RELAIS

User's Guide

Version 2.3 (Preprocessing)

Editors:

Monica Scannapieco (DCMT/SVS/C)

Laura Tosco (DCMT/SVS/C)

Luca Valentino (DCMT/SVS/C)

Nicoletta Cibella (DCMT/U)

Tiziana Tuoto (DCMT/1)

Marco Fortini (DCCG/MTO/A)

Index

1. Pre-processing Functionalities

Preprocessing functionalities are available in the menu named “Preprocessing”. This menu is enabled after the dataset loading.

We distinguish two types of functions:

- *Check functions*
- *Conversion functions*

The detail of each function is described in the following sections.

1.1 Check Functions

Given a variable for which a format rule is available, a check function in RELAIS allows to create a list of the variable values that do not match the expected format rule.

This list is saved in a text file where the first line is the name of the variable and the remaining lines report inaccurate values. The inaccurate value list does not contain duplicated values or, if present, the NA value.

An extract of the inaccurate value list file looks like:

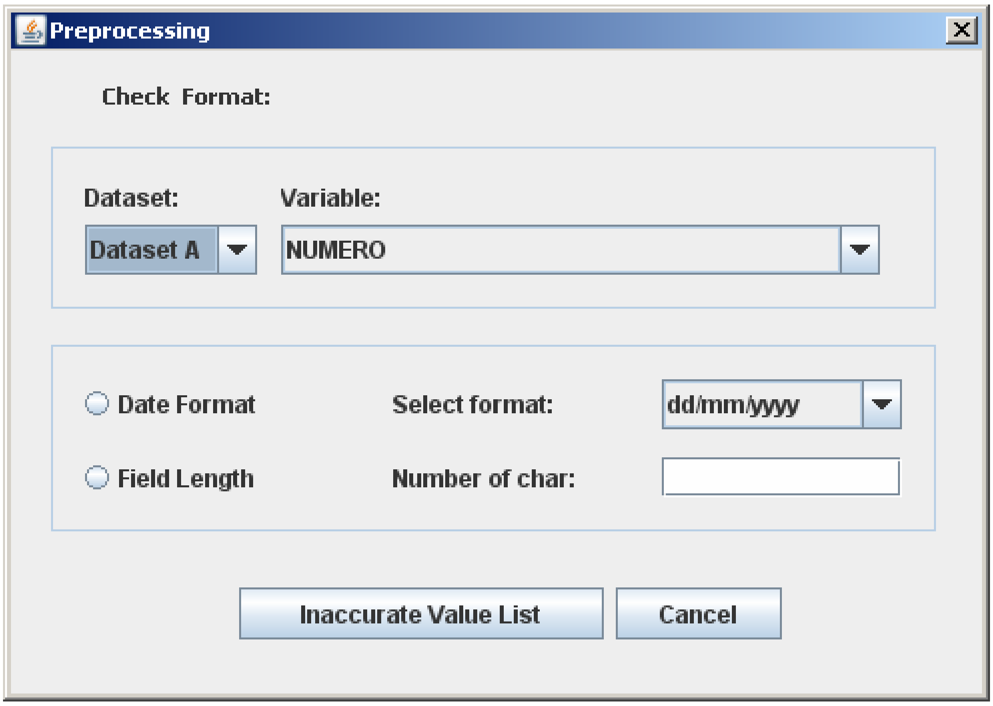
```
NATION
FRANC
FRANSZE
MESICO
NEERLANDS
```

The name of the file and the directory where it must be saved are specified by the RELAIS user.

The Check functions are available by choosing the item menu:

Preprocessing -> Check Format

The following window to select the format rule is open:



In the top of the window it is possible to select the dataset and the variable on which the check will be performed.

In the bottom of the window it is possible to specify the rule format. In this release you can specify a specific date format or a fixed value for field length.

Using the "Inaccurate Value List" button, you can choose the output file name and then the list will be created.

1.2 Conversion Functions

The output of a single conversion function operating on a variable is a new variable. The new variable is added to the dataset (while also maintaining the old variable). The new variable must have a name that is different from the original variable name.

