PharmaSUG China 2023 - Paper AP-178

CRF Translation and aCRF Generation

Fei Chen, Jie Wei, Chengwei Li, Kewei Yuan, Eli Lilly and Company

ABSTRACT

As the Clinical Trial Data Submission Guideline was issued by CDE in 2020, it requires foreign CRF for clinical trial submission to NMPA translated into Chinese. Additionally, generation of SDTM annotation has become common practice for agency submissions. To meet these new regulatory requirements and streamline the process of CRF translation and annotation creation, this paper will discuss the method to achieve the above the goal based on CRF designed in Veeva Vault EDC system.

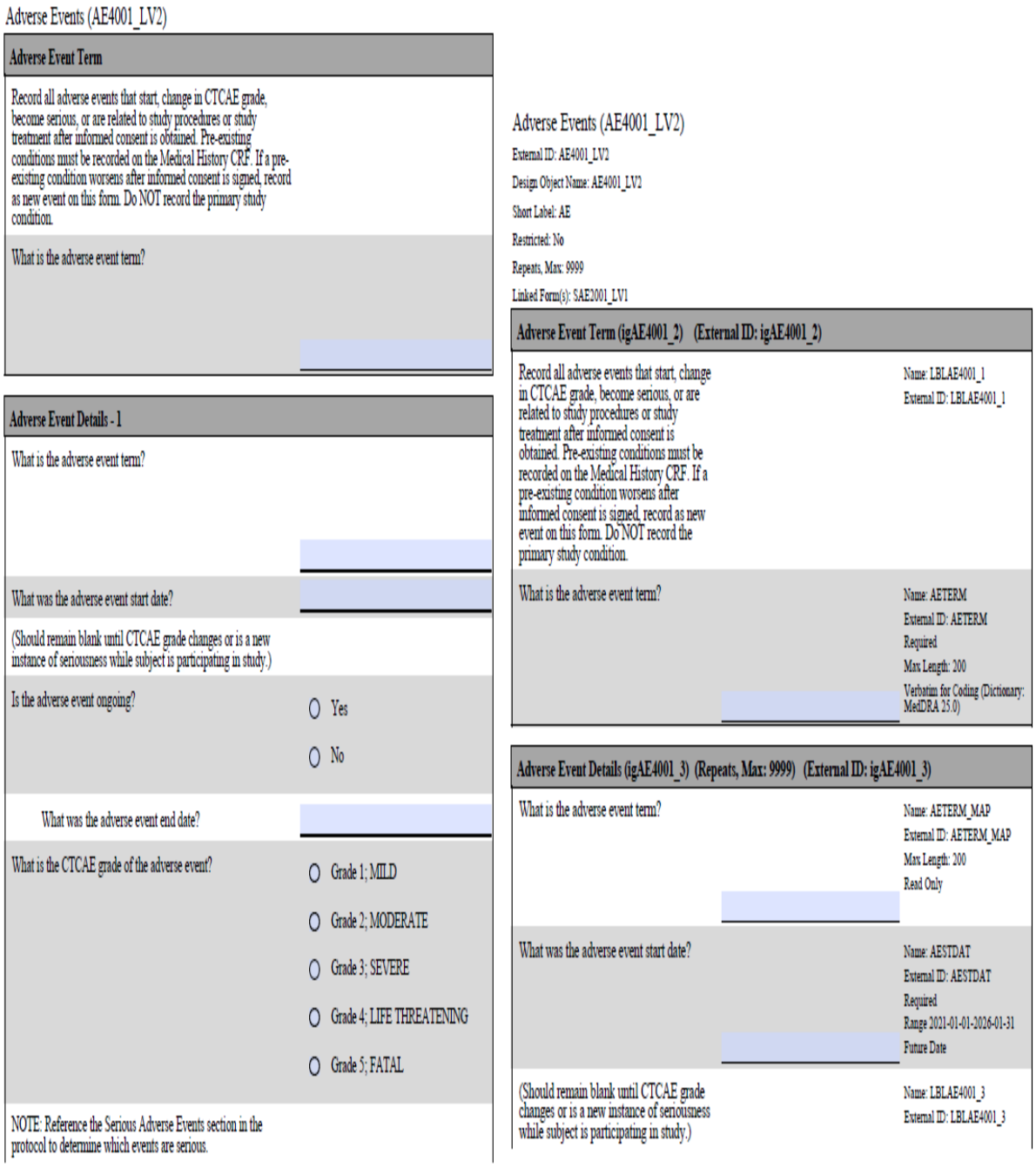
As of CRF Chinese translation, a paper CRF Chinese Translation – Mask Plan in PharmaSUG China 2022 (Chunpeng Zhao etc.) was released to introduce the detailed technical skills based on CRF in RAVE EDC system. As our approach largely aligns with that one, the paper will not detail those specific skills, Instead, we will focus on Vault CRF sentence assembly, background color capture and mapping achieved through Python programming to extend the mask plan to other CRF system. Moreover, this paper illustrates how to optimally utilize automation for generating aCRF through Vault EDC document study design specification (SDS), The essential methodology involves the following steps:

