PharmaSUG 2023 - Paper XX-###

Automatic CRF Annotations Using Python Together JSON Format File

Pengpeng Li, BioEnvo Corporation;

# readme

Light tan highlighting (the background in this paragraph) is used for all comments that tell you how to use this template and specify guidelines that you should follow. For more detailed instructions about how to use the template and writing guidelines that you should follow, see the section “Basic Instructions .“ (Every cross-reference include a link.)

When you use this template, many formatting issues are automatically controlled for you. For example, you do not need to add blank lines between paragraphs because all defined paragraph styles (for example, headings and PaperBody) automatically include the proper space between paragraphs.

Replace with your own text all text that is not highlighted in light tan except for the last two paragraphs (about trademarks) and any headings that have “Do not change” after them.

When you have finished writing your paper, delete all text that is highlighted in light tan, including the README heading.

# Abstract

Do not change the heading style or the text “ABSTRACT” of the preceding heading. Replace the following text with the edited abstract from the conference system.

# Of utmost importance to both sponsor and CRO companies is ensuring high quality and efficiency when generating SDTM aCRFs. This is because SDTM aCRF is a crucial element in the SDTM submission package, and many pharmaceutical MNCs have established their own company-level SDTM aCRF standards in addition to CDISC guidelines. To guarantee success in the highly competitive pharmaceutical industry, it is essential to carefully follow both CDISC and company-level SDTM standards when generating SDTM aCRFs. In this paper, we present a novel way which relies on a meticulously crafted infinite-dimensional and regular expression compatible hashing data structure to automated generated annotated aCRF with quality and efficiency. This method can automatic positioning and locating of comment location and additionally possesses loose coupling capability, making it convenient for multiple collaborators to modularly handle lengthy documents.

# Introduction <heading 1>

Do not change the heading style or the text “INTRODUCTION” of the preceding heading. Replace the following text with one or more paragraphs of an introduction that explain the purpose and scope of your paper and provide readers with any general information that they need to understand your paper.

The genesis of SDTM production lies in the annotation of vacant CRF pages, which must be packaged alongside SDTM datasets as an integral element of the clinical data submission package to the Food and Drug Administration(FDA). For statisticians and programmers, the laborious task of manually annotating CRFs using the Adobe Acrobat comment tool presents a hurdle. They have struggled to mechanize this process, yet, most of the current techniques or packages rely on the use of multiple software. These methods generate Forms Data Format (FDF) files for import substitution into blank CRFs. These approaches record poor annotation performance when applied to unprecedented CRFs or versions of CRFs that undergo radical transformations. Other method such as using R tm package or python's PyMUF2 packages to capture text blocks in PDF file and simply arrange a number to each question solely by vertical coordinate of neighbored line break and organizing data which finally store in EXCEL. For most case ,it will contain mistakes for PDF parsing result whose structure in not consistent with Study Design Specification (SDS) file which is not always available for annotating CRF and bring unexpected annotation mistakes . To surmount this pitfall, we have developed a Python module that capitalizes on the influential characteristics of Python to annotate CRF pages automatically. Our package supersedes customary methods and other computerized tools by providing significant attributes, including:

* Abandoning coordinate systems, EXCEL, or FDF files, we propose utilizing a meticulously designed data structure based on the "infinite dimensional hash" to record the entire structure of the CRF form. This data structure will generate a JSON file containing all possible annotation positions. Users can directly modify the JSON file to enable automatic annotation of the desired text. Annotation width and position are automatically calculated by the system and integrated into the generated JSON file without human intervention .
* There is no need to ensure the correctness of CRF structure parsing. As long as the input parameters are accurate, all possible annotation positions will be traversed, even if the document structure parsing results do not match expectations. The annotation effect will not be affected by any discrepancies.
* Users do not need to pay attention to the PDF file itself, only edit the parsed JSON file to achieve highly reliable document annotation results and greatly improve efficiency. Furthermore, related editors can format and display the JSON file, facilitating a better understanding of the document by workers and enabling them to work without reference to the source document.
* This task is highly modular and portable. A large CRF PDF file can be divided into multiple sub-documents and completed collaboratively by multiple personnel. The data structure then combines the annotation results without losing any data integrity. This means that the once cumbersome and labor-intensive work can now be completed by teams working by divide-and-conquer strategy. Additionally, if necessary, the project can even be open-sourced on Github.
* Annotation migration becomes possible. Based on Python's build-in functions, annotation updates and comparisons for multiple different versions of the same document can be performed through key-value pair conversion, enabling seamless annotation migration to the required result without data loss.
* The system supports regular expression matching, and the annotated text can be synchronized based on specific variables. Standard regular expressions can be used to enable the annotated text to change automatically with the matching variables.This greatly simplifies the workload for some tasks.
* The annotation records are stored in the JSON format file and can be easily dumped into databases such as MongoDB, Redis, and jsonDB, enabling remote management and remote work needs at any time. Furthermore, database management can also bring benefits to task management.
* The system has a complete command-line tool and graphical user interface. Apart from editing the JSON file, all other steps are basically automated.

