

Rectifying Irregular Text Data a case for using *Regular Expressions* in SAS®

Jayshree Garade, Cytel, Pune, India
Manjusha Gode, Cytel, Pune, India

SUMMARY

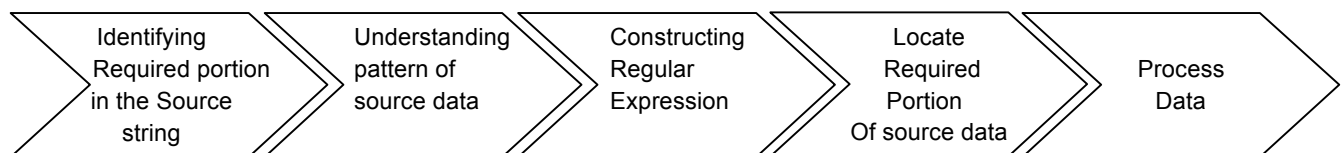
Users of the Clinical trial data frequently encounter problems when data are collected in 'free text' fields – in other words, loosely structured data. In many situations, these fields contain valuable insights and so it is worth to convert this inconsistent data into a more meaningful data. This can be well achieved using the technique of 'Regular expressions' by 'Pattern Matching'. In short, this conversion deals with the 'Irregular data' and transforms it into clean data.

Regular expressions provide a mechanism for describing character-string patterns. A host language, like SAS uses these patterns to match and perhaps change or delete the character string found during the pattern search.

RATIONALE

Many of the string processing tasks can be performed either with the traditional character functions or regular expressions in SAS. The advantage of using the regular expressions is the compact solution that they offer to the string manipulation problem. Regular expressions are especially useful for reading highly unstructured data streams. For example, you may have text and numbers all jumbled up in a data file and you want to extract all of the numbers on each line that contains numbers. One needs to define a pattern to search for, locate the position of the pattern by matching the pattern and then do the manipulation such as extracting a substring, or substituting the string with another string. Unlike the use of 'If then' or filtering the data using 'Where' clause, this technique allows the user to search for and extract multiple matching patterns from a character string in one step, as well as to make several substitutions in a string in one step.

STEPS TO USE REGULAR EXPRESSIONS:



PhUSE 2011

Paper #####

A REAL LIFE EXAMPLE – TIME POINTS

Most of the times, the time points data for safety studies is received in Character format. The screenshot below describes the data received in irregular format and then converted into systematic, well descriptive time variable using regular expressions in SAS.

BEFORE → **AFTER**

USUBJID	VISIT	VSDT	PRSDTLTM	VNTR_RT	VNTRTUN	Time_desc
1	1	17-Oct-08	Per 1 D01 Predose	47	/min	Predose
1	2	3-Nov-08	Per 1 D01 .5 hr	58	/min	Day 1 , 0.5 Hour
1	2	3-Nov-08	Per 1 D 01 01 hr	51	/min	Day 1 , 1 Hour
1	2	3-Nov-08	Per 1 d01 02 hr	49	/min	Day 1 , 2 Hours
1	3	4-Nov-08	Day2	53	/min	Day 2
1	90	3-Feb-09	Poststudy	56	/min	Poststudy