1. Manually populate SDTM annotations in the initial study in SDS file, then SDTM annotations will be iterated for upcoming studies based on previous saved SDS.
2. Precisely calculate each blank CRF items' position.
3. Merge the source data annotated CRF (SRCCRF, same as below) with SDS file to capture the SDTM annotations by external ID.
4. Merge blank CRF and SRCCRF by form name, section head and item to capture the SDTM annotations beneath each item in blank CRF.

The CRF translation and SDTM annotations files are both generated in .xfdf format, which are then imported into the blank CRF, thereby completing the main body of the translation and annotation process.

INTRODUCTION

There have been tremendous papers exploring this topic to figure out how to efficiently generate aCRF, but few examples discuss the translation of English-language CRF in Chinese. One such example is the paper titled "CRF Chinese Translation – Mask Plan," which was shared at China PharmaSUG2022. This paper inspired us to adopt this methodology in the Veeva Vault system.
It is important to understand the structure of the blank CRF (BLANKCRF) and the source annotated CRF (SRCCRF) in the Vault system, as they have distinct features that are utilized in, for example, CRF items coordinator and background color mapping. The Vault CRF, as shown in Display 1, consists of structured forms with two or three columns and different sections with mixed background colors. This design enables reviewers to conveniently capture item information and its attributes, it would also contribute to create SDTM variables annotations mapping in a specified alignment. However, it also brings challenges presents challenges in masking the background color. Does background color information could be extracted? As we are known aCRF comprises with two parts: foreign language CRF translation and SDTM variables annotations, currently, there are very few papers being integrative to cover both topics in Vault CRF. In this discussion, we will explore how to go beyond the mask plan and leverage structured forms to automate SDTM annotations more effectively.

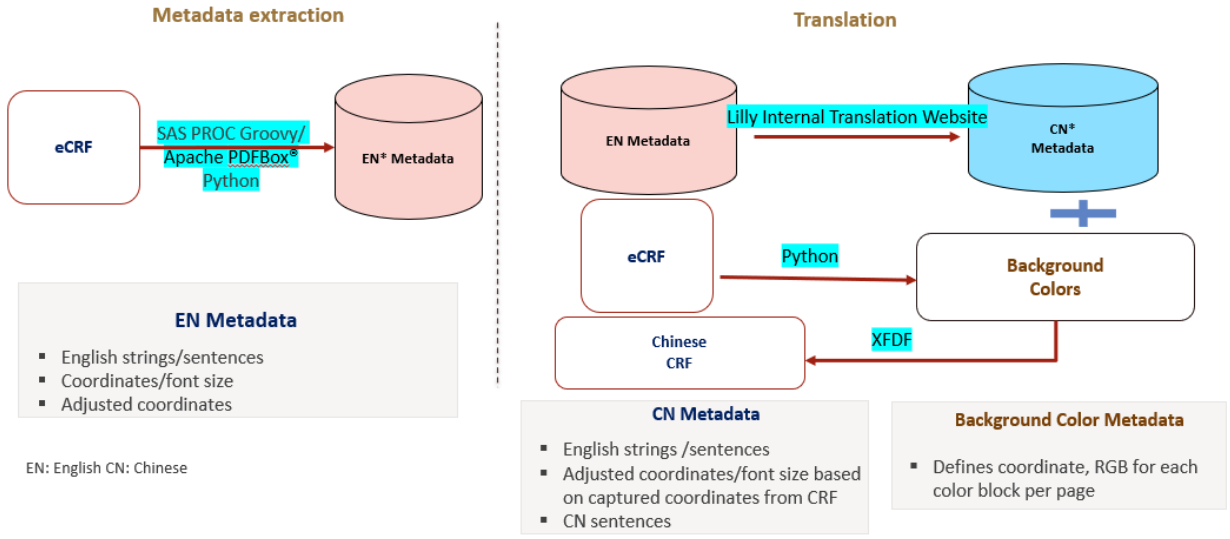


Display: 1 Veeva Vault CRF Layout (The left one is BLANKCRF, the right one is SRCCRF)

