# MODULE 4:

# DEALING WITH MESSY DATA

BY AGGREY MUTIMBA – DATA RESEARCHER

13TH JUNE, 2014

Local voices. Global change.

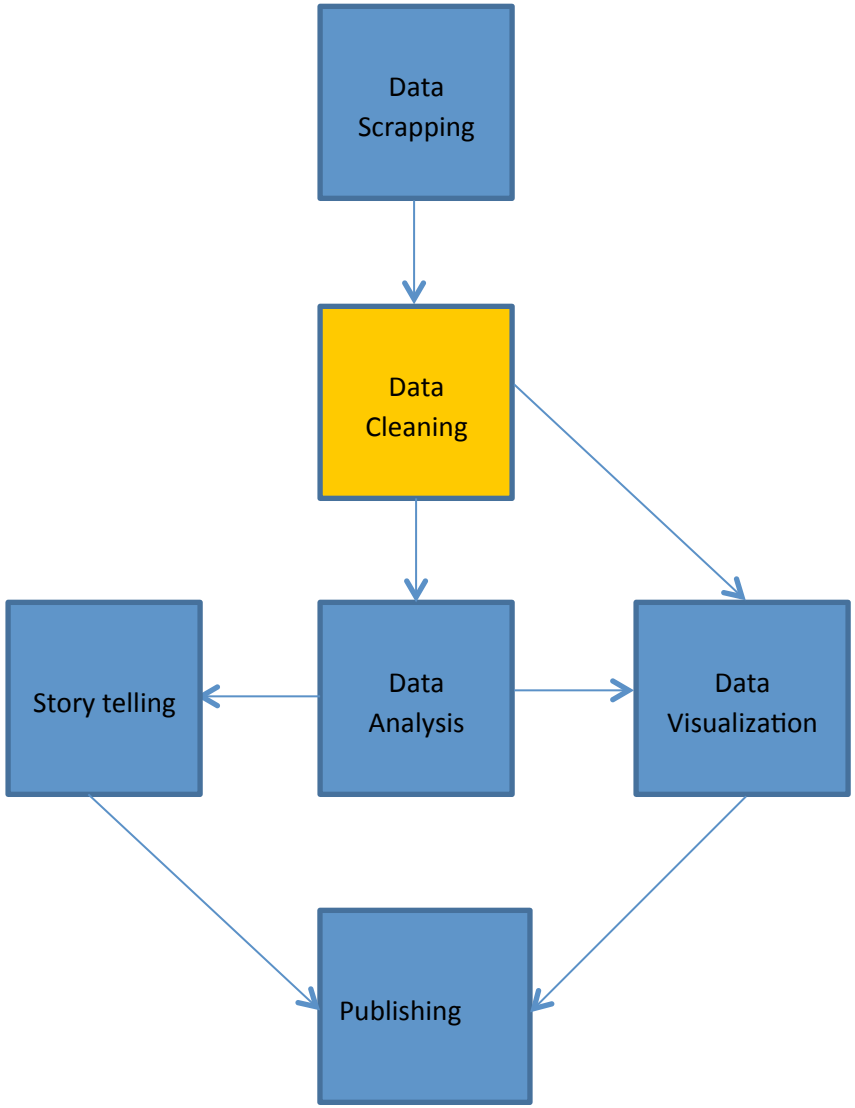
# Introduction

Before you can analyze the data, you often need to clean it up (for Human and Non human errors) especially when you have imported the data from a database, text file, or a Web page i.e. (xml, *.dta, *.csv, tab, *.sav, defined formats, dump.sql).

Data that is not clean could have the following:

- Missing data.
- Spaces before and after values.
- Outliers - data that is beyond range.
- Inconsistency.
- Different encoding.

# Figure 1: Data Journalism




Internews
Local voices. Global change.

# Tools used for cleaning data;

1. Spreadsheets like MS Excel
2. Open Refine (also called Google Refine)
3. Statistical tools: SPSS, SAS, R, STATA.
4. Writing code using scripting languages like Ruby, Perl and Python.

# Contents

**In this lesson, we are going to focus on the following;**

- Removing duplicate rows
- Finding and replacing text
- Removing spaces and nonprinting characters from text
- Fixing numbers and number signs
- Transforming and rearranging columns and rows
- Cleaning big data (Crowdsourcing)
- Navigating through Open Refine.