As you write, consider the following:

* Keep your audience members (their skill levels, industries, and job roles) in mind.
* Identify what skills or knowledge your audience will gain.

# WROKFLOW AND PROCESS

## Data Structure for storing data

First of all, we design a specific data structure to store the parsing result of blank CRF PDF file. Our ultimate goal is to create a self-expandable infinite-dimensional hash table that can perform regular expression fuzzy matching we package this data structure as a class named **MultiRegexDict** After testing, we could clearly find we fulfill the envisioning that the MultiRegexDict class supports automatic self-expandable and regular expression fuzzy matching.



Figure 1. Testing code for data structure to store PDF parsing result

## workflow

Based on the **MultiRegexDict** class, we design an automatic workflow for CRF annotating. The whole process consists of 3 steps , and a series of python scripts have written to fulfill the function.

1. Parsing PDF file and map corresponding Heading to **MultiRegexDict** class and dump it into a JSON file.
2. Edit the JSON file and add annotation .
3. Map back the Annotated JSON to Blank CRF PDF file to generate Annotated PDF File.

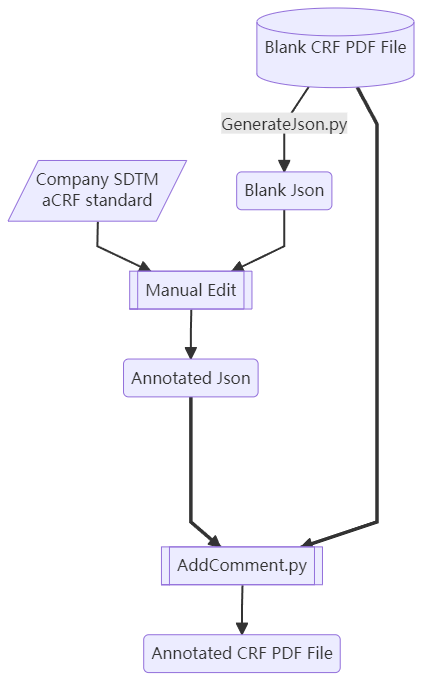


Figure 2. Brief workflow overview

Always introduce a table by inserting a cross-reference. For instructions, see the section “To insert a cross-reference.”

| SAS Variable Format | DB2 Data Type |
| --- | --- |
| $*w*.  $CHAR*w*. | CHARACTER |
| any date format | DATE |

Table 1. DBLOAD Procedure: Default DB2 Data Types for SAS Variable Formats

Always use the Caption paragraph style for table captions. See the section “To insert a caption.” Following those instructions ensures that the tables are automatically numbered (even if you rearrange them later). Use title style capitalization for captions (as shown in this example).

# Second Main Topic <heading 1>

This is a main topic in the paper. This paragraph uses the PaperBody style. This paragraph uses the PaperBody style.

If you need to include a numbered (ordered) list, copy the following list, paste it, and modify the text. Be sure to introduce your list with a complete sentence that ends with a colon. For example, “Here are the required steps:”.

1. This is a sample numbered or ordered list item. This is list item text. This is list item text. This is list item text.
2. This is a sample numbered or ordered list item. This is list item text.

This is another sample paragraph. This paragraph uses the PaperBody style. This paragraph uses the PaperBody style.