OVERVIEW OF CRF TRANSLATION INTO CHINESE

The implementation of CRF translation and SDTM annotation is inevitable following the guidelines set forth by the NMPA for the submission of clinical trial data, regardless of China alone study or global study with China subgroup. Various methods exist for translating CRF, one of which is the "mask plan" approach, where the Chinese language is chosen to seamlessly replace the original language. This paper generally adopts the mask plan method. However, instead of delving into the technical aspects that have

already been described in existing papers, our focus is on the overall process and specific considerations related to the Veeva Vault CRF. Display: 2 provides a refined workflow process that illustrates these aspects.



Display: 2 Workflow on CRF Translation

Left part split by vertical dotted line is English metadata extraction from eCRF (SRCCRF as example) by utilization of SAS GROOVY to call tools of Apache PDFBox to generate txt file which includes all the characters (alphabet, punctuation mark, number) (See Display: 3), by utilization of Python to assemble each character into word, phrase (See Display: 4), then into sentence/paragraph (See Display: 5) as well as its position, the final result will be saved into translated metadata excel file for Chinese translation.

The right part is to a) translate each entry item in CRF into Chinese, b) extract background color and its coordinators of SRCCRF and BLANKCRF (See Display: 6, the background color will have introduction in next section) by Python, c) SAS program is eventually created to integrate both metadata information and exported into XFDF file.

```
String[128.964,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=2.25]
String[131.214,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=6.498001]A
String[137.712,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.5]n
String[142.212,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.5]n
String[146.712,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.5]o
String[151.212,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=2.4929962]t
String[153.705,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=3.9869995]a
String[157.692,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=2.4929962]t
String[160.185,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=2.4929962]i
String[162.678,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.5]o
String[167.178,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.5]n
String[171.678,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=3.501007]s
String[309.0,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=6.497986]N
String[315.498,355.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.5]o
String[102.0,368.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=5.9940033]C
String[107.994,368.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=3.9869995]a
String[111.981,368.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=3.5009995]s
String[115.482,368.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=3.9869995]e
String[119.469,368.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.5]b
String[123.969,368.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.4999924]o
String[128.469,368.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.5]o
String[132.969,368.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=4.5]k
String[137.469,368.0 fs=9.0 xscale=9.0 height=4.086 space=2.25 width=2.25]
```

Display: 3 Text Extraction from Source Annotated CRF

page	phrase	x axis	y axis	length	height
2	Record all adverse events that start, change	90.7	218	172.8	4.54
2	External ID: LBLAE4001_1	446.37	218	91.464	3.632
2	in CTCAE grade, become serious, or are	90.7	228	162.59	4.54
2	related to study procedures or study	90.7	238	142.92	4.54
2	treatment after informed consent is	90.7	248	139.8	4.54
2	obtained. Pre-existing conditions must be	90.7	258	165.67	4.54
2	recorded on the Medical History CRF. If a	90.7	268	169.82	4.54
2	pre-existing condition worsens after	90.7	278	144.01	4.54
2	informed consent is signed, record as new	90.7	288	167.89	4.54
2	event on this form. Do NOT record the	90.7	298	155.97	4.54
2	primary study condition.	90.7	308	98.53	4.54
2	What is the adverse event term?	90.7	328	127.86	4.54
2	External ID: AETERM	454.81	328	74.576	3.632

Display: 4 Phrases Assemble from Character Text/Word

page	sentence	x_s	y_s	x_e	y_e	label	xs	ys	xe	ye
2	External ID: LBLAE4001_1	446.37	214.37	537.83	222.132	External ID: LBLAE4001_1	446.37	214.37	550	313
2	Record all adverse events that start, change in CTCAE grade, become serious, or are related to study procedures or study treatment after informed consent is obtained. Pre-existing conditions must be recorded on the Medical History CRF. If a pre-existing condition worsens after informed consent is signed, record as new event on this form. Do NOT record the primary study condition.	90.7	213.46	263.5	313.04	记录所有在获得知情同意后开始的、CTCAE分级改变的、变得严重的、或与研究程序或研究治疗有关的不良事件。必须在病历CRF中记录原有的状况。如果在签署知情同意书后，原有病情恶化，在此表中记录为新事件。不要记录主要的研究状况。	90.7	213.46	272	313
2	What is the adverse event term?	90.7	323.46	218.56	333.04	什么是不良事件术语？	90.7	323.46	272	381