**Always create back-up before you embark on data cleaning**

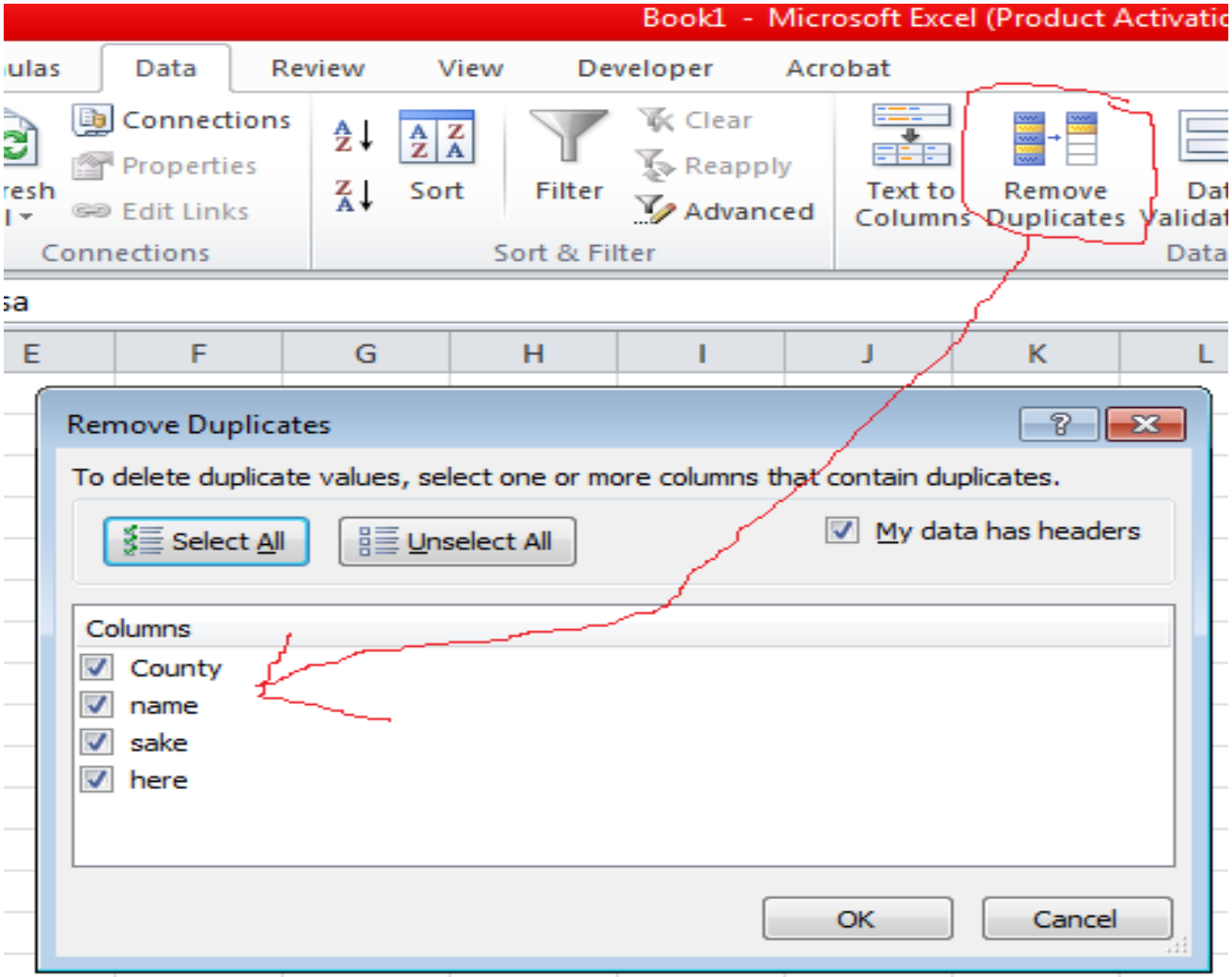
Internews
Local voices. Global change.

# Removing duplicate rows

Duplicate rows are a common problem when you import data. It is a good idea to filter for unique values first to confirm that the results are what you want before you remove duplicate values.

