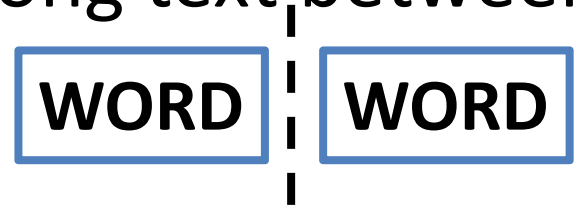
"SAS Macro to divide long text between words for CDISC compliance",

by Maksim Kazanski,
PAREXEL Informatics

Solution is:

1. Built for CDISC compliance for long text

2. Splits long text between words



3. Used in DATA step to divide into COLUMNS or RECORDS

SDTM IG v3.2

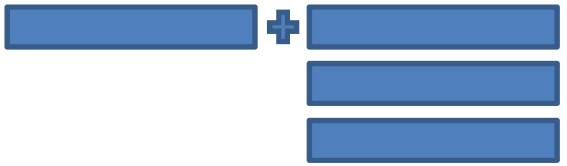
4.1.5.3.2 Text Strings> 200 Characters in Other Variables

- Because of the current requirement for Version 5 SAS transport file format, it will not be possible to store those long text strings using only one variable

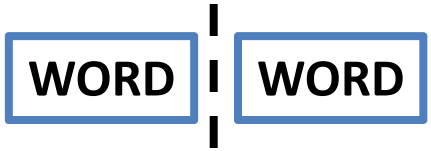
SDTM IG v3.2

4.1.5.3.2 Text Strings> 200 Characters in Other Variables

- The first 200 characters of text should be stored in the standard domain variable and each additional 200 characters of text should be stored as a record in the SUPP-- dataset …

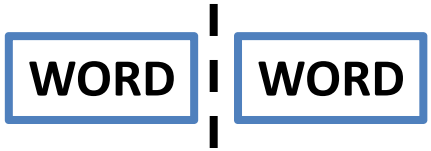


- When splitting a text string into several records, the text should be **split between words** to improve readability.



Split betwee
n words to i
mprove reada
bility

Split between
words to
improve
readability



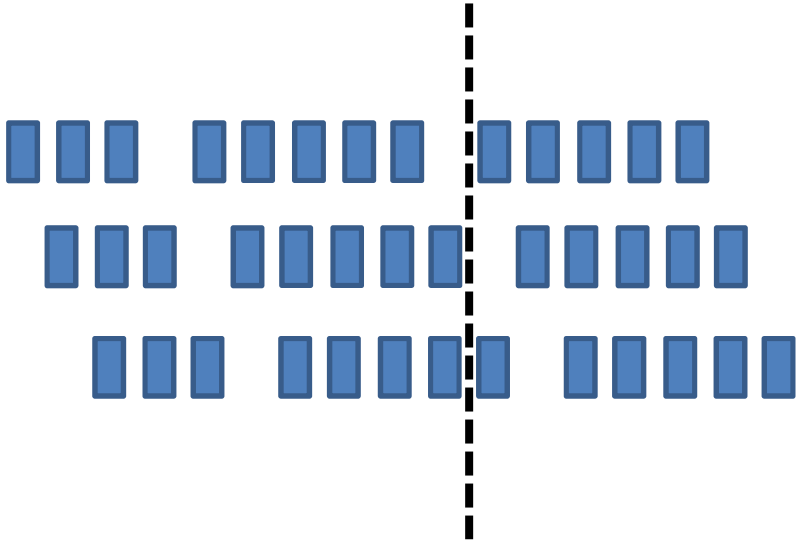
SDTM IG v3.2 Section-5 CO Domain CO - Assumptions for the Comments Domain Model

- When the comment text is longer than 200 characters, the first 200 characters of the comment will be in COVAL, the next 200 in COVAL1, and additional text stored as needed to COVALn. See example, Rows 3-4.

ROW	COVAL	COVAL1	COVAL2
3	First 200 characters	Next 200 characters	Remaining text
4	First 200 characters	Remaining text	

Close look at the task

1. Space is at 200 – cut at 200.
2. Space is right after 200. At 201 – cut at 200.
3. Last space is at 196 and word continues after position 200. Cut at 196.



The Longest Word
in the English Language

Methionylglutaminylarginyltyrosylglutamylserylleucylphenylalanylalanyl-
glutaminylleucyllysylglutamylarginyllysylglutamylglysylalanylphenylalanylvalyl-
Prolylphenylalanylyalylthreonylleucylglcycylaspartylprolylglicylisoleucy-
glutamylglutaminlserylleucyllysylisoleucylaspartylthreonylleucylisoleu-
cylglutamylalanylglyclyalanylaspartylalanylleucyglutamylleucylgluycyliso-
leucylproluylphenylalanyserylaspartyprolylleucelalanylaspartylglycylprolylthreo-
Nylisolleucyglutaminylasparaginylalanythreonylleucylarginylalanylphenylalany-
Lalanylalanylglycylvalylthreonylprolylalanylglutaminylcysteinylphenylalanyl-
Glglutamylmethionylleucyalanylleucylisolcucylarginylglutaminyllysylhistidyl-
Prolyuthreonylisoleucylprolylisoleucylglycylleucylleucylmethionyltyrosylalany-
Lasbaraginylleucylvalylphenylalanylsparaginyyllysylglycylisoleucylaspartyl-
Glutamylphenylalanylyltyrosylalanylglutaminylcysteinylglutamyllysylvalyl-
Glycylvalylspartylserylvalylleucylvallalanylaspartylvalylprolylvalvlglutaminyl-
Glutamylserylalanylprolylphenylalalrginylglutaminylalanylalanylleucylarginyl-
Histidylasparaginylvalylalalprolylisoleucylphenylalanylisoleucylcysteinylproly-
Prolylaspartylalanylaspartylaspartyspartyleucylleucylarginylglutaminylisoleucy-
Lalanylseryltyroslglycylarginylglycyltyrosylthreonyltyrosylleucylleucylsery-
larginlalanylglycylvalylthreonylglycylalanylglutamylasparaginylarginyla-
nylalanylleucylprolylleucylaspaaginylhistidylleucylvalylalanyllysylleucyllys-
sylglutamyltyrosylasparagimylalanylalanyprolylprolylleucylglutaminylglycyl-
phenlalanylglycylisoleyucylserylalanylprolylaspartylglutaminylvalyllysylalany-
lalanylisoleucylalspartylalanylglycylalanylalanylglycylalanylasoleucylse-
rylglycylserylalanylisoleucylbalyllysylisoleucylisoleucylglutamylglutaminyl-
histidylasparaginylisoleucylglutamylpronylglu-0-tamyllysylmethionylluecy-
lalanylalanyoeucyllysylvalylphenylalanylvalylglutamilylprolylmethionyllys-
sylalanylalanylthreonylarginylserine.