Display: 5 Sentence/paragraph per Entry Item Assemble from Multiple Rows of Phrases

page	left x	left y	right x	right y	RGB	X_2	X_3
6	85	167	550	185	(165, 165, 165)	272	436
6	85	185	550	289	(255, 255, 255)	272	436
7	85	195	550	213	(165, 165, 165)	272	436
7	85	213	550	324	(255, 255, 255)	272	436
7	85	324	550	400	(216, 216, 216)	272	436
7	85	413	550	431	(165, 165, 165)	272	436
7	85	431	550	502	(255, 255, 255)	272	436
7	85	502	550	571	(216, 216, 216)	272	436
7	85	571	550	611	(255, 255, 255)	272	436

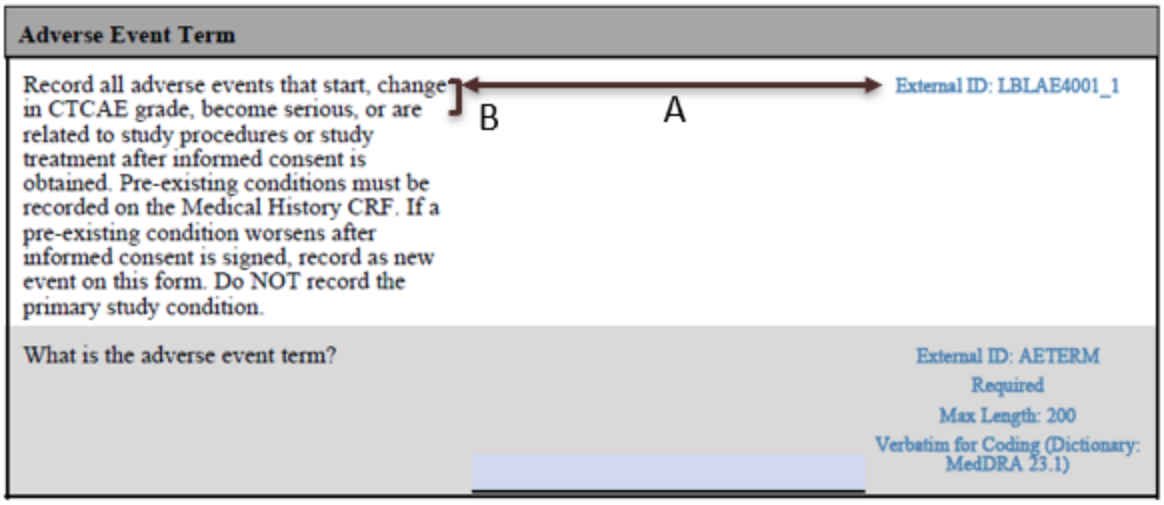
Display: 6 Background Color Information from SRCCRF

The Methodology of Assemble of Word/Phrase/Sentence/Paragraph

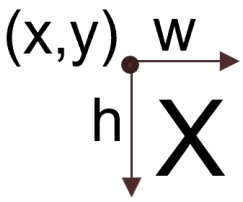
The delimiter used to identify words is a space. If there is not space between last character and next character, then last/current/next character will be concatenated together, this loop will go ahead until space is encountered.

The way to identify a row within a specified column in a phrase/sentence is the distance between current character and next character and difference of height of character. In Display: 7, when the loop of scanning each character go into last character ‘e’ in the ‘change’ and the next scanning is ‘E’ in ‘External’, the distance indicated by A is much longer than a specified value, then we consider the two words would be split into different range. The identifier of different rows within a column is the height of each character. The height of Character ‘e’ in ‘change’ is different from the height of ‘i’ in ‘in’ which is indicated by B, so ‘change’ and ‘in’ would set into different row. By such analogy, all the rows in different columns will be print out as the like of Display: 4.

