E07 – ETL (OpenRefine)

Business Intelligence

Exercise

Winter Term 2025/2026

Agenda

- Introduction
  - Introduction
  - Notes on Software
- Exercise
  - Tasks
- Credits and Materials

Introduction

The main goal of the “ETL” exercise is to introduce students in the world of the business intelligence and data analytics using one of the available out there data analytics software tools.

There are a number of ETL tools available on the market and you should already be familiar with core features of these ETL tools – extract, transform and load of a data into a data warehouse or a database.

Despite this set of exercises focused on a limited number of software tools, obtained knowledge and hands-on experience could be very useful and could be easily re-used for other similar software solutions.

In this exercise you will be using “OpenRefine” software.

Notes on Software

See the materials from the previous exercise **E02, E03** and **E05**

Additional instructions on the infrastructure (online platform) for this course can be found in
E00 - Infrastructure

Start

- In this exercise the “OpenRefine” data wrangling and transformation software will be used
- There are 3 possibilities to get access to this tool
  - Use docker
    - https://github.com/OpenRefine/OpenRefine/wiki/Alternative-Installation-Types
    - https://github.com/openculturereconsulting/openrefine-docker
  - Use jar file - https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions
  - Use provide to you instance of OpenRefine - for further details refer to “**E00 – Infrastructure.pdf”**
- The special datasets are also required to work on this exercise
  - See “Supplementary Materials” of further details

To save computation resources on the server, all running services are stopped at midnight.

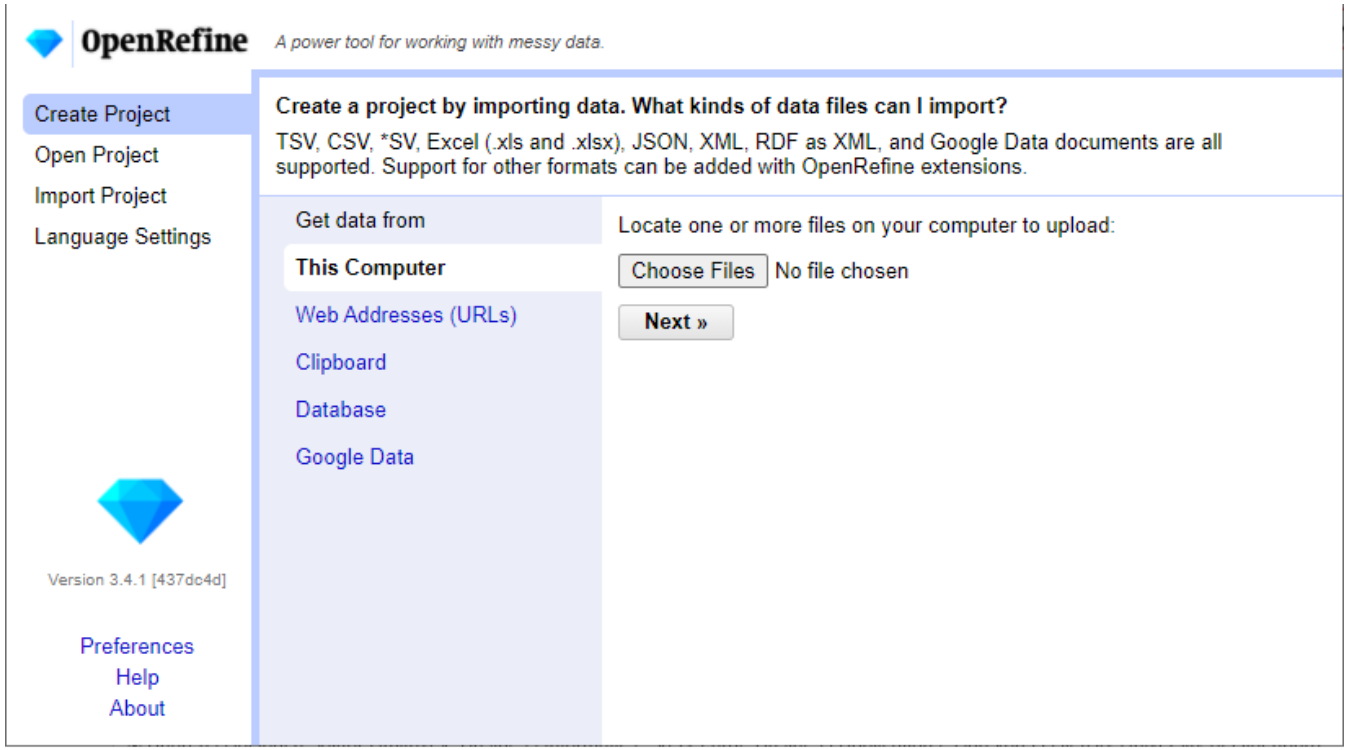
So, before you start to work on your exercise, you have to start your service manually.