This word consists of 1,909 letters. It is the term for the formula C1289H2051N343O375S8. A Tryptophan synthetase A protein, an enzyme that has 267 amino acids.

Solution consists of four small macros to insert into DATA step

```
DATA … ;
```

1) `%co_vars(Length);` ➔ Create temporary variables

2) `%co_init(In Var);` ➔ Pass long text

3) `%co_piece(Out Var1..N);` ➔ Get first and following pieces

4) `%co_drop();` ➔ Drop temporary variables

```
RUN;
```

Used SAS Character Functions

- STRIP – Strip off blanks from both sides
- COMPBL – Compress multi-blanks
- TRANSLATE – Replace by characters
- SUBSTRN – Substring
- FINDC – Searches a string for any character in a list of characters. Can search backwards from any position.

1. Create temporary variables “w_..” Once per Data step.

```
%macro co_vars(pLen);
  %GLOBAL w_len;
  %LET w_len = &pLen;
  FORMAT
    w_cmm_rest $1000.
    w_last_space 8.
    w_end_pos 8.;
%mend;
```

➔ Global. Max length of string

➔ Buffer. Rest of the long text

➔ Position of the last space

➔ Position where to cut buffer

2. Pass long text.
Once per Long Text variable.

```
%macro co_init(pLongText);
  w_end_pos = 0;
  w_cmm_rest = &pLongText.;
  /* Replace   '0D'x Carriage Return,  '0A'x Line Feed,  '09'x Horizontal Tab
  with Space */
  w_cmm_rest = TRANSLATE(w_cmm_rest,'   ','090A0D'x);
  /* Remove double-spaces,
  leading and trailing blanks */
  w_cmm_rest = STRIP(COMPBL(w_cmm_rest));
%mend;
```

3. Get first, second, etc. piece.

```
%macro co_piece(pVarName);
  /* Find last space prior to w_len+1 */
  w_last_space = FINDC(w_cmm_rest,' ',-(&w_len.+1));
  /* If no spaces or space at w_len+1 use w_len */
  IF w_last_space > &w_len. or w_last_space = 0
    THEN w_end_pos = &w_len.;
    ELSE w_end_pos = w_last_space;
  /* Cut piece at calculated position */
  &pVarName. = SUBSTRN(w_cmm_rest, 1, w_end_pos);
  /* Remove piece from the buffer */
  w_cmm_rest = STRIP(SUBSTRN(w_cmm_rest, w_end_pos+1));
%mend;
```

4. Drop temporary variables.

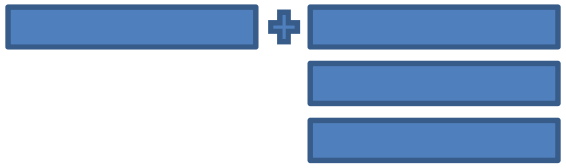
```
%macro co_drop();
  DROP w_cmm_rest w_last_space w_end_pos;
%mend;
```

Use for CO – domain.
Divide into columns.

```
DATA CO;
  SET CO_TMP;
  %co_vars(200);
    %co_init(COVAL_TMP);
      %co_piece(COVAL1);
      %co_piece(COVAL2);
      %co_piece(COVAL3);
      %co_piece(COVAL4);
      %co_piece(COVAL5);
  %co_drop();
  DROP COVAL_TMP;
RUN;
```

Use for SUPP-- qualifier. Divide into records.

```
DATA MH() SUPPMH();
  SET CO_TMP;
  %co_vars(20);
  %co_init(COVAL_TMP);
  %co_piece(MHTERM);
  OUTPUT MH;
  DO i=1 TO 5;
    %co_piece(QVAL);
    IF LENGTHN(QVAL)=0 THEN LEAVE;
    QNAM = 'MHTERM'||STRIP(PUT(i,Best2.));
    OUTPUT SUPPMH;
  END;
  %co_drop();
RUN;
```



Data Example

	id	MHTERM
1	1	____This is a very
2	2	_____This is a very
3	3	This is a very long

	id	QVAL	QNAM
1	1	long comment with	MHTERM1
2	1	space at position	MHTERM2
3	1	200. This is a very	MHTERM3
4	1	long comment with	MHTERM4
5	1	space at position	MHTERM5
6	2	long comment with	MHTERM1
7	2	space at position	MHTERM2
8	2	201. This is a very	MHTERM3
9	2	long comment with	MHTERM4
10	2	space at position	MHTERM5
11	3	comment with space	MHTERM1
12	3	at position 196.	MHTERM2
13	3	This is a very long	MHTERM3
14	3	comment with space	MHTERM4
15	3	at position 196.	MHTERM5

THANK YOU!