To recognize the end of a paragraph, the difference in height between rows is considered. In Display: 7, the last character of paragraph is period ‘.’, the different of this period with next character ‘W’ in ‘What’ is significantly larger than a specified value. This indicates the end of a paragraph. Assembling the paragraphs is the final step in assembling the characters. In Display: 5, it includes the original rectangle coordinator indicated by (x_s, x_e, y_s, y_e) as well as the assembled sentence and translated sentence, and the other coordinator (xs, xe, ys, ye) is the adjusted one to update endpoint (xe, ye), it stretches endpoint from original location into right bottom of the cell. This adjustment is to prevent from the exposure of the original English text. It’s noticeable that location of identifying start point of each character is left top (see Display: 8).



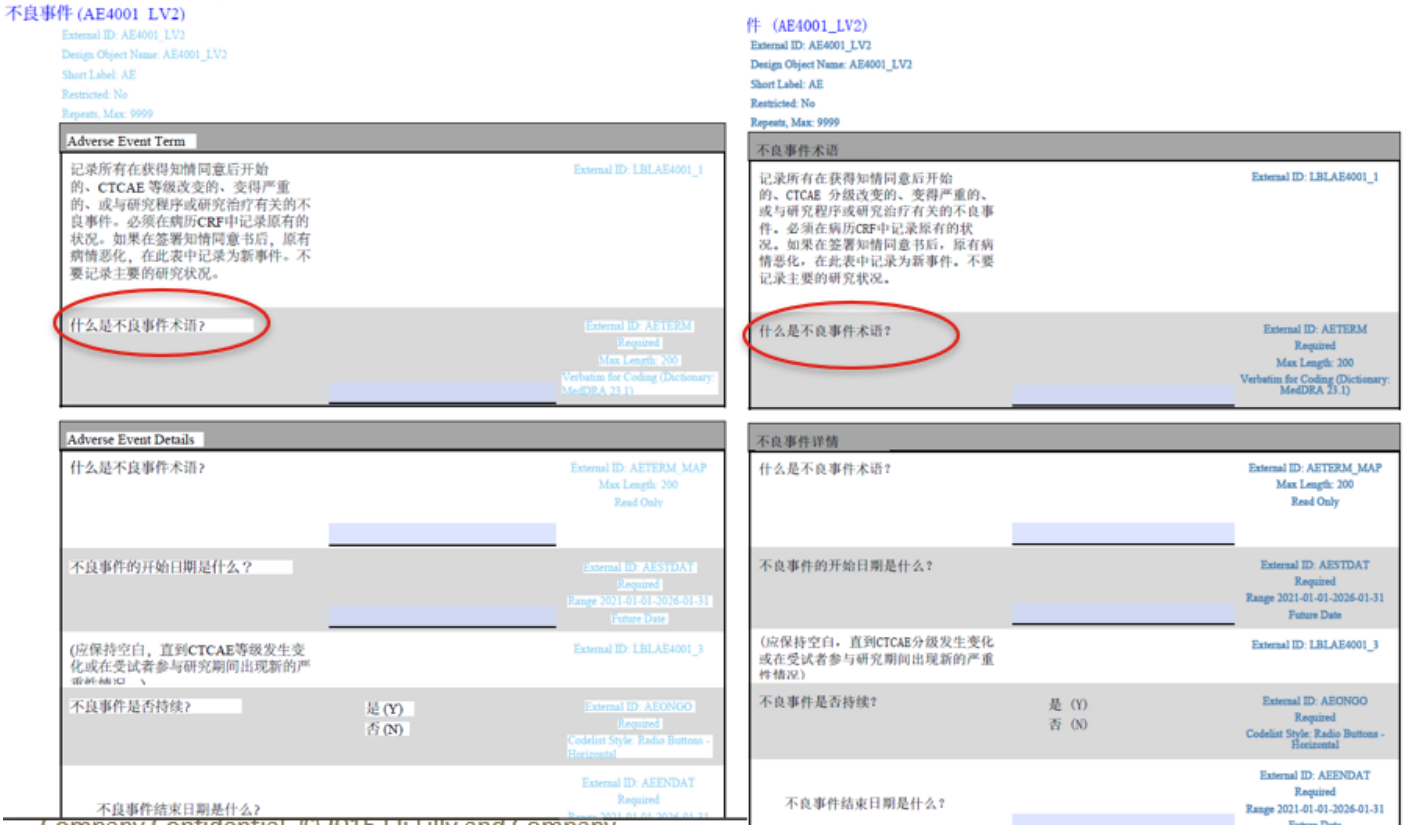
Display: 7 Identify Word/Sentence/Paragraph