The following features for data manipulation are available in the sub-menu 'Conversion function':

1.2.1 Field Standardization

The "Field Standardization" function allows performing a series of simple, but often very useful, cleanup tasks of the values of a variable.

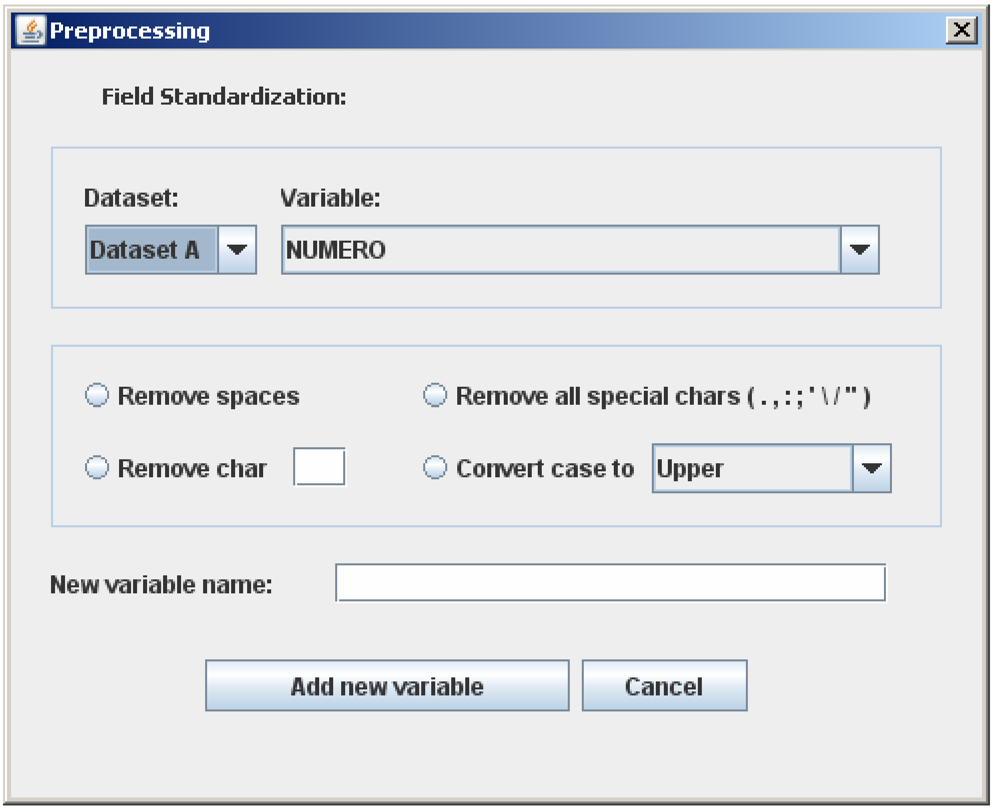
These cleanup tasks are:

- Remove spaces
- Remove special chars (pre-defined)
- Remove a char (user-specified)
- Case conversion

The Field Standardization functions are available by choosing the item menu:

Preprocessing -> Conversion Functions -> Field Standardization

The following window is open:



Preprocessing

Field Standardization:

Dataset: Dataset A

Variable: NUMERO

○ Remove spaces

○ Remove all special chars (. , : ; ' \ / ")

○ Remove char

○ Convert case to Upper

New variable name:

In the top of the window it is possible to select the dataset and the variable on which the check will be performed.

In the middle of the window it is possible select one ore more standardization task to perform.

In the bottom of the window it is required the name of the new variable.