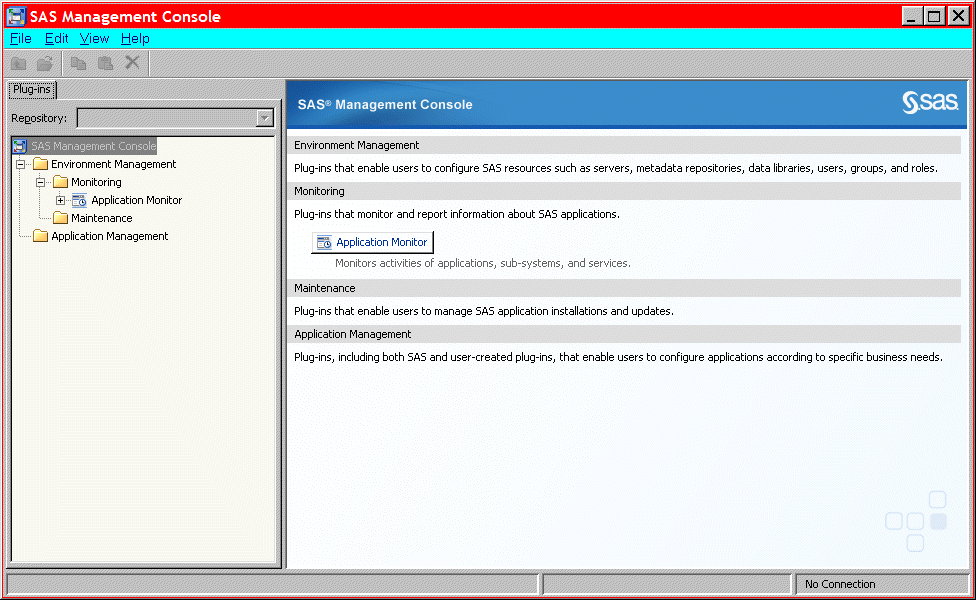
If you need to include a bulleted (unordered) list, copy the following list, paste it, and modify the text. Be sure to introduce your list with a complete sentence that ends with a colon. For example, “Systems that are supported by the product include the following:”.

* This is a sample bulleted list item. This is list item text. This is list item text. This is list item text. This is list item text.
* This is a sample bulleted list item. This is list item text.

This is a continuation of the body of the paper—after an unordered list. This paragraph uses the PaperBody style.

Display 1 is sample display or screen capture.

Always introduce a display by inserting a cross-reference. For instructions, see the section “To insert a cross-reference.”



Display 1. Former Main Interface for SAS Management Console

Always use the Caption paragraph style for display captions. See the section “To insert a cross-reference.” Following those instructions ensures that the displays are automatically numbered (even if you rearrange them later). Use title style capitalization for captions (as shown in this example).

## Subhead A Level <heading 2>

This paragraph uses the PaperBody style.

Output 1 shows an example of how to present output.

Always introduce output by inserting a cross-reference. For instructions, see the section “To insert a cross-reference.”

CREATE TABLE ALLACCTX(SourceSystem varchar(4),

cctnum numeric(18,5) CONSTRAINT "ALLACCT\_PK" PRIMARY KEY,

ccttype numeric(18,5),balance numeric(18,5),clientid numeric(18,5),

losedate date,opendate date,primary\_cd numeric(18,5),status varchar(1))

Output 1. Output from a CREATE TABLE Statement

Always use the Caption paragraph style for output captions. See the section “To insert a caption.” Following those instructions ensures that the outputs are automatically numbered (even if you rearrange them later). Use title style capitalization for captions (as shown in this example). Note that output has the same font as source code, but it is in a box. (This box is not a text box.) The easiest way to properly format output is to copy the preceding output (including the box) and its caption, paste it, and then modify the text.

## Subhead A Level <heading 2>

This paragraph uses the PaperBody style.

### Subhead B Level <Heading 3>

This is a subtopic of a subtopic. This paragraph uses the PaperBody style.

#### Subhead C Level <Heading 4>

This paragraph uses the PaperBody style.

# Conclusion <heading 1>

Do not change the heading style or the text “CONCLUSION” of the preceding heading

This paragraph uses the PaperBody style.

The conclusion summarizes your paper and ties together any loose ends. You can use the conclusion to make any final points such as recommendations, predictions, or judgments.

# References <heading 1>

This section is not required. If you include this section, do not change the heading style or the text “REFERENCES” of the preceding heading.

The following references are examples and do not cover the spectrum of references that might be included. The important thing is to be consistent in formatting and organizing your references. If you prefer, you can follow a bibliographic approach such as the American Psychological Association (<http://www.apastyle.org/manual/index.aspx>) or the American Statistical Association (<http://amstat.tfjournals.com/asa-style-guide/>).

This is the format for references if you use the author-date format, in which citations are inserted in parentheses in text, such as (Smith 2014) or (Smith, 2014; Doe and Johnson, 2016). If the same author has several references, use a, b, c, and so on. For example, (SAS Institute Inc. 2015a, 2016b, and 2016c).

Book <Author name: last name, first name (or initials)>. <Publication date>. <*Book title*>. <City, State (abbrev) of publication> : <Publisher name>.

Journal article <Author name: last name, first name>. <Publication date>. “<Article title>.” <*Journal title*>, <volume no.:page numbers>.

