDF PROCESSING

An open-source software, found https://www.xpdfreader.com/index.html, XPDF includes several command line tools that allow for the processing of PDF files as alternative file formats. Using the "x" command in SAS, these utilities can render the PDF into a format (e.g. HTML) that can be read into a SAS dataset and parsed for key information, such as CRF Form Page Names, Form IDs, and Page Numbers; examples of this parsing will be shown later in this paper.

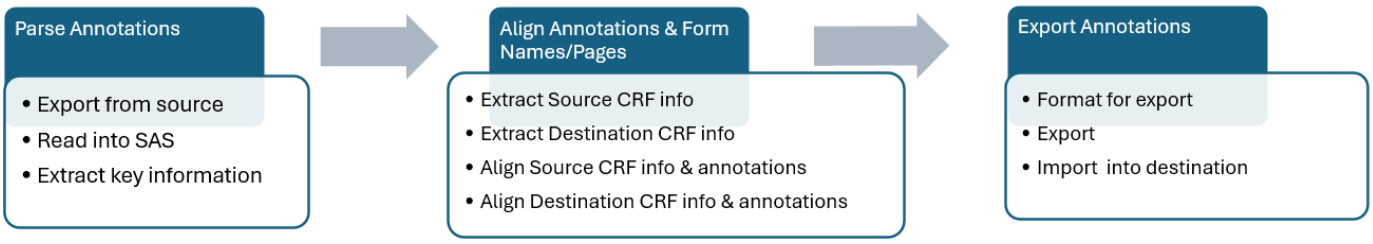

Figure 1

Figure 1. Process Flow Diagram

PARSE ANNOTATIONS

To facilitate the selection of annotations from the Source CRF, for eventual import into the Destination CRF, the user must first transform the annotation into a programmatically parsable form; once transformed, key information can be extracted from the annotation record to assist in the selection process.

EXPORT ANNOTATION FROM SOURCE

To initiate this process, the user must first extract the annotations from the "source" CRF as an FDF file—this process may vary per software used but is typically associated with the Comments tool; it is important to note that the XFDF (XML Forms Data Format) file format is not sufficient for this process, as key annotation information, like color and location, are not retained in this output file type.

The FDF file is composed of three critical parts, hereafter referred to as the "header", "body", and "footer". The "header" delineates the precise count of annotations within the FDF file-- the excerpts Figure 2 and Figure 3 show the beginning and end of the "header" for an aCRF with 1252 individual annotations—while the footer (Figure 4) serves as a standardized piece of text to indicate the file's end. The "body" contains all the essential information for each individual annotation; it is recommended the user first focus on extracting the individual annotations from the "body" and storing them as individual SAS records, as later modifications to the "header" are dependent on the final count of exported annotations and should be reserved for the "Export Annotations" stage of the process.

```
%FDF-1.2
%âãÏÓ
1 0 obj
<</FDF<</Annots[2 0 R 3 0 R 4 0 R 5 0 R 6 0 R 7 0 R 8 0 R 9 0 R
10 0 R 11 0 R 12 0 R 13 0 R 14 0 R 15 0 R 16 0 R 17 0 R 18 0 R 19
0 R 20 0 R 21 0 R 22 0 R 23 0 R 24 0 R 25 0 R 26 0 R 27 0 R 28 0
```

Figure 2. Start of FDF Header

[..]

```
1251 0 R 1252 0 R 1253 0
R]/F(/bos1.vrtx.com/SAS/e_share/macrolib/Development/Automate
Annotate/Vertex_Template_Study_v2_DEV1_1
_Unique_CRF_annotation_FINAL.pdf)/ID[<27E92FDD817D9E633C3711D9CEF
01D66>
<2390CD88364E0645A2F0142BAACC8638>]/UF(/bos1.vrtx.com/SAS/e_share
/macrolib/Development/Automate Annotate/Vertex_Template_Study_v2
_DEV1_1_Unique_CRF_annotation_FINAL.pdf)>>/Type/Catalog>>
endobj
```