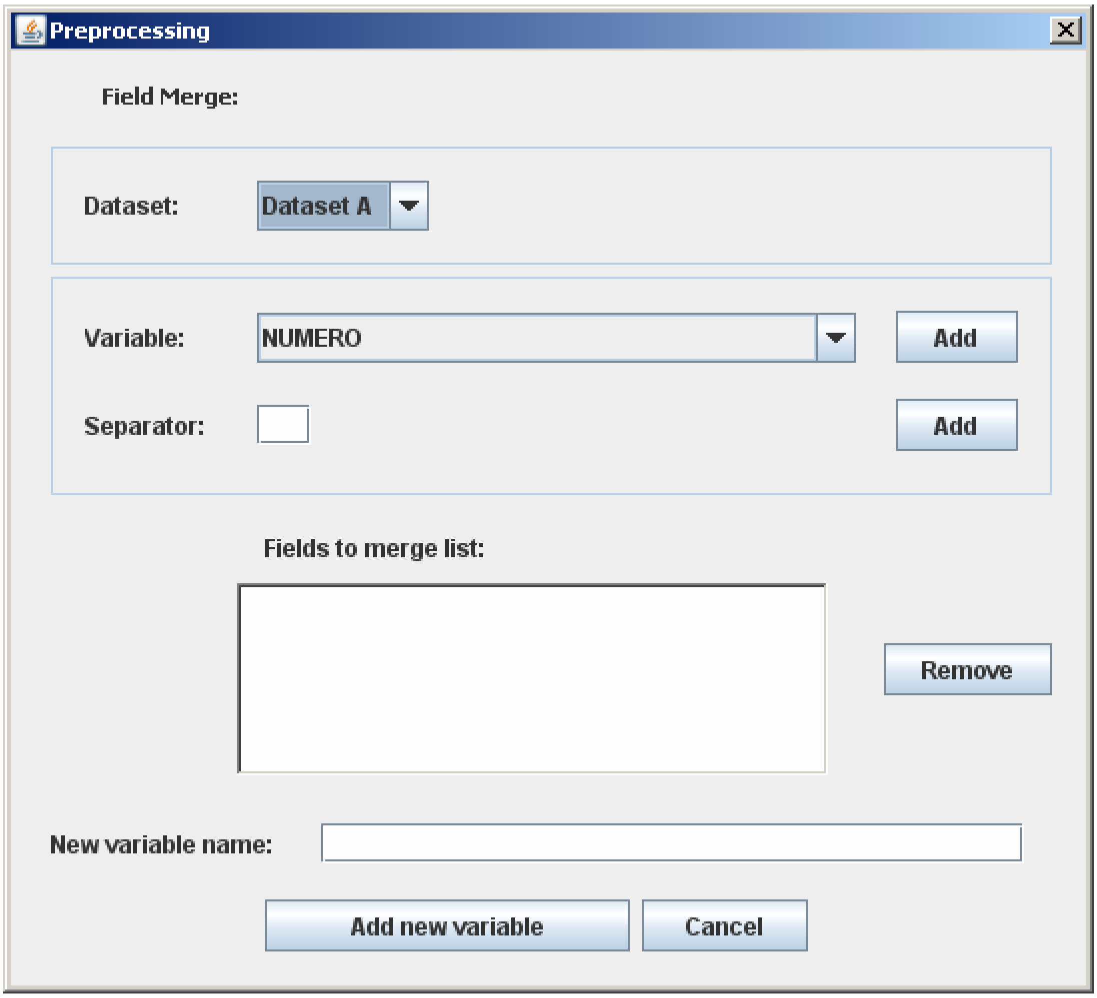
1.2.2 Fields Merge

The "Fields Merge" function allows creating a new variable by concatenation of filler and/or the existing variables of the dataset.

The Field Merge functions are available by choosing the item menu:

Preprocessing -> Conversion Functions -> Fields Merge

The following window is open:



In the top of the window it is possible to select the dataset.

In the middle of the window it is possible select one ore more existing variables to merge and, if desired, a filler as variable separator. The Add button commits the choice.

In the bottom of the window are shown the actual elements of concatenation and the required name of the new variable.