Display: 8 Location of Character Coordinator

Background Color Metadata

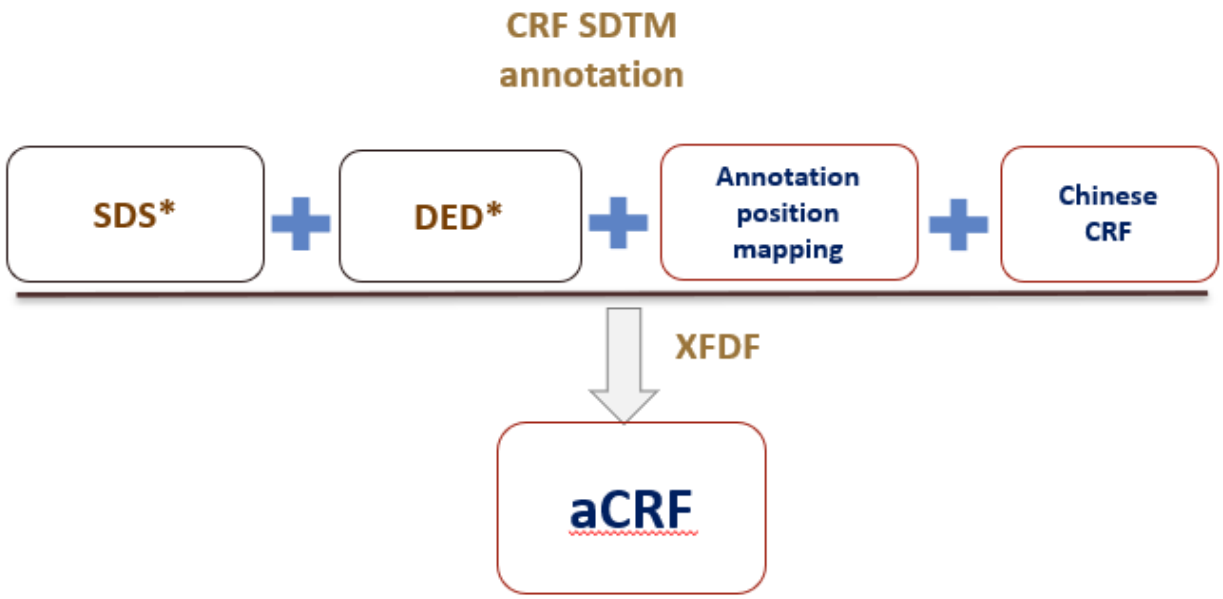
Background Color Metadata is comprised with background color coordinator, RGB and reference line to split columns. See Display: 6, Variable Page means actual CRF page number, left x left y right x right y represent the coordinator of each section from left top to right bottom, and RGB indicates each section background color, x_2 and x_3 is the vertical lines between column 1 and column2, column2 and colum3. Generation of background color file from BLANKCRF is to mask original background to make sure new color is identical to the original one. Generation of background color file from SRCCRF is to achieve the one-to-one link between column1 and column3, it's aimed to get the alignment of location of SDTM annotations with the linkage of SRCCRF and external SDTM metadata file. The detail explanation could be seen in the part of aCRF generation.



Display: 9 Contrast Effect with Background Color Mapping

OVERVIEW OF SDTM VARIALBES ANNOTATIONS

From Display: 10, aCRF need prerequisites: SDS, DED, Annotation position mapping and Chinese CRF as the source CRF. Veeva Vault EDC system could extract SDS as well as its CRF, it defines CRF attributes for each entry item, including form name, label, items group name, label, item name, label, external ID, data type, codelist, etc. As we know item may repeat with one form, it will bring the challenge map SDS file with CRF, so external ID is good choice since it's unique. DED could consider as SDTM variables' annotations population per item, it's noticeable that there are only 2 groups annotations (SDTM Variables Name/SDTM SuppQual Desc, Second SDTM Variables Name/Second SDTM SuppQual Desc). What we want to highlight is Annotation position mapping, this would maximumly reduce manual movement of annotations.