To manually start your service, use provided to you web interface of the infrastructure platform.

Exercise

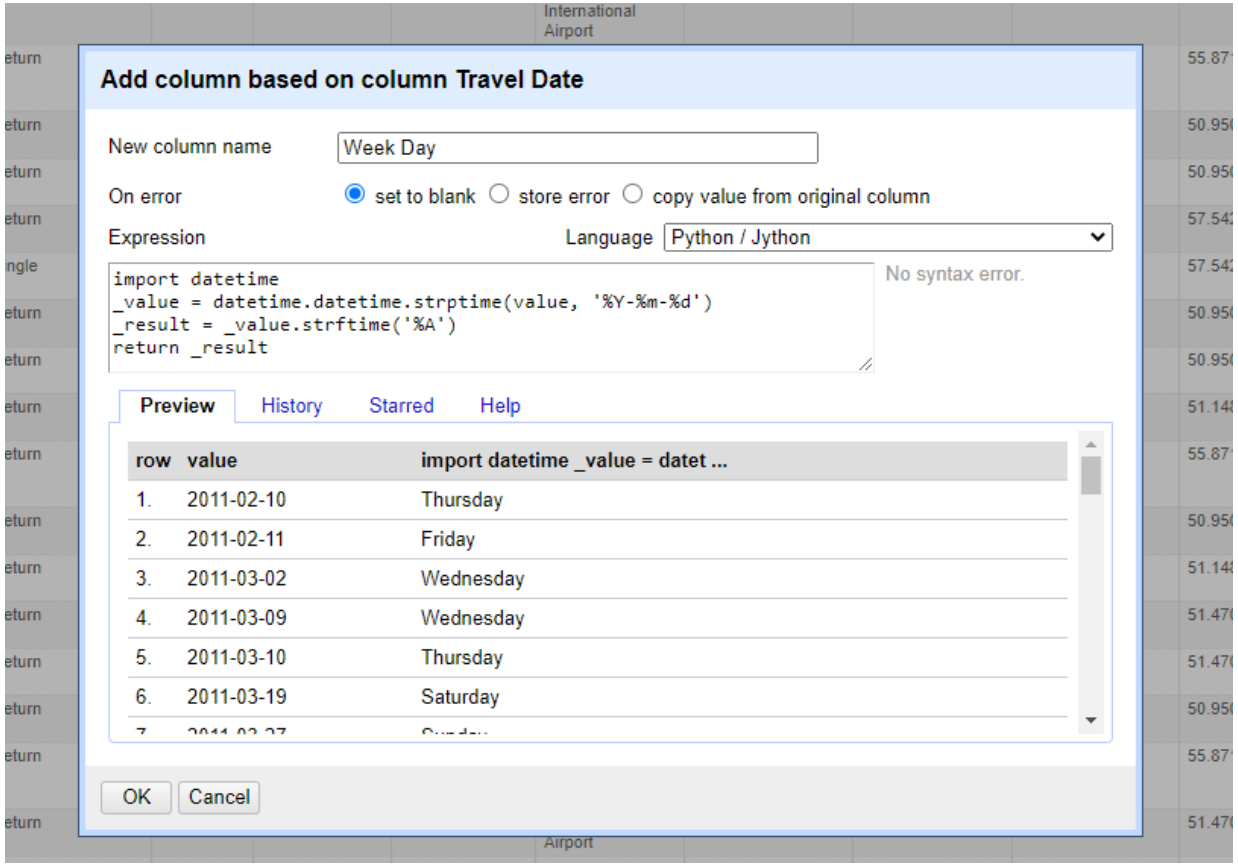
Description - OpenRefine

OpenRefine (previously Google Refine) is a standalone open source desktop / web application for data cleanup and transformation to other formats. It also could be refereed as a data wrangling software tool, which work similar to once that are primary focused on tabular data. OpenRefine empowers data wrangling, simplifies process and offers a number of useful feature for data driven projects out of the box.



Technical Hint – OpenRefine + Python

This example shows, how a column could be used to create a new column, where new values will be based on the old column. In this particular case Python is used to get the day of the week based on the given date from the old column.



Tasks

Tasks – Part I

1. In order to get familiar with the tool follow this tutorial on OpenRefine “Using OpenRefine to Clean Your Data”
   - https://multimedia.report/tutorials/openrefine/
2. In addition to the tutorial perform following manipulations on the dataset provided to you:
   - Create a new column that contains length of the “Name of project/news package”
   - Create a new column that contains length of the “given name”
   - Create a new column that contains length of the “surname”
3. Answer following questions
   - What is the longest given name in this dataset?
   - What is the longest “Name of project/news package”
   - What is the average length of “Name of project/news package”

Tasks – Part II

1. Import data from the "dates_synthetic.csv" file into your OpenRefine project
2. Create a new column that contains the first 2 octets of the provided IP address
3. Create multiple new columns with the following content for each date in the column “datetime_eu”
   - day of the week
   - week of the year
   - month of the year (as number)
   - month of the year (as string - 'January')
4. Homogenize given date and time for all columns with datetime value, use the following format "%Y-%m-%d %H:%M"