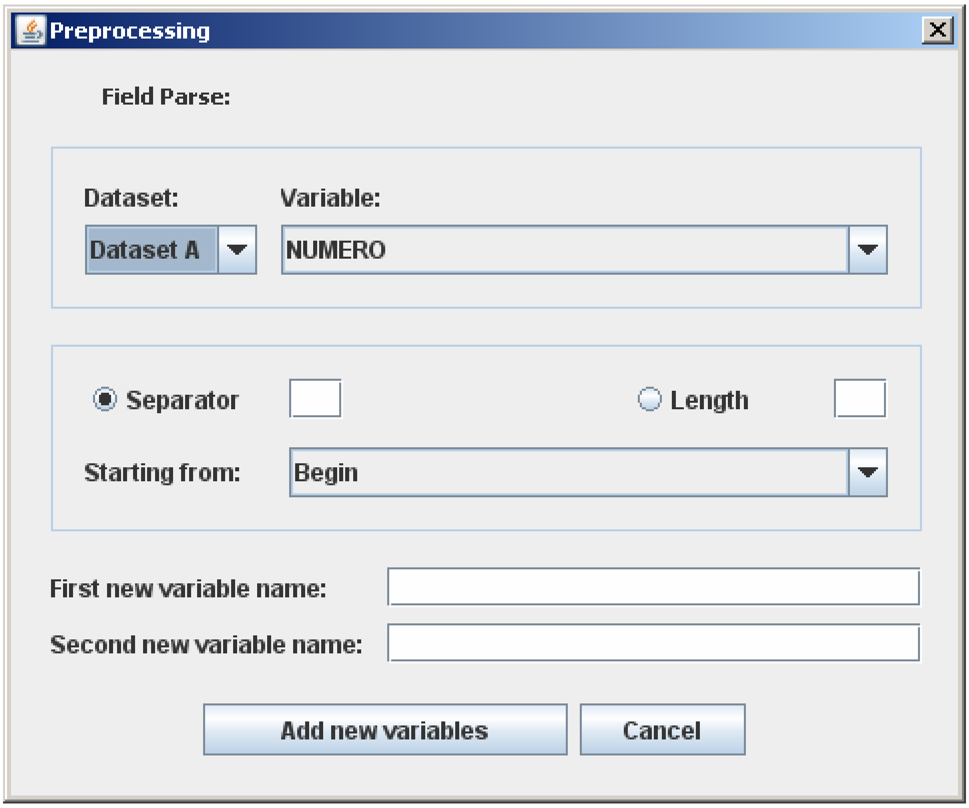
1.2.3 Field Parse

The "Field Parse" function allows splitting a variable creating two new fields in the dataset.

The Field Parse functions are available by choosing the item menu:

Preprocessing -> Conversion Functions -> Field Parse

The following window is open:



Preprocessing

Field Parse:

Dataset: Dataset A

Variable: NUMERO

● Separator

○ Length

Starting from: Begin

First new variable name:

Second new variable name:

In the top of the window is possible to select the dataset and the variable to parse.

In the middle of the window it is possible select the parsing rule. The variable can be split by specifying the number of chars from the beginning or from the end. In alternative, a separator char can be specified.

In the bottom of the window the names of the two new variables are required.

1.2.4 Inaccuracy Repair

The "Inaccuracy Repair" function allows creating a new variable for a dataset as a copy of an existing variable where the inaccurate values are repaired by corresponding "correct" values. The conversion of an inaccurate value with the corresponding correct value must be provided as an input by an external text file named "Conversion Input File".

The format of this file is the following:

- The first row of the file is a header with the name of the two fields (old variable with inaccurate values and the new variable).
- In each subsequent row, an inaccurate value has a corresponding value proposed as correct (separated by the char separator).