Figure 3. End of FDF Header

```
trailer<</Root 1 0 R>>%%EOF
```

Figure 4. FDF Footer

As shown in Figure 5, an annotation within the “body” of the FDF file is a complex tangle of information that encodes key style elements, including font size, color, weight, style, type (e.g. Arial), location and dimensions of the “word box”, border style and background color of the “word box”, as well as the text itself; each annotation begins with a string of “X 0 obj” and almost universally ends with “endobj”.

```
1250 0 obj
<</C[0.749023 1.0 1.0]/Contents(VSORRES / VSORRESU when VSTESTCD
= PULSE)/CreationDate(D:000000000000000Z)/DA(0 0 0 rg
/arial,italic 9 Tf)/DS(font: italic arial 9.0pt; text-align:left;
color:#000000 )/F
4/M(D:20220715144758-04'00')/NM(a98694aa-4fd2-4ad7-9409-
e1ec1e2d3f16)/Page 230/RC(<?xml version="1.0"?><body
xmlns="http://www.w3.org/1999/xhtml"
xmlns:xfa="http://www.xfa.org/schema/xfa-data/1.0/"
xfa:APIVersion="Acrobat:22.1.0" xfa:spec="2.0.2"  style="font-
size:9.0pt;text-align:left;color:#000000;font-weight:normal;font-
style:italic\
;font-family:arial;font-stretch:normal"><p dir="ltr">VSORRES<span
style="font-style:normal"> / </span>VSORRESU<span style="font-
style:normal"> when VSTESTCD = PULSE</span></p>
</body>)/Rect[136.698 136.148 355.028
149.184]/Subj(VS)/Subtype/FreeText/T(VS_WT_OXSAT)/Type/Annot>>
endobj
```

Figure 5. Example Annotation from FDF Body

READ ANNOTATION INTO SAS

The FDF file can be read into SAS (e.g. with a INFILE statement) and the “body” processed, using the “X 0 obj” and “endobj” character strings to split each annotation into an individual record.

Once the FDF file has been separated into individual annotation records, it is recommended that each record be processed as follows:

Key style elements {e.g. font size, color, background color, etc.} are extracted as distinct variables (Figure 6) and their values within the annotation replaced with placeholder values (Figure 7), as this processing facilitates user review and the ability to directly modify key values for one or more annotations programmatically;

The page number of the annotation is extracted for each record (and replaced with a placeholder);

The annotation’s identification number is replaced with placeholders at (a) the start of the annotation (e.g. the numeric value preceding the “0 obj”) and (b) when an annotation contains the string “Parent”, indicating it is a “child” annotation (Figure 8) tied to the previous annotation;

If an annotation represents an Origin = “Assigned-Sponsor” variable, MSG 2.0 guidelines dictate that a dashed border be used. In the FDF file, dashed borders are identified by the string “/BS” at the start of the annotation (Figure 9 and Figure 10) and it is recommended that these records be flagged for additional processing during the “Export annotations” stage.

CONTENT	BACKCOLOR_D	BACKCOLOR	BORDER	FONT	FONT_SIZE	FONT_STYLE	FONT_WEIGHT	FONTCOLOR
DM \(Demographics\)	BLUE	[0.749023 1.0 1.0]	Solid	arial	9.0pt	normal	bold	#000000
ANGIRST in SUPPDM	BLUE	[0.749023 1.0 1.0]	Solid	arial	9.0pt	italic	normal	#000000
ANGIRYN in SUPPDM	BLUE	[0.749023 1.0 1.0]	Solid	arial	9.0pt	italic	normal	#000000
ANGIR4YN in SUPP...	BLUE	[0.749023 1.0 1.0]	Solid	arial	9.0pt	italic	normal	#000000
AE \(Adverse Events\)	BLUE	[0.749023 1.0 1.0]	Solid	arial	9.0pt	normal	bold	#000000