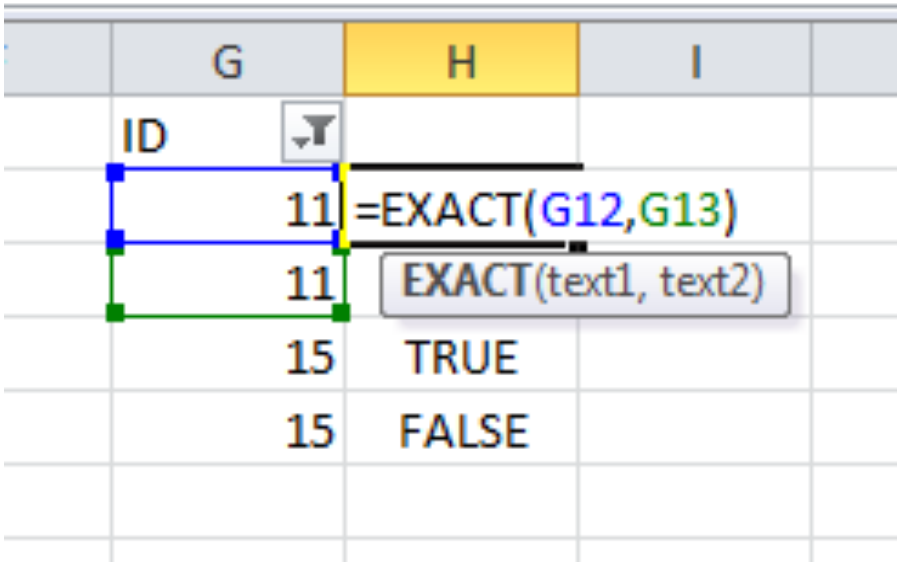
1. Download data from here.
2. Click in any cell in the worksheet.
3. Click on “remove duplicates” button as shown below.
4. Check columns that may have duplicates.

# Figure 2: Removing duplicate rows




Internews
Local voices. Global change.

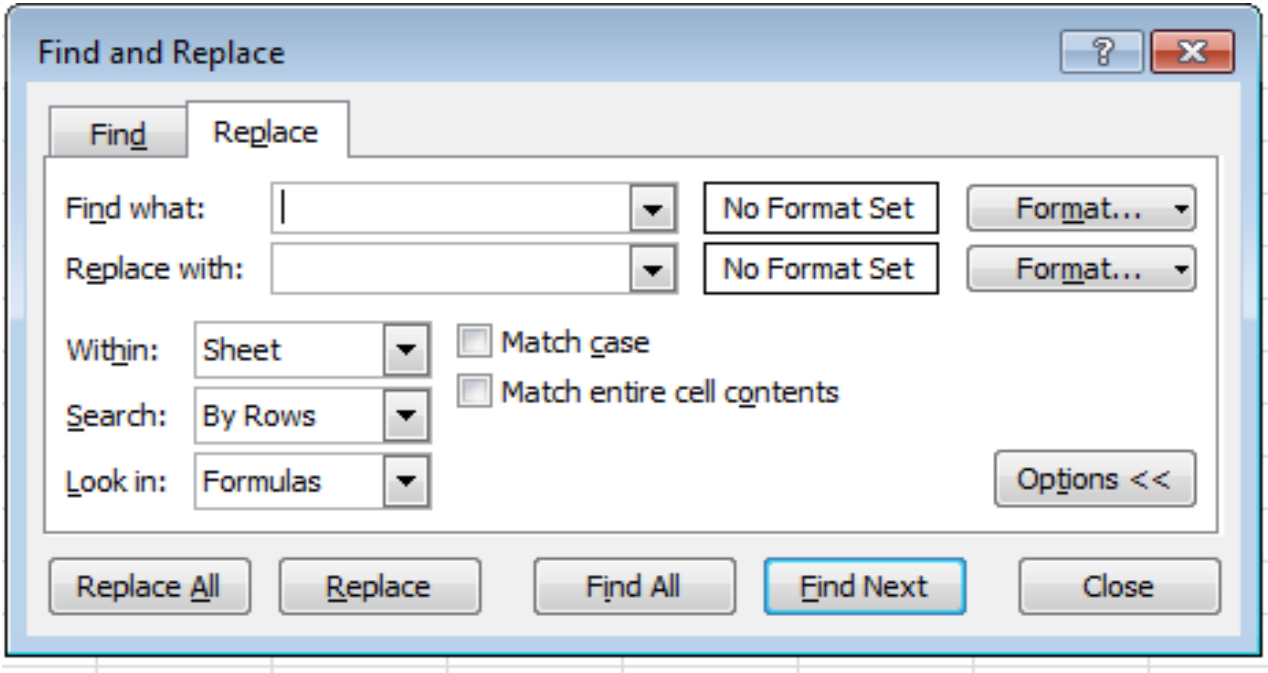
# Figure 3: Using EXACT() to remove duplicates




1. Sort the column you want to check for duplicates from lowest to highest.
2. Use exact function to check values that are repeated as shown.