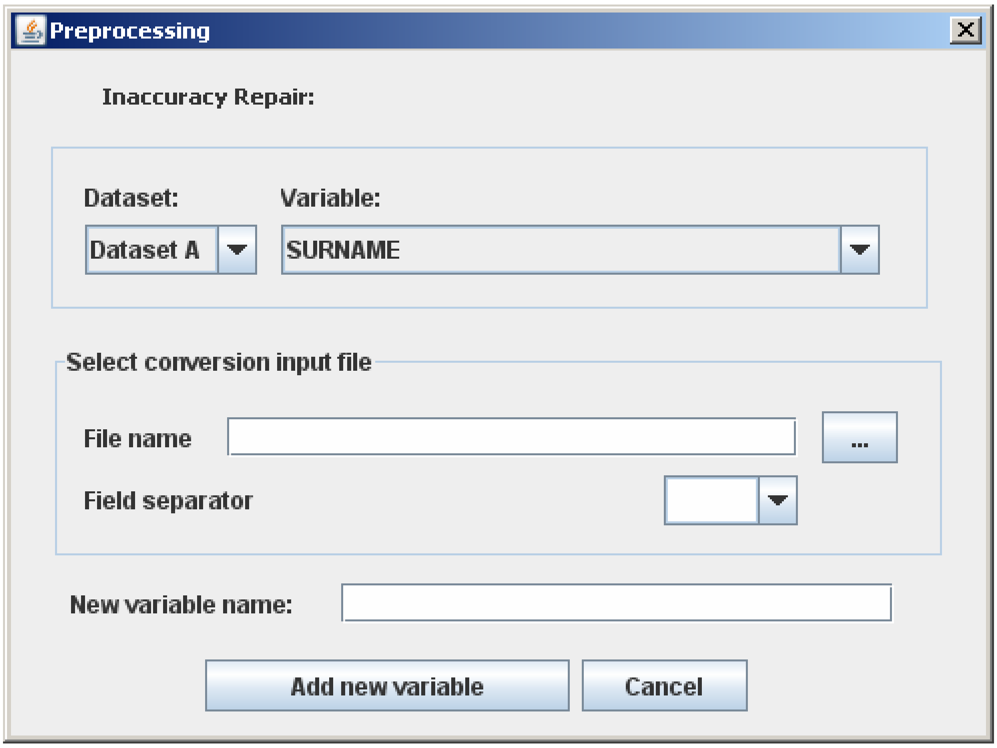
An extract of the conversion input file looks like:

```
NATION;NATION_CORRECT

FRANC;FRANCE

FRANSZE;FRANCE

MESICO;MEXICO

NEERLANDS;NETHERLANDS
```

The Inaccurate Repair function is available by choosing the item menu:

Preprocessing -> *Conversion Functions* -> *Inaccuracy Repair*

The following window is open:



Dataset: Dataset A

Variable: SURNAME

In the top of the window it is possible to select the dataset and the variable on which the check will be performed.

In the middle of the window it is possible to select the conversion file and the character to be used as field separator.

In the bottom of the window it is required the name of the new variable.

1.2.5 Schema Reconciliation

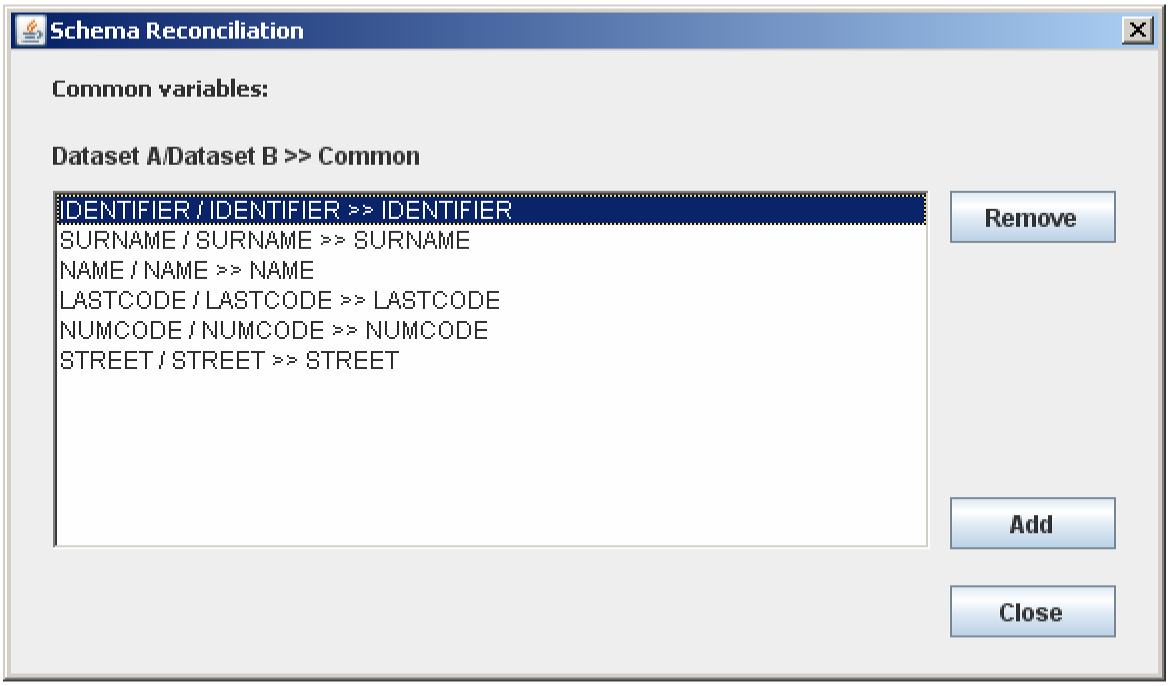
RELAIS performs automatically a schema reconciliation combining the variables of the two input datasets according to their names (read in the headers of the input files). The same reconciliation is performed for the new variables, created from pre-processing functions.

