Webinar: Introduction to OpenRefine

- A powerful tool for cleaning and preparing data before performing your data analysis

My first encounter … was via The Carpentries
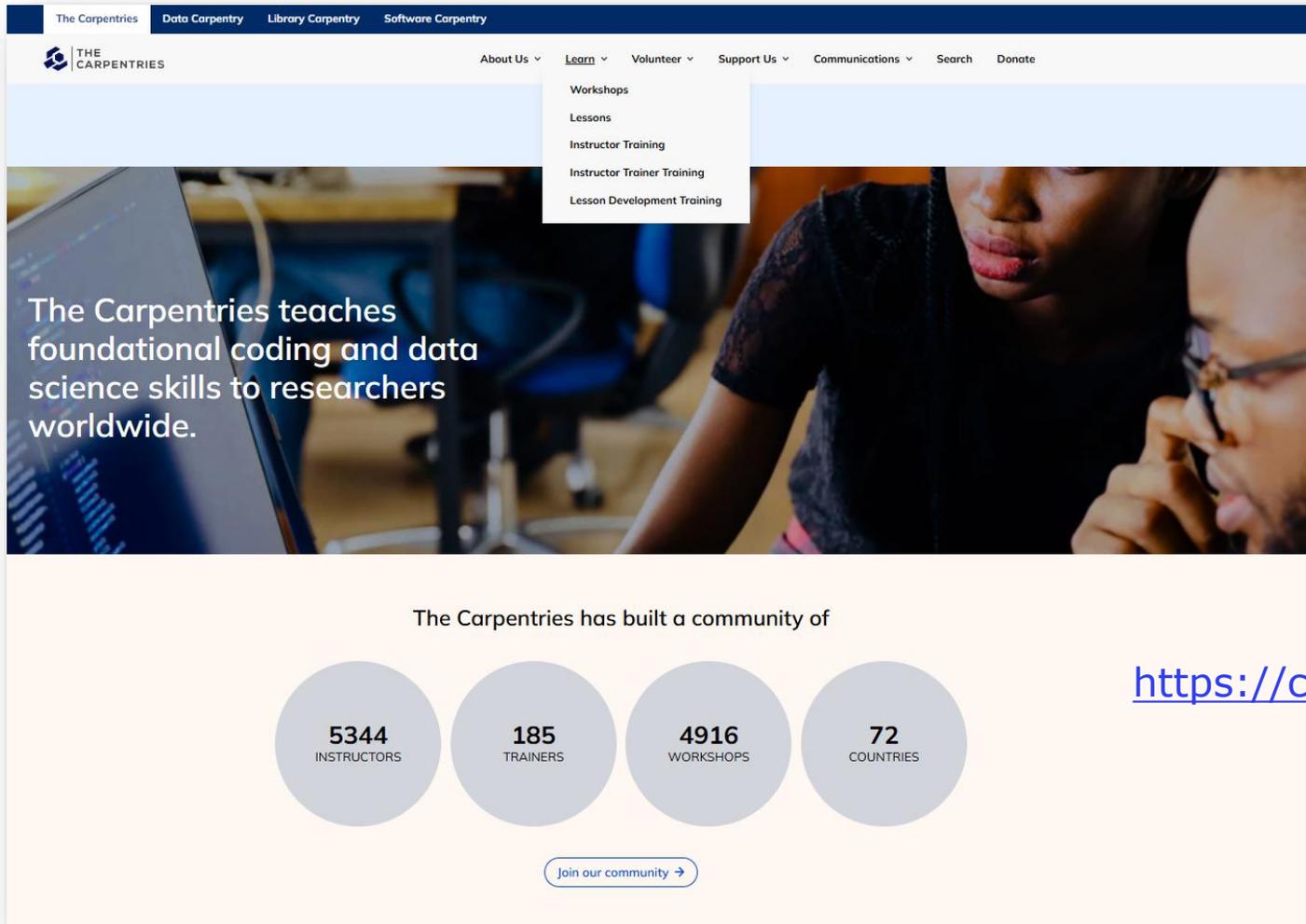
https://carpentries.org/

About OpenRefine

- OpenRefine is an **open source tool**
  - For cleaning and transforming **tabular data**
  - **The documentation** is available at https://github.com/OpenRefine/OpenRefine

- Began in 2009 as Freebase Gridworks, acquired by Google in 2010 & renamed as Google Refine. In 2012 it became known as OpenRefine, now a community-driven open source project
  [from: Using OpenRefine : the essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web (2013) - https://findit.dtu.dk/en/catalog/5e3c0e77d9001d016f56848a]

- It's **freely available**
  - …. Keeps track of all made changes/operations (reproducible & transparent)
  - …. Handles medium-sized datasets efficiently (up to a few hundred thousand rows)

- All changes take place locally and do NOT 'damage' the original data…

Cleaning data can be time consuming…

- Almost every dataset you'll encounter will be messy.

