

# Data-driven journalism: Step-by-step guide

Define **lead question**/  
hypothesis

**Find data**

Approach 1: Who could have such data  
Approach 2: Use a search engine to find datasets  
Approach 3: Use a search engine to find databases

**Check source**

Why was this data collected?  
How was this data collected?

**Understand data**

How can the data be read?  
What do abbreviations stand for?  
When have they been collected vs published?

Save original, **make a copy**, use that copy continuing

determine **data format**

Tables  
on **websites**

BrowserPlugin  
Copytables

Tables  
in **pdfs**

OCR Software  
**Tabula**  
(Abbyy Fine Reader €)

**csv**

**xlsx**

open in Excel:  
Text to columns

**Clean data**

Remove freeze  
Delete formates  
Delete unnecessary cells  
Remove references:  
Copy -> Paste Special -> Values  
Table head clean:  
1 row, columns named unambiguously

Trouble-Shooting

Text to columns did not work as it should have —> Use regular expressions to clean it further  
Slightly different spellings for the same thing —> Open Refine -> Text Facet -> Cluster -> Merge  
Charachters like ä, ö, ü not displayed properly —> Import data again, make sure to select the right encoding  
XYZ doesn't work —> Have you cleaned your dataset properly?  
I get a reference error #REF —> Copy -> Right click: Paste Special -> Values  
I'd like to exchanges rows and columns —> Copy-> Right click: Paste Special -> Transpose

**Analyse data**

Min/Max  
Median/Meand  
...  
5 more questions  
to ask your dataset

Merging datasets

**VLOOKUP** (both datasets need to have  
one column with identical content)

Summarize dataset

**Pivot-Table**

**Analysis results**

Start "normal" journalistic work  
find out the **cause** for your results,  
what **consequences** it has, talk with  
**people concerned/affected**

Talk about it

to **validate** your findings at least with  
- data provider  
- experts who know/use the data

**Visualize data**

find the right form:  
Abela Chart Chooser  
FT: Visual Vocabulary  
Dataviz Catalogue