Display: 10 Workflow of SDTM Variables Annotations

ANNOTATION POSITION MAPPING

To reduce the manual movement of annotations, we think putting annotations beneath each item and aligning with start of the item position is good choice (see Display: 11). How could we automatically achieve this, and if annotations are much larger than the allowed space to put the rectangle boxes, how could we make the annotations wrap out into multiple rows like Display: 12? We need to involve SRCCRF as the bridge to link the DED into end file BLANKCRF. The detailed steps followed as below:

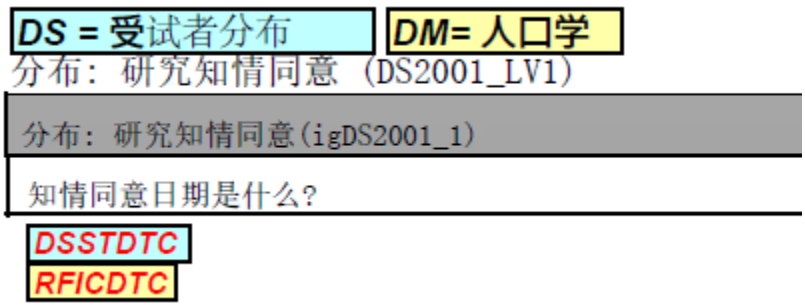
- Link CRF CN metadata in excel file with SRCCRF by each form name and external ID to capture DEDs.
- Link each Entry item and its external ID in CRF CN metadata by background color coordinators.

Merge SRCCRF with BLANKCRF by each form name and sentences, if there are sentences duplicate, then new variable SEQ is derived to count the sentences per form name to make sure one-one match, then precisely calculate every annotation box length and width.

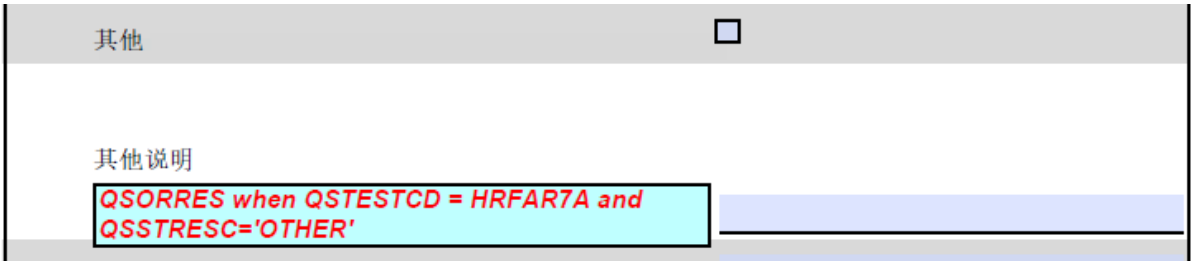
Pertaining to calculation of every annotation box length and width, firstly we need to figure out how could we find every character (alphabet in both lowercase and uppercase, numbers and other non-alphabet symbols) width, so a character list is created see Display: 13 with different font (Arial, Courier New, Times New Roman) / font size (8, 9,10,11,12) or other font family combination as you're interested, then each character information is extracted by same method as above. Those characters are categorized into four groups, A-Z, a-z, average of other symbols and space and set them as macro variables. Secondly, count character and multiply corresponding macro variables' value within whole annotation, then sum them up to get the annotation width and set maximum height as the annotation height (See Display: 14). As we set the start point the beginning beneath of the item, so the primary annotation coordinate of start point is known, if annotation does not wrap into two or more rows, and annotation width and height are calculated

out, so the position of rectangle boxes for one row annotation is greatly established, if there are secondary annotation, use the similar method to position annotations. For annotations requires two or more rows, we set the reference line for splitting column1 and column2 as the horizontal cut point, number of rows could be calculated based on annotation whole width/(the x-axis reference line x_2 – annotation start point), therefore, endpoint coordinate is derived out as well. Finally, after above information is well confirmed, we could create XFDF file and import this file into translated BLANKCRF to become aCRF. It's noticeable that this annotation position does not completely replace the manual movement adjustment, relatively less manual review and update would make the aCRF more perfect.

Once we complete the initiate study translation and SDTM annotations, and to borrow that well-established information into other studies, two SAS dataset transmeta for translation and sdsmeta for annotations are generated as the initialized library dataset to support upcoming studies. To make the library productive to cover more information, both datasets could be iterative and accumulative from other studies.