- Data might be entered inconsistently, including misspellings or extra white spaces, and that can be challenging to analyze the data afterward.

- It is important to be well acquainted with your dataset - and 'clean up your data' before using it in any way.
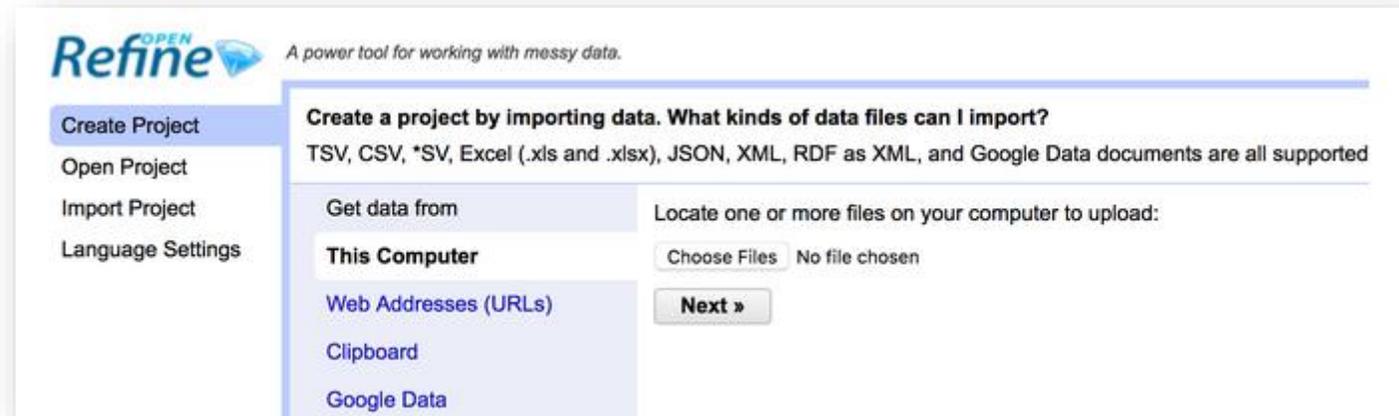
OpenRefine help you identify, explore and clean..

Easy documentation and reuse

....easily document how you cleaned your data and reuse the same "recipe" or script for similar datasets

Many publishers, journals, funding agencies, and institutions now **require clear documentation of the steps** taken in your data processing and analysis.

– Your cleaning process can be shared too

– This supports good data management practices and aligns with FAIR data principles

Your data stays private – since OpenRefine runs locally on your own computer

OpenRefine **never alters your raw dataset** directly—all transformations happen in a separate working copy

OpenRefine vs. Excel / Google Sheets
(Simple Comparison)

Feature	OpenRefine	Excel / Google Sheets
Ease of use	More technical; requires learning facets & transformations	Very easy; familiar spreadsheet interface
Handling large datasets	Excellent performance with big data	Slows down or crashes with large files
Automation & reproducibillity	Tracks every step; exportable workflows	Mostly manual; limited reproducibility
Power for cleaning messy data	Strong: clustering, bulk edits, complex transformations	Good for simple cleaning tasks
Collaboration	Limited real-time collaboration	Excellent live collaboration (Sheets especially)
Supported data formats	Broad: CSV, JSON, XML, RDF, API data	Primarily spreadsheet and CSV

Illustration created by use of AI by Jeannette Ekstrøm - from Copilot M365 2026-02-10

Examples on how Data can be inconsistent ..

Data you have	Desired data
1st January 2014	2014-01-01
01/01/2014	2014-01-01
Jan 1 2014	2014-01-01
2014-01-01	2014-01-01

Data you have	Desired data
London	London
London]	London
London,]	London
london	London

Address in single field	Institution	Library name	Address 1	Address 2	Town/City	Region	Country	Postcode
University of Wales, Llyfrgell Thomas Parry Library, Llanbadarn Fawr, ABERYSTWYTH, Ceredigion, SY23 3AS, United Kingdom	University of Wales	Llyfrgell Thomas Parry Library	Llanbadarn Fawr		Aberystwyth	Ceredigion	United Kingdom	SY23 3AS
University of Aberdeen, Queen Mother Library, Meston Walk,	University of Abderdeen	Queen Mother Library	Meston Walk		Aberdeen		United Kingdom	AB24 3UE

https://librarycarpentry.github.io/lc-open-refine/01-introduction.html

Getting started

1. Download OpenRefine from the official website

    https://openrefine.org/download

2. Install and 'launch' the program [runs locally – NOT in the cloud]

3. It opens your browser – into a 'localhost' window - http://127.0.0.1:3333

4. Get familiar with the interface, and **create a new project** with 'your dataset'

The Carpentries have dataset to download and use for 'self-paced learning'
   ❑https://librarycarpentry.github.io/lc-open-refine/
   ❑https://datacarpentry.github.io/openrefine-socialsci/ eller
   ❑https://datacarpentry.github.io/OpenRefine-ecology-lesson/

My examples are from taken from this lesson ☺

Local host – på egen PC http://127.0.0.1:3333/

https://librarycarpentry.github.io/lc-open-refine/01-introduction.html

I will only have time to do a few examples in this webinar

The actual schedule may vary slightly depending on the topics and exercises chosen by the instructor.

