

Only keep certain characters in all Character Fields.

When you are being provided data from various data sources and systems there will be occasions when control characters appear in your data.

This data may be caused by moving the data between operating systems and the difference in character sets can cause this. Mainframes have a tendency to store spaces in hex 20 and when this data is migrated to a UNIX or PC environment the characters display as little boxes.

You may decide to have this removed at source but this is not always possible or expedient enough so having some way of removing unwanted characters or keeping only certain characters is an alternative and provides clean data from there on forward.

The simplest way of doing this is when reading the dataset for the first time perform a variation of the following code.

CODE:

```
data new;  
  set old;  
  
  array T_CHAR_VARS {*} _CHARACTER_;  
  do T_I = 1 to dim(T_CHAR_VARS);  
    T_CHAR_VARS(T_I) =  
compress(T_CHAR_VARS(T_I),  
compress(T_CHAR_VARS(T_I),  
  'ABCDEFGHIJKLMNOPQRSTUVWXYZ  
abcdefghijklmnopqrstuvwxyz01234567890'));  
  end;  
  drop T_;;  
run;
```

The array T_CHAR_VARS contains each of the Character variables from the OLD dataset and the DO loop cycles through each of these variables and performs the COMPRESS function twice. The first compress removes the characters from the list which are the characters you want to keep and the result (the unwanted characters) are then compressed out of the

original variable – leaving only the required characters.

This solution provides the technique – you may decide that rather than compressing these characters you want them converting to spaces and then you would use the TRANSLATE function.

I hope this is of use.