Article in conference proceedings <Author name: last name, first name>. <Publication date>. “<Article title>.” <*Title of proceedings such as Proceedings of the SAS Global 2010 Conference*>, <City, State (abbrev) of publication> : <Publisher name>. Optional: You can add a URL to access available online proceedings. For example: Available at <http://support.sas.com/resources/papers/proceedings09/TOC.html>.

Website <Author name: last name, first name>. “<Title>.” <*Source*>. <Date>. Available at <URL>.

Reference examples:

Book Agresti, A. 2013. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Journal article Akaike, H. 1979. “A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting.” *Biometrika,* 66:237–242.

Article in conference proceedings Dorfman, A. H. and R. Valliant. 1993. “Quantile Variance Estimators in Complex Surveys.” *Proceedings of the Survey Research Methods Section*, 866–871. Alexandria, VA: American Statistical Association.

Website Federal Reserve Bank of St. Louis. 2012. “Economic Research.” Accessed November 7, 2012. <http://research.stlouisfed.org>.

# Acknowledgments <heading 1>

This section is not required. If you include this section, do not change the heading style or the text “ACKNOWLEDGMENTS” of the preceding heading.

This is the text for the acknowledgments. This paragraph uses the PaperBody style.

# Recommended Reading <heading 1>

This section is not required. If you include this section, do not change the heading style or the text “RECOMMENDED READING” of the preceding heading.

This is the format for recommended reading.

* *Base SAS® Procedures Guide*
* *SAS® For Dummies®*

# Contact Information <heading 1>

Do not change the heading style or the text “CONTACT INFORMATION” of the preceding heading. Do not change the text of the following paragraph. Replace all fields that are shown in angle brackets.

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

<Name>

<Enterprise (optional)>

<Phone (optional)>

<E-mail>

<Web (optional)

The next sentence is **optional** and applicable if you are including references to trademarked programs in your paper.

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

Remember to delete the following instructions and all other text that is highlighted in light tan.

# Basic Instructions

## Writing Guidelines

### Writing style

* Use active voice. (Use passive voice only if the recipient of the action needs to be emphasized.) For example:

The product creates reports. (active)  
Reports are created by the product. (passive)

* Use second person and present tense as much as possible. For example:

You get accurate results from this product. (second person, present tense)  
The user will get accurate results from this product. (future tense)

* Run spellcheck, and fix errors in grammar and punctuation.

### Citing references

All published work that is cited in your paper must be listed in the REFERENCES section.

If you include text or visuals that were written or developed by someone other than yourself, you must use the following guidelines to cite the sources:

* If you use material that is copyrighted, you must mention that you have permission from the copyright holder or the publisher, who might also require you to include a copyright notice. For example: “Reprinted with permission of SAS Institute Inc. from *SAS® Risk Dimensions®: Examples and Exercises*. Copyright 2004. SAS Institute Inc.”
* If you use information from a previously printed source from which you haven’t requested copyright permission, you must cite the source in parentheses after the paraphrased text. For example: “The minimum variance defines the distance between cluster (Ward 1984, p. 23)

## Tips for using Word

These instructions are written for MS Word 2007 and MS Word 2010. The steps are similar for MS Word 2003.

### To select a paragraph style

1. Click the HOME tab. The most common styles in your document are displayed in the top right area of the Microsoft ribbon. If you don’t see a style that you want, click the slanted down arrow at the bottom right corner of the Styles area, and scroll through the list. The main styles for this template are headings 1 through 4, PaperBody, and Caption. Avoid using other styles.
2. To change a paragraph style, click the paragraph to which you want to apply a style, and then click the style that you want in the ribbon.
3. PaperBody (used for most text) is automatically applied when you press Enter at the end of any heading style or the Caption style.

### To insert a caption

1. Click **REFERENCES** on the main Word menu.
2. Click **Insert Caption**.
3. Select the **Label** type that you want.
4. Click **OK**.

### To insert a cross-reference

1. Click **REFERENCES** on the main Word menu.
2. Click **Cross-reference**.
3. In the **Reference type** list box, select Heading, Figure, Table, Display, or Output.
4. For a heading:
   1. In the **For which heading** list, select the heading that you want.
   2. From the **Insert reference to** list, select **Heading text.**
5. For a figure, table, display, or output:
   1. In the **For which caption** list, select the caption that you want.
   2. From the **Insert reference to** list, select **Only label and number**.

### To insert a graphic from a file

1. Click **INSERT** on the main Word menu.
2. Click **Picture**.
3. In the Insert Picture dialog box, navigate to the file that you want to insert.
4. When the name of the file that you want to insert is displayed in the **File name** box, click **Insert**.