.io/lc-open-refine/index.html

These are some of the features I will show – LIVE…

Facetting:
Filter and analyze data by specific attributes.
https://librarycarpentry.github.io/lc-open-refine/04-faceting-and-filtering.html

Clustering:
Identify and merge variations (e.g., typos).
https://librarycarpentry.github.io/lc-open-refine/05-clustering.html

Sorting:
Reorder rows for better understanding.
https://librarycarpentry.github.io/lc-open-refine/06-working-with-columns.html

Transforming data:
Using GREL transformation (en smule kode-kendskab)
https://librarycarpentry.github.io/lc-open-refine/10-data-transformation.html

Working with 'Transformations' like

❑ splitting columns,

❑ trimming whitespace

❑ changing upper / lower case..

❑ add values, split values, ..

❑ replace values

… and several other advanced features

Common Transformation	Action	GREL expression
Trim leading and trailing whitespace	Removes any 'whitespace' characters (e.g. spaces, tabs) from the start and end of the current value	value.trim()
To titlecase	Converts the current value to titlecase (i.e. each word starts with an uppercase character and all other characters are converted to lowercase)	value.toTitlecase()
To uppercase	Converts the current value to uppercase	value.toUppercase()
To lowercase	Converts the current value to lowercase	value.toLowercase()

- Understand and learn some simple **GREL commands (General Refine Expression Language)**
https://openrefine.org/docs/manual/grel

Live presentation of

OpenRefine

A few take-aways

- **Undo/Redo:** Makes it easy to track the 'process' (reverse operations….)

- **Export** the cleaned dataset and export to other formats (CSV, JSON, etc.)

- **A simple overview of the 'data cleaning proces'**
  – step by step can be saved, and re-used for a similar dataset

- Easy to **save and share** the process you just went through
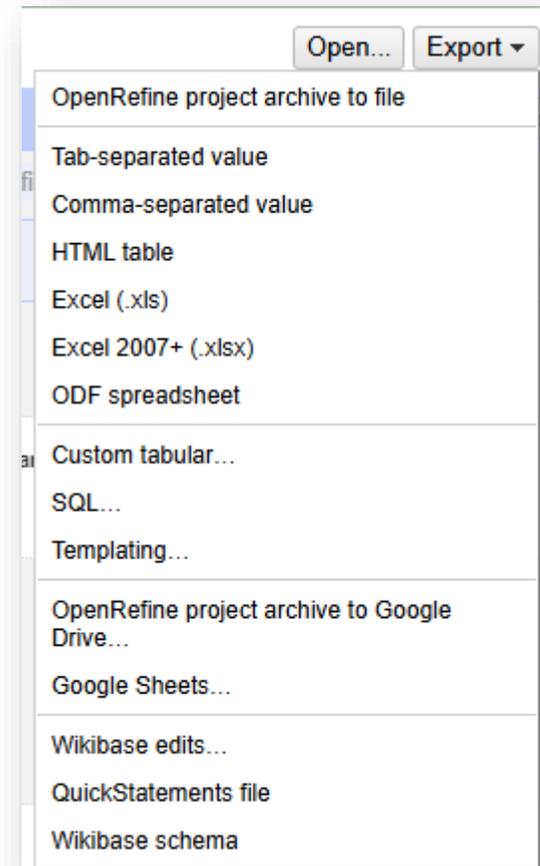
Supplementary materials & 'self paces learning'

- All about OpenRefine – incl nice introductory videos (and download links)
  https://openrefine.org/

- Official Forum (Q AND A for both beginner questions & problems) -https://forum.openrefine.org/

- Several community-developed add-ons to discover
  https://openrefine.org/extensions

- Digital Humanities too:
  https://programminghistorian.org/en/lessons/fetch-and-parse-data-with-openrefine

- The Open Educational Lessons from the Carpentries
  - Library Carpentry / OpenRefine - https://librarycarpentry.github.io/lc-open-refine/
  - Data Carpentry / OpenRefine for Ecologists - https://datacarpentry.github.io/OpenRefine-ecology-lesson/
  - Data Carpentry / OpenRefine for Social Science Data – https://datacarpentry.github.io/openrefine-socialsci/

Thank you all for dropping by our zoom room …

Time to wrap up !