In addition to these automatic associations, RELAIS user has the opportunity to review them by the functionality "Schema Reconciliation".

These functions are available by choosing the item menu:

Preprocessing -> Schema Reconciliation

The following window is open:



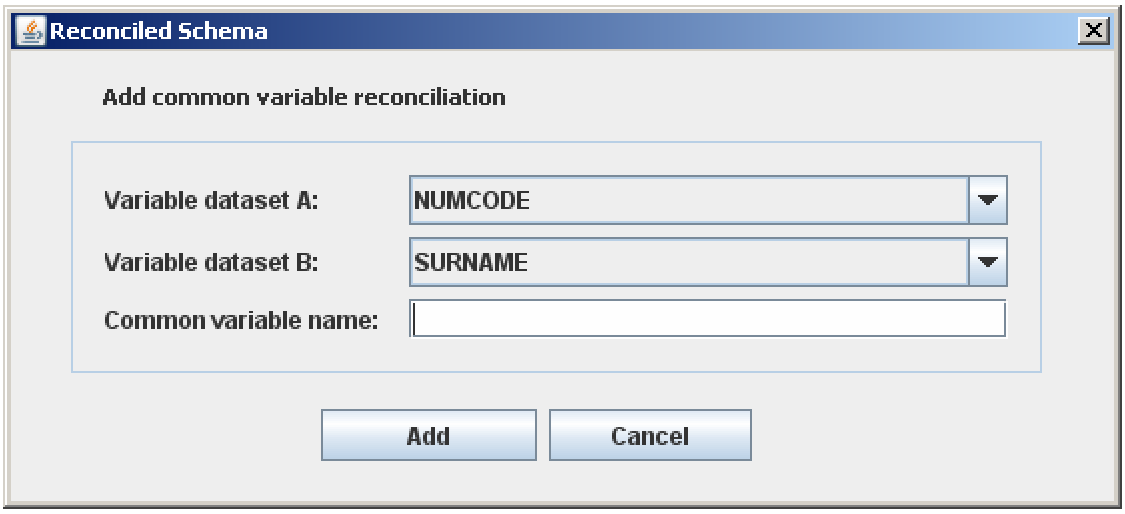
Schema Reconciliation

Common variables:

Dataset A/Dataset B >> Common

IDENTIFIER / IDENTIFIER >> IDENTIFIER
SURNAME / SURNAME >> SURNAME
NAME / NAME >> NAME
LASTCODE / LASTCODE >> LASTCODE
NUMCODE / NUMCODE >> NUMCODE
STREET / STREET >> STREET

Using the “Remove” button, you can remove the selected association. Using “Add” button you can create a new common variable by selecting “data set A variable”, “data set B variable” and the common name in the following window.



Reconciled Schema

Add common variable reconciliation

Variable dataset A: NUMCODE

Variable dataset B: SURNAME

Common variable name: