BATCH INGESTION WITH OPEN REFINE

A kind of Holding Pen

http://openrefine.org

More from https://datacarpentry.org

TUTORIAL

based on https://tinyurl.com/jqrdmzen developed by Owen Stephens on behalf of the British Library

- Get an overview of a data set
- Resolve inconsistencies in a data set
- Help you split data up into more granular parts/statistics
- Match local data up to other data sets
- Enhance a data set with data from other sources

→ PREPARE DATA FOR UPLOAD INTO YOUR REPO

STEPS Summary

- Install and Run OpenRefine on your laptops: it a “local web app”

- Create a project in OpenRefine

- Some simple manipulations

- Amending data through facets, filters, manipulation of cells

- Regular Expressions to enrich the Data

Install and Run OpenRefine

Install OpenRefine 3.0 on your laptop :

[**http://openrefine.org/download.html**](http://openrefine.org/download.html)

Windows , Mac or Linux kit

Create a project

Get the Participant list in .csv format:
https://tinyurl.com/yap4n5ny (in .ods format:
https://tinyurl.com/ycm4slaq)

Create Project:

- Load the file
- Check the preview
  - Check the ‘Character encoding’
  - Ensure the first row is used to create the column headings
  - Doesn’t try to automatically detect numbers and dates
- Create Project

Simple manipulations

- **Reorder columns**: Switch “Name” and “Surname”

- **Rename column**: “Speciality” → “Profession”

- **Sort data** by “Countries”

- Filter the data using **Facets** : exclude the “Lecturer” - keep people “with/without a telephone”

- Filter the data using **Filters**: keep only people with a gmail account

PLEASE RESET !

Except the first action: column switch :-)

UNDO / REDO OPTION

Amending data

- Change “KENYA” into “Kenya” in all cells of the “Countries” column: ‘Edit’ a cell → ‘Apply to All Identical Cells’

- in Surnames: Edit cells’ → ‘Common Transforms’ → Remove leading/trailing space

- in telephone numbers: → Collapse consecutive white space

- Harmonize “Assistant Library” & “Library Assistant” into - using ‘Cluster and Edit’

- Remove “University” in the City column

Enriching data with ‘GREL’ (Google Refine Expression Language)

- **Remove** “University” in the City column -> Edit Cells -> Transform: type in *replace(value,"University","")*
or *value.split(“ ”)[0]*

- Create **new column “Fullname”** with the values “Surname, Name”

From Surname → Edit Column → “Add column from this column” → type in: *value + ", " + cells["Name"].value*

Enriching data looking up from a URL

- Adding the ORCID values for participants !
- From Surname→ ‘Edit column’ -> ‘Add column by fetching URLs’. Expression to be entered: *“https://oaa.tind.io/search?cc=Authorities&of=t&ot=100__g&p="+ escape(value, 'url')*
- Let running the fetching of ORCIDs

Export your data

- Select the columns you want to export
- Make sure the labels are the ones you want →
- ‘Export’ in ‘Comma separated values’
- ‘Export’ with ‘Templating’