**How to clean your data**

One of the biggest challenges when working with any data is dealing with errors. Often errors are not even noticed by data publishers because the data can change over many years. In other cases, errors can be the result of human mistakes in data entry, like mistyping or incorrect abbreviations.

When working with any data, it is important to know how to find errors and correct them to make the data more useful.

**In this module we'll explore the following:**

Common data errors

Useful data cleaning tools

Reasons for cleaning data

**Data cleaning - in 2 minutes**

ODI Trainer David Tarrant explains why clean open data is important, how to clean open data you find and what tools are available to support you.

**Revealing the value: what do we mean by clean data?**

Knowing that data is clean is important to ensure you have a reliable starting point when working with it and creating new value from it.

Clean data is data that is consistent, free from duplication and ready for machine consumption.

If data is clean, it is easier to combine it with different datasets and gain deeper insights.

**Common errors in data**

There are a number of common errors in data that you should look out for in any dataset you are working with.

# Wrong formats

Dates can be written in inconsistent ways. Mixed date formats are sometimes simple to spot:

8-sep-2013 vs 8/9/2013

and sometimes less obvious:

8/9/2013 vs 9/8/2013.

Another challenge is that some applications, like Excel, will ‘guess’ the format when a dataset is opened, which can hide how dates are actually written in the file.

The most common error is the mixed use of American (MM/DD/YYYY) and European (DD/MM/YYYY) formats.

# Multiple representation

People often try to save time when entering data by abbreviating terms. If these abbreviations are not consistent, it can cause errors in the dataset.

Differences in capitalisation, spacing and genders of adjectives can all cause errors.

# Duplicate records

A duplicate record is where the same piece of data has been entered more than once. Duplicate records often occur when datasets have been combined or because it was not known there was already an entry.

# Redundant data

Redundant data is anything that is not relevant to your work with the dataset. Often a dataset has been created for a specific purpose which requires details you may not need.

Common occurrences of redundant data include rows that represent total amounts.

Other times columns of data have been combined or replicated in order to assist human readability.

# Mixed numerical scales

Numerical values in datasets often use different scales to make it easier for a human to read. In budget datasets, for example, the units are often in the millions. 1,200,000 often becomes 1.2m. However, smaller amounts like 800,000 are still written in full. For a machine, this means they read the larger figure as 1.2, which causes errors.

# Mixed ranges

Data is sometimes measured in ranges, such as age or salary range. In order for a machine to understand these ranges it is important to separate the high and low values.

# Spelling errors

Spelling errors are often difficult to identify in tabular data.

Don’t forget to run your data through a spell-checker to catch spelling errors.

**The data-cleaning toolkit**

Without clean data, we can’t build information. There are a number of useful tools that can help the process of cleaning data.

# The data-cleaning toolkit

When you are looking for errors in data, it may be necessary to download and upload datasets in many tools for cleaning and processing.

It is also important to keep a note of what changes have been made and share these openly with others so that everyone can benefit from your work, particularly if the data you are cleaning is existing open data that has been published.

In this section we look at two of the main tools and their uses:

Open Refine

Excel and similar spreadsheet programmes

# Open Refine

Open Refine is a software tool designed to deal with uncleaned data. The tool is a column-based browser that allows you to fix errors across an entire open dataset in a single action. The errors that can be fixed include:

Date formats

Multiple representations

Duplicate records

Redundant data

Mixed numerical scales

Mixed ranges

# Spreadsheet programmes

Open Refine is a key tool for cleaning data. However, it is sometimes easier to fix some errors in a spreadsheet programme:

Spelling errors

Redundant data

Numerical verification

Fixing shifted data

**What are the benefits of cleaning data?**

In addition to the tools that can be used to clean open data, it is important to understand the benefits that cleaning brings.

# Prioritise cleaning

If data is not clean then decisions made on the basis of it could be wrong. This can affect products or insights you develop from data and, in some cases, have an impact upon thousands or even millions of people.

A [recent study](http://www.ercim.eu/publication/ws-proceedings/12th-EDRG/EDRG12_JeDiRe.pdf) found that the process of preparing data for analysis can take anywhere between 60% and 80% of the time of a data-driven project.

# Enhance usability

Sharing a clean open dataset will make the data more usable. Tools like Open Refine allow you to share change history, letting people know what has been cleaned and how. Making the process open as well as the data adds to the sustainability and trust a user can put in a dataset.

There are also data-cleaning communities emerging that could be supported more actively, such as OpenStreetMap communities.

[Take me to OSM Denmark Community Wiki](http://wiki.openstreetmap.org/wiki/WikiProject_Denmark)

# Discover insights

Many data-cleaning tools, including Open Refine, help you to combine and enrich datasets. This can lead to valuable insights being discovered.

For example, enriching a dataset that refers to businesses (by name or some other identifier) can inform you if the business is still trading and its current financial position.

[Get started with enriching data.](http://training.theodi.org/resources/odt/Enrichingdata.pdf)

**Are you ready to clean data?**

Can you recall the reasons why data needs to be cleaned, the useful tools to use and which common errors you should look for in datasets?

**Why does data need to be cleaned?**

**Uncleaned open data is a problem because:**

it cannot be plotted on a chart it cannot be published it can lead to wrong conclusions

**That's right!**

Unclean data is a problem because it can result in the wrong conclusions being made. This should not prevent the publication or use of the data, however cleaning must be performed.

**Are you sure?**

Unclean data is a problem because it can result in the wrong conclusions being made. This should not prevent the publication or use of the data, however cleaning must be performed.

**What are two of the most useful tools for cleaning data?**

Microsoft Access and Adobe Photoshop

Apple Notes and Microsoft Word

Microsoft Excel and Open Refine

**That's right!**

Open Refine and Excel (or equivalent spreadsheet editors) are two of the Module 9most useful tools for cleaning open data.

**Are you sure?**

Open Refine and Excel (or equivalent spreadsheet editors) are two of the most useful tools for cleaning open data.

**Recognising errors**

**Which of the following are errors in data that require cleaning?**

Mixed numerical scales

Lack of granularity

Redundant data

Lack of licence

Multiple representations

**That's right!**

Mixed use of scales, redundant data and multiple representations of the same data are just some of the errors that need to be fixed in unclean data. While the granularity and lack of licence are a problem, they are not related to the data itself being unclean **Are you sure?**

Mixed use of scales, redundant data and multiple representations of the same data are just some of the errors that need to be fixed in unclean data. While the granularity and lack of licence are a problem, they are not related to the data itself being unclean

**How much time can be spent cleaning and preparing data?**

**In any given data-driven project, how much time can be spent cleaning and preparing data?**

Between 20% and 40%

Between 40% and 60%

Between 60% and 80%

**That's right!**

Evidence suggests that between 60% and 80% of a data project's time will be spent cleaning, transforming and preparing data for analysis.

**Are you sure?**

Evidence suggests that between 60% and 80% of a data project's time will be spent cleaning, transforming and preparing data for analysis.

**Cleaning data**

Data-cleaning is essential to ensure that when the data is used, the right conclusions are made.

Often errors are not even noticed by data publishers because the data can change over many years. Errors can be the result of human mistakes in data entry, like mistyping or incorrect abbreviations.

Be aware that cleaning may take longer than you think. A careful cleaning process will, however, save substantial time in the long-run.

Follow our data cleaning exercise to get started with Open Refine and discover how easy it is for yourself.

Next module

Main menu

**Module 11 – video**

I’m Dave Tarrant, I’m the Trainer and Data Scientist here at the open data institute. One dataset will usually not be enough in order to develop new knowledge. It will be necessary to combine it and enrich it with further data in order to derive the insight. One of the biggest problems is the number of errors that exist in data. In an age full of computers, we still have extremely messy data and it is critical to clean these errors before you start to develop a story, for instance. Studies have found that actually up to 80% of the time can be used in preparing the data before you even get to telling a story and deriving this insight. Fortunately, there are a number of fantastic tools out there which can ease this burden. So one of the first things to do with any dataset is to ensure it’s clean. This may involve fixing spelling errors, removing misplaced data or normalising terminology. With the dataset clean, it may be then necessary to enrich the data to add other datasets to it, and could involve also normalising code so finding out what that number means, or to take a list of say company names, and find out the company number for instance that you can reference other sorts of data. And then, once you have this clean and enriched dataset you can really start to derive the insight. So fortunately, there is a fantastic open-source and free tool available to help you do this, called OpenRefine. OpenRefine is a power tool for dealing with messy data. Not only does it allow you to clean up the data and to export it into a number of different formats but it also enables you to enrich the data in a number of clicks so that you can add company data, location data and other general knowledge data to your dataset to really find the value and derive the insight.