Internews
Local voices. Global change.

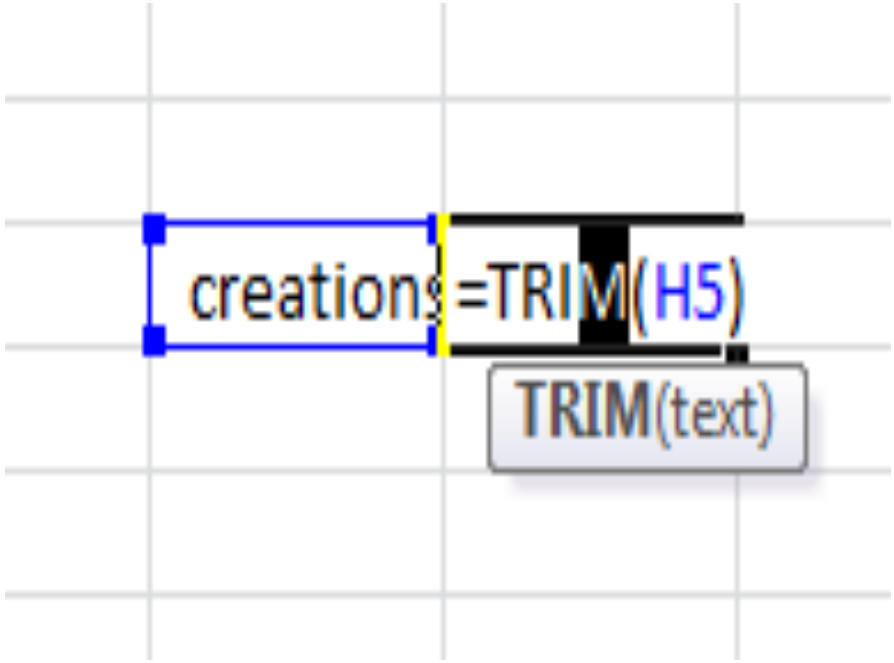
# Figure 4: Finding and replacing texts




The most common way to find and replace is to press ***Ctrl + F*** on your computer.

Internews
Local voices. Global change.

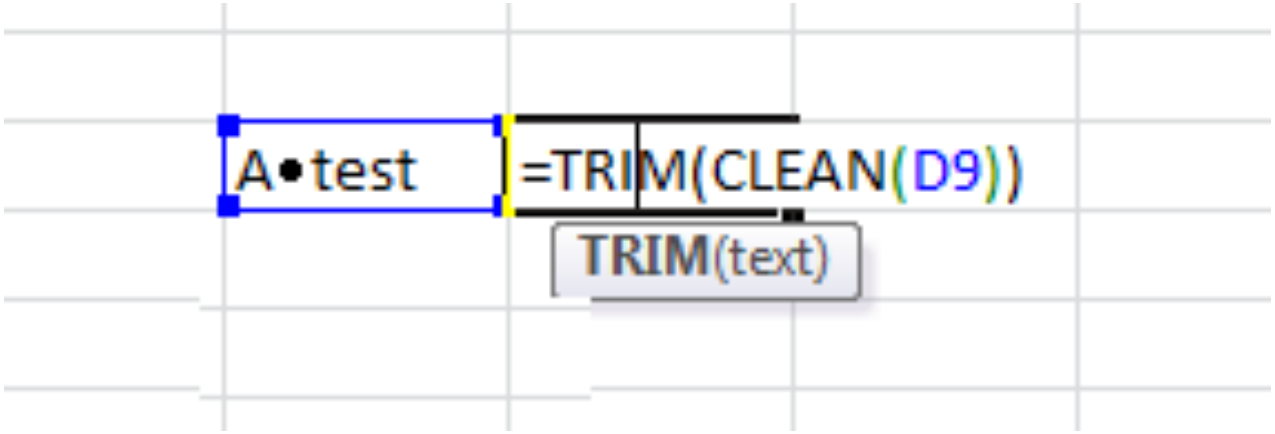
# Figure 5: Removing spaces from text using TRIM()




Using TRIM removes white spaces

Internews
Local voices. Global change.