Figure 6. SAS records representing individual annotations

ANNOT_STORE
[XX] 0 obj<</C[BACKCOLOR]/Contents(AE \(Adverse Events\))/CreationDate(D:20140815104812-04'00')/DA(0 0 0 rg /Helv 10 Tf)/DS(font: [FONT] [FONTSIZE]; text-align:left; color:[FONTCOLOR])/F 4/M(D:20230123110830-08'00')/NM(381f293e-281f-4dff-942c-bd2b7dde8fc2)/Page [Z]/RC(<?xml version="1.0"?><body xmlns="http://www.w3.org/1999/xhtml" xmlns:xfa="http://www.xfa.org/schema/xfa-data/1.0/" xfa:APIVersion="Acrobat:22.3.0" xfa:spec="2.0.2" style="font-size:[FONTSIZE];text-align:left;color:[FONTCOLOR];font-weight:[FONTWEIGHT];font-style:normal;font-family:[FONT];font-stretch:normal"> AE \(Adverse Events\) </body>)/Rect[84.8411 705.517 188.705 720.788]/Subj(Text Box)/Subtype/FreeText/T([CRFPAGENAME])/Type/Annot>>endobj

Figure 7. Full annotation as SAS variable with key information replaced with [PLACEHOLDER] values

```
500 0 obj
<</F 28/Open false/Page 129/Parent 499 0 R/Rect[200.975 672
380.975 792]/Subtype/Popup/Type/Annot>>
endobj
```

Figure 8. "Child" annotation, from FDF, showing the Parent ID number of "499"

```
466 0 obj
<</BS 527 0 R/C[0.74902 1 1]/Contents(NOT
```

Figure 9. Starting fragment of an annotation, from FDF, with a dashed border

```
527 0 obj
<</D[3.0 3.0]/S/D>>
endobj
```

Figure 10. Dashed border object, from FDF, associated with annotation from Figure 9

(OPTIONAL) UPDATE ANNOTATION

As an optional step, the user may choose to programmatically modify one or more key information variables in order to comply with the latest guidelines. As an example, Figure 11 shows an aCRF’s annotations programmatically updated from MSG v1.0 to MSG v2.0 and imported into the destination aCRF, including text box color, font color, and border style.