5. Calculate “time delta” with today's date for the column with the date "datetime_eu"

Tasks – Part III

1. Import data from the "UOL News.csv" file into your OpenRefine project
2. Check, if provided within the dataset URLs are still available
3. Fetch data from a URL using "Add Column By Fetching URL“ feature
4. For the each “item” below, create a separate column based on the downloaded content for each URL:
   - page title
   - publication date
   - content of the main article for each given URL
   - *(optional - remove HTML tags)

Tasks – Part IV

1. Import data from the "phm-collection-4000.tsv" file into your OpenRefine project
2. In the column “Categories” change values into uppercase
3. What is the oldest item in the dataset according to the provided information/data?
4. Create a new column, which contains the length of the "Description“ column
5. Create multiple new columns and keep only numbers inside newly created columns. Do it for the following columns “Height”, ”Width”, ”Depth”
6. Create a new column that will calculate volume of each item (if measurements are provided). Which measurement unit should be used here?

Submission

- Use StudIP to upload your solution (PDF report)

- You should upload one of the following document
  - Report as a PDF file with your solutions
  - Each answer (e.g., SQL query) should be annotated by you
    - Original question (+ number of the task)
    - Answer

- Name convention for your submission file (without extension)
  - **E07**_FIRSTNAME LASTNAME

- Submission deadline (it is a “soft” deadline)
  - 10 days after this exercise starts
  - Some exercises could take a bit more time and could be submitted later
  - NOTE: to receive feedback, you should first submit your progress

Credits and Materials

- https://multimedia.journalism.berkeley.edu/tutorials/openrefine/
- *https://openrefine.org/docs/manual/starting*
- https://openrefine.org/docs/manual/columnediting#add-column-by-fetching-urls
- *https://guides.library.illinois.edu/openrefine/duplicates*
- OpenRefine Beginners Tutorial (2019) - https://www.youtube.com/watch?v=wfS1qTKFQoI