Display: 11 Example of Annotations Position



Display: 12 Example of Annotations Wrap into Multiple Rows

```
ABCDEFGHIJKLMNOPQRSTUVWXYZ

abcdefghijklmnopqrstuvwxyz

1234567890,./?_\():;'@#$%&*+-=


ABCDEFGHIJKLMNOPQRSTUVWXYZ

abcdefghijklmnopqrstuvwxyz

1234567890,./?_\():;'@#$%&*+-=
```

Display: 13 Example of List of Character

```
%macro annomap(indat=,outdat=,var=,id=);
data &outdat.;
     set &indat.;
     if annotats1^='' or annotats2^='';
     array alpha[26] a b c d e f g h i j k l m n o p q r s t u v w x y
z ;
     array _bigalpha[26] _A _B _C _D _E _F _G _H _I _J _K _L _M _N _O
_P _Q _R _S _T _U _V _W _X _Y _Z;

     array wid_alpha[26] wid_a wid_b wid_c wid_d wid_e wid_f wid_g
wid_h wid_i wid_j wid_k wid_l wid_m wid_n
          wid_o wid_p wid_q wid_r wid_s wid_t wid_u wid_v wid_w wid_x
wid_y wid_z ;
     array wid_bigalpha[26] wid_bigA wid_bigB wid_bigC wid_bigD
wid_bigE wid_bigF wid_bigG wid_bigH
          wid_bigI wid_bigJ wid_bigK wid_bigL wid_bigM wid_bigN
wid_bigO wid_bigP wid_bigQ wid_bigR
          wid_bigS wid_bigT wid_bigU wid_bigV wid_bigW wid_bigX
wid_bigY wid_bigZ;
   if fontsize=10 then do;
    do kk=1 to 26;
       if index(&var.,byte(kk+96)) then
alpha[kk]=countc(&var.,byte(kk+96));
        else alpha[kk]=0;

     end;

     do kkk=1 to 26;
       if index(&var.,byte(kkk+64)) then
_bigalpha[kkk]=countc(&var.,byte(kkk+64));
        else _bigalpha[kkk]=0;

     end;
```

```
      l_l_charen&id.=sum(a*wid_a, b*wid_b, c*wid_c, d*wid_d, e*wid_e,
f*wid_f, g*wid_g, h*wid_h, i*wid_i,
     j*wid_j, k*wid_k, l*wid_l, m*wid_m, n*wid_n, o*wid_o, p*wid_p,
q*wid_q, r*wid_r, s*wid_s, t*wid_t, u*wid_u,
     v*wid_v, w*wid_x, x*wid_x,y*wid_y, z*wid_z);

      u_l_charen&id.=sum(_a*wid_bigA, _b*wid_bigB, _c*wid_bigC,
_d*wid_bigD, _e*wid_bigE, _f*wid_bigF, _g*wid_bigG, _h*wid_bigH,
     _i*wid_bigI, _j*wid_bigJ, _k*wid_bigK, _l*wid_bigL, _m*wid_bigM,
_n*wid_bigN, _o*wid_bigO, _p*wid_bigP, _q*wid_bigQ,
     _r*wid_bigR, _s*wid_bigS, _t*wid_bigT, _u*wid_bigU, _v*wid_bigV,
_w*wid_bigW, _x*wid_bigX,_y*windbag, _z*wid_bigZ);

     l_char&id.=u_l_charen&id. + l_l_charen&id. + _lsp&id.*spwid +
_loth&id.*%sysevalf(&Ai10O+2);
     end;
     keep page x_2 x_s y_s y_e _annotats: annotats: fontsize key
_charalp: l_char: u_l_charen: l_l_charen: spwid _lsp: _loth: wid_:;
run;
%mend;
%annomap(indat=annot2,outdat=annot3,var=_charalp1,id=1);
```

Display: 14 Program for Calculation Annotation Width

CONCLUSION

This paper presents a unique perspective and method for managing CRF Chinese translation and CRF annotation using the Veeva Vault system. The proposed approach has practical value and is worth trying. As the tool becomes more mature and more studies adopt it, there will be more resources and economic benefits gained. However, as the tool is a combination of Python and SAS technical skills, it may present challenges for individuals who are solely focused on one language. Nonetheless, I believe that it could be fully developed using just one software alone.

Contact Information

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

Fei Chen
Eli Lilly and Company
188-1759-4458
CHEN_FEI1@LILLY.COM

Jie Wei
Eli Lilly and Company
WEI_JIE@LILLY.COM

Chengwei Li
Eli Lilly and Company
LI_CHENG_WEI@LILLY.com

Kewei Yuan
Eli Lilly and Company
YUAN_KEWEI@LILLY.com