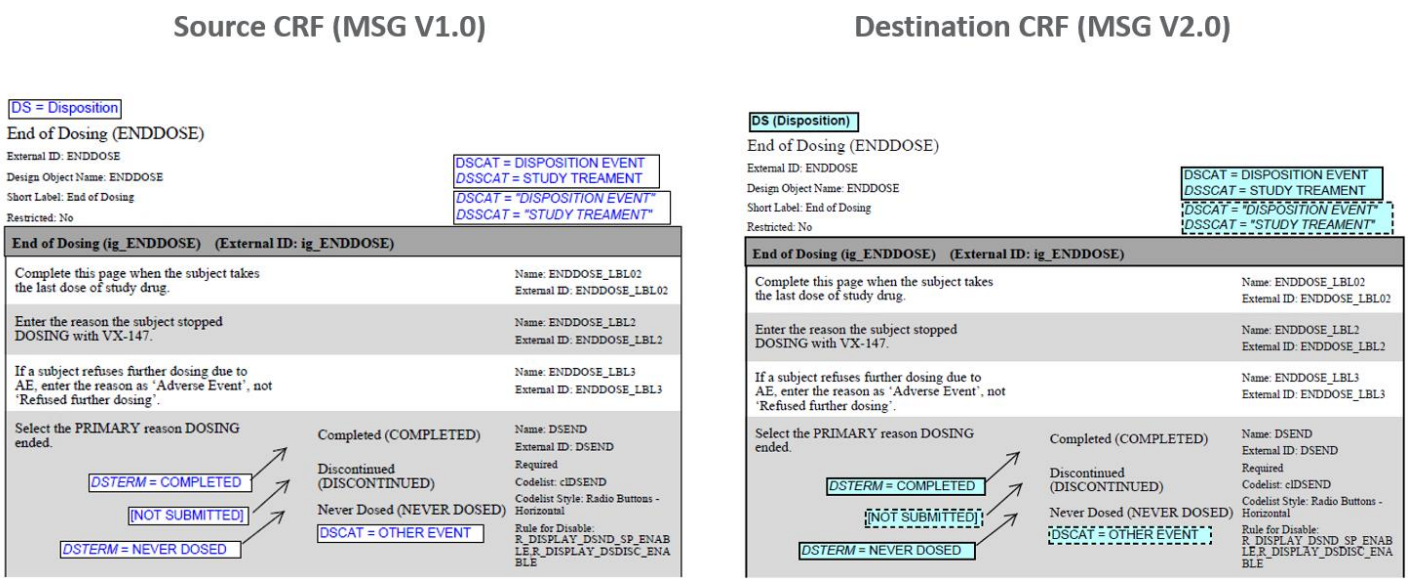

Figure 11. Sample annotations showing programmatic updating

ALIGN ANNOTATIONS AND FORM NAMES/PAGES

Once all annotations have been parsed, the next challenge is to programmatically align each annotation with its target page number on the Destination CRF. To do so, each annotation must be aligned with the CRF Form Name (or ID) of its Source CRF, in order to then match with the Form Name (ID) of the Destination CRF.

EXTRACT SOURCE & DESTINATION CRF INFORMATION

In order to capture the Source CRF's key information, it must be processed into a SAS-parsable format; it is recommended that the user use the XPDF's pdftohtml utility to convert the Source CRF to HTML and read this file into SAS. Once the user has read the Source CRF into SAS, it is recommended that the following information be extracted (Figure 12): the long form name of each CRF Form page (FORMNAME), the short form name (FORMEID), the starting page number of each Form (PAGE), and the number of pages representing each Form (FORM_PGCT). This same information should be extracted for the Destination CRF in the same manner (Figure 13).

PAGE	FORMNAME	FORMEID	FORM_PGCT
123	End of Follow-Up	ENDFUP	3
126	Inclusion and Exclusion	IE	3
129	Informed Consent	IC	1
130	Medical History	MH	2
132	Non-Pharmacological Treatments or Procedures	NT	1
133	Physical Exam	PE	5
138	Prior and Concomitant Medications	CM	2
140	Pulse Oximetry	PULSEOX	1
141	Pulse Oximetry - Unscheduled	PULSEOX_UNS	1
142	Randomization	RAND	2
144	Reconsent	RIC	1

Figure 12. Key information from a Source CRF, captured in SAS

121	End of Follow-Up	ENDFUP	3
124	Inclusion and Exclusion	IE	3
127	Informed Consent	IC	1
128	Medical History	MH	2
130	Non-Pharmacological Treatments or Procedures	NT	1
131	Phone Call	PH	1
132	Physical Exam	PE	5
137	Prior and Concomitant Medications	CM	2
139	Randomization	RAND	1
140	Reconsent	RIC	1

Figure 13. Key information from a Destination CRF, captured in SAS

ALIGN SOURCE CRF WITH ANNOTATIONS

Once the user has extracted the Source CRF information, the individual annotations can be matched with their corresponding Source CRF Form names (FORMNAME/FORMEID) by matching the annotation page number with the page ranges of each Source CRF Form (PAGE/FORM_PAGE_CT); it is recommended that the annotation SAS dataset differentiates between an annotation that occurs on the starting page of its form and one that occurs on the subsequent pages, to ensure that annotations are mapped correctly to the Destination aCRF.

ALIGN DESTINATION CRF WITH ANNOTATIONS

Once the user has aligned each annotation with its corresponding form name (Source CRF's FORMNAME or FORMEID), this information can then be used to align the annotation with the corresponding form in the Destination CRF. Once these have been matched, the user can programmatically update the target page for each annotation. For example, the Source CRF's ENDFUP form (Figure 12) began on the PDF's page 123, while the Destination CRF's ENDFUP form (Figure 13) started on page 121, as all result all annotations associated with the ENDFUP form should be re-mapped to occur on page 121 or its subsequent pages.

EXPORT ANNOTATIONS

Once all annotations have been assigned their target page number in the Destination CRF, the user must recreate the FDF file containing selected annotations, export the file, and import it into the Destination CRF.

FORMATTING ANNOTATIONS FOR EXPORT

To reconstruct the FDF file's "body", the user should take the following steps to format and prepare the annotation dataset for export:

1. Retain the annotations matched with the destination aCRF and update the internal "page placeholder" with the new target page (see Figure 14).