# Figure 6: Removing nonprinting characters from text



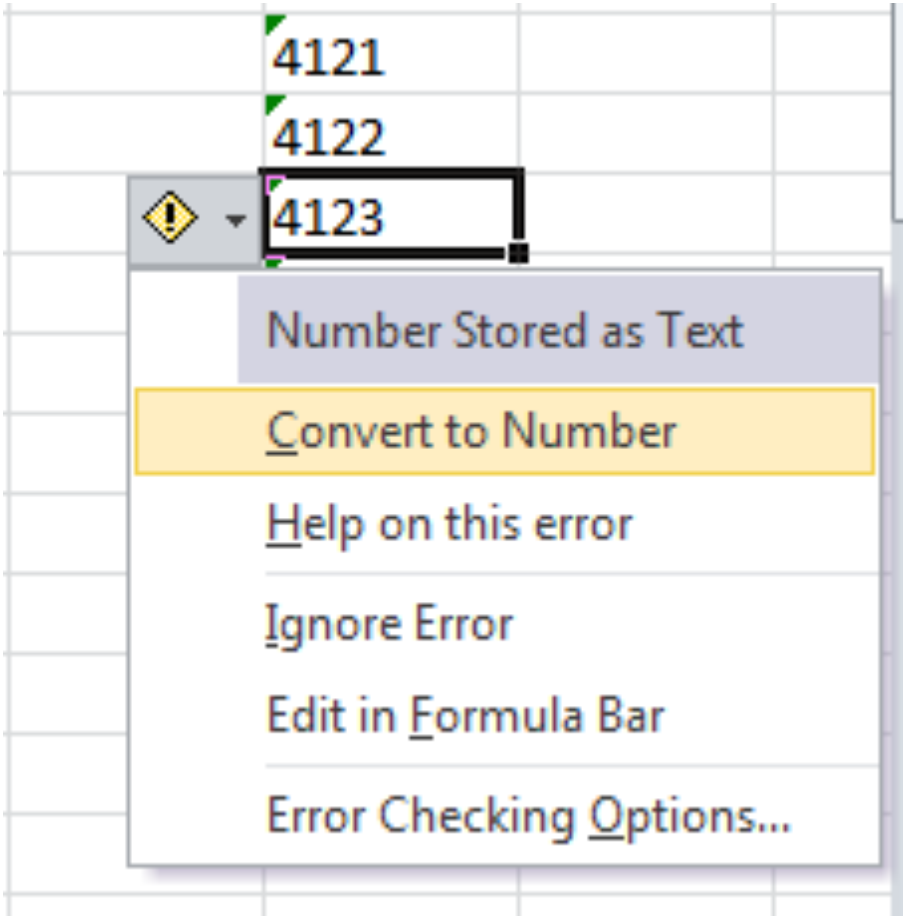

Adding CLEAN removes nonprinting characters i.e. CHAR(7) shown.

# Fixing number formats

One of the main issues with numbers that may require you to clean the data is when the number was inadvertently imported as text.

How to change it is shown in Figure 7.