ANNOT_STORE	CONTENT	CRF_PAGE
[XX] 0 obj<</C[BACKCOLOR]/Contents(DS \(Disposition\))/CreationDate(D:00000000000000Z)/DA(0 0 0 rg /[FONT]-[FONTWEIGHT]MT [FONTSIZEINT] Tf)/DS(font: [FONTWEIGHT] [FONT] [FONTSIZE]; text-align:left; color:[FONTCOLOR])/F 4/M(D:20231226230640-05'00')/NM(a94f42ee-ec09-4d6c-9943-80f13f2dfb08)/Page [Z]/RC(<?xml version="1.0"?><body xmlns="http://www.w3.org/1999/xhtml" xmlns:xfa="http://www.xfa.org/schema/xfa-data/1.0/" xfa:APIVersion="Acrobat:23.6.0" xfa:spec="2.0.2" style="font-size:[FONTSIZE];text-align:left;color:[FONTCOLOR];font-weight:[FONTWEIGHT];font-style:[FONTSTYLE];\font-family:[FONT];font-stretc dir="ltr">DS \(Disposition\)</p></body>)/Rect[85.9666 698.568 169.02 712.578]/Subi(VS)/Subtype/FreeText/T(ENDFUP)/Type/Annot>>endobi	DS \(Disposition\)	ENDFUP

Figure 14. An annotation in SAS, showing the page placeholder, the associated CRF Form Name, and the target page

2. Restore all key style element placeholders (e.g. font size, background color, font weight), using the corresponding variables' values in the annotation dataset.
3. Restore the placeholders for each annotation's identification number (and Parent number, where applicable). The first annotation should start with "2" ("1" is reserved for the "header") and iterate consecutively for all subsequent annotations.
4. Create an "associated annotation" for all annotations formatted with dashed borders (see Figure 9 and Figure 10). These "associated annotation" should be located after all textbox/line annotations (e.g. if there are 250 annotations, the last "text" annotation will be numbered as "251" and the first dashed border "associate" will be numbered as "252").
5. Modify the "header", as a dataset extracted from the original FDF file, to represent the correct number of annotations in the dataset (numbered as XX 0 R; see Figure 2).
6. Concatenate the "header", "body" and "footer" datasets (see Figure 15).

Figure 15. Example records in annotation dataset, ready for export

EXPORTING THE ANNOTATIONS AS AN FDF

The user can output the final annotation dataset as an FDF file, using the FILE statement.

IMPORTING THE FDF FILE INTO THE DESTINATION CRF

The user can use the Comments tool to import the FDF file into their Destination CRF and should confirm all annotations have been properly migrated. It should be noted that, when using Adobe as the PDF reader, importing an FDF file into the Destination CRF may produce a pop-up warning (Figure 16); this notification is due to a different file name in the FDF "header" than the Destination CRF. As modifying this value in the FDF file can cause corruption and import failures, it is recommended to click "Yes" when the pop-up message appears.

Figure 16. Expected and harmless pop-up from importing FDF file

CONCLUSION

Although the proposed macro for exporting, modifying, and **IMPORTING** CRF annotations is far from a "one-click" solution, its off-loading of copying, editing, and moving annotations from programmer to program has reduced the number of both hours and errors involved in this critical but time-consuming process.

REFERENCES

Hufford, Walter. 2014. "Automating Production of blankcrf.pdf." PharmaSUG 2014 Proceedings, Available at https://pharmasug.org/proceedings/2014/CC/PharmaSUG-2014-CC21.pdf.

ACKNOWLEDGMENTS

Thank you to Susheel Arkala, who served as beta tester and troubleshooter in the development of this macro and who pointed me in the direction of Hufford's paper, and to Todd Case, Emily Seem, and Dr. Valerie Finnemeyer for your valuable feedback on this paper.

CONTACT INFORMATION

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

Matthew Finnemeyer
Vertex Pharmaceuticals, Inc.
Matthew_finnemeyer@vrtx.com