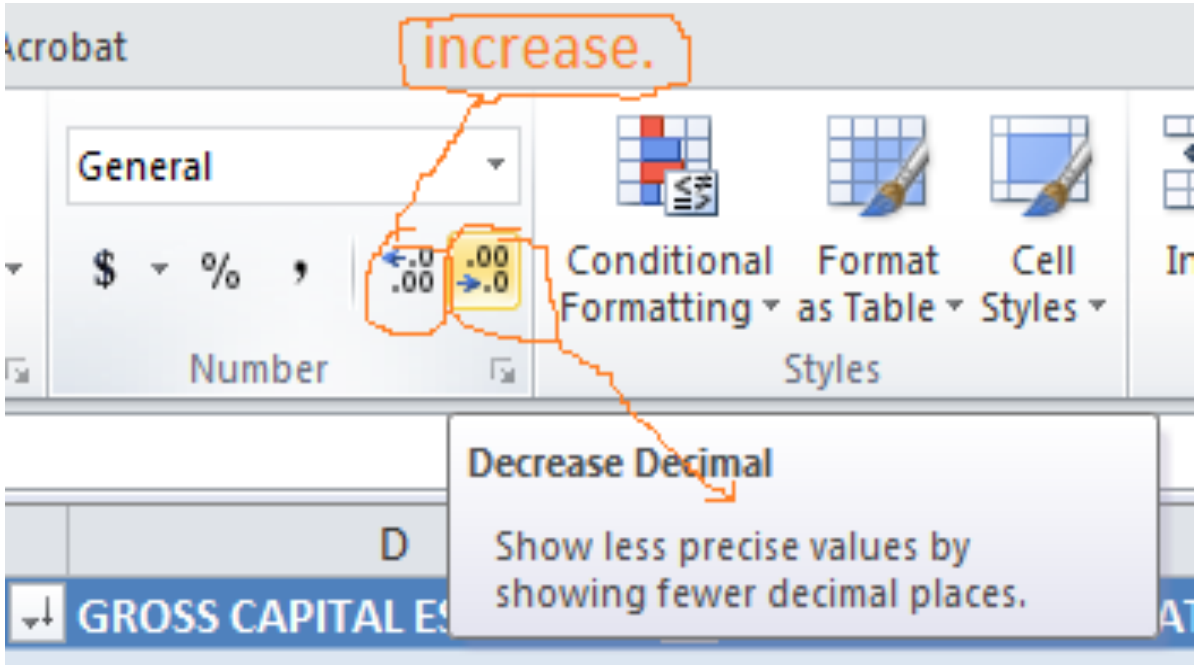
# Figure 7: Fixing number formats




Internews
Local voices. Global change.

# Figure 8: Changing the number of decimal places

Highlight the column you want to increase or reduce the decimal places and click on the buttons shown.

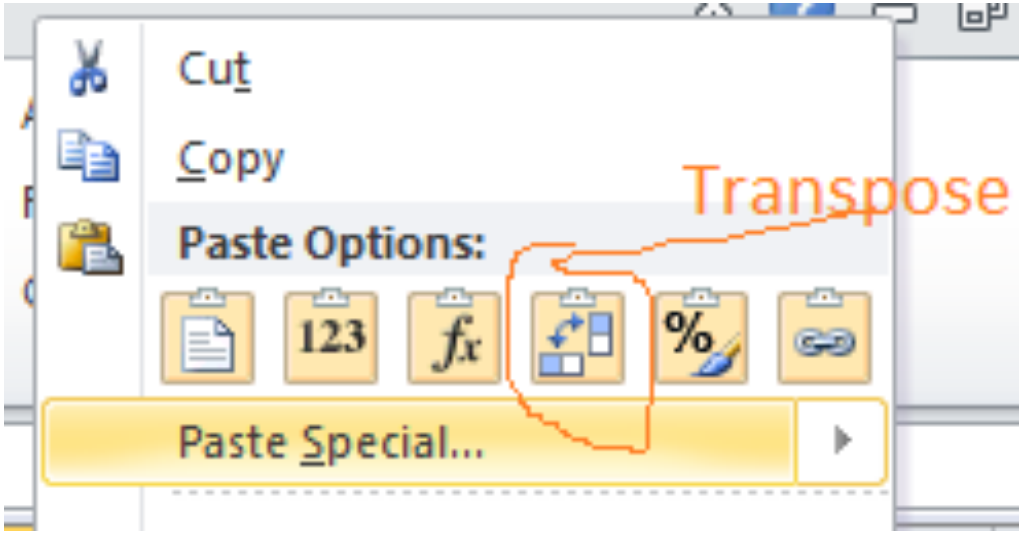



Internews
Local voices. Global change.

# Figure 9: Transforming and rearranging columns and rows

Sometimes you may want to change rows to columns or change columns to rows. You therefore have to transpose:

=TRANSPOSE(array)




Internews
Local voices. Global change.

# Cleaning Big Data: (CROWDSOURCING)

Asking crowd to clean existing data for you;

Always provide clear instructions otherwise data will be messed up! Examples of where this method is being used.

Free the Files: Help ProPublica Unlock Political Ad Spending (ProPublica)

Crowdsourcing app for data collection and cleaning - Login required (ProPublica)

# Navigating through Open Refine

# THANK YOU FOR YOUR TIME

**ANY QUESTIONS???**

Internews
Local voices. Global change.