Assignment 1
General instructions

Goal of the project:

The assignment consists in processing different datasets shared by researchers/library staff of the KU Leuven. The goal of this assignment is that the students

1. Familiarize themselves with OpenRefine and, in particular, its “reconciliation function”
2. Establish an efficient workflow
3. Manipulate regular expressions

The goal of the resulting dataset won’t be directly evaluated, however it will be a direct consequence of the previous points.

The datasets:

The students have been assigned 2 different typologies of datasets: datasets containing metadata about ancient books present at the Library of Leuven and datasets containing metadata about paper periodicals catalogued by Kadoc (https://kadoc.kuleuven.be/english/1_KADOC/index). The metadata, originally provided in MARCXML have been transformed into a CSV format. All the datasets will be uploaded on Toledo and every group will receive the information of what dataset has been assigned to them. Additional documentation will also be uploaded to Toledo.

Methodology:

Create a project in OpenRefine and upload the dataset (see the documentation specific to each dataset for this passage). The functions of OpenRefine are documented on the Website of the tool (see links below), and most of them are available just by clicking on the icon at the top of the column. The specific tasks are designated in the separate documentation but common functions to use are:

- Facets/clusters to standardize information in the columns
- Top column icon -> Edit cells -> Transform. There you can use regular expressions (here indicated as GREL) to edit the text column
- Switching between record mode and row mode (esp. useful for Cornelia project)
- Undo/Redo function where you can go back to any previous stage of the project
- Create a new column based on the value of another (Top column icon->Edit column -> add column based on this column)
- Explore the rest!

Reconciliation with Wikidata: you can use this function to link to Wikidata entries.

- Top column icon -> reconcile -> start reconciling.
- Of course, this has to be done on a clean text, so without special characters indicating the metadata structure: for this, you will need to split the column accordingly

Workflow:

it is important that you coordinate and document the changes performed to the dataset. For this, you can set up a shared folder/directory and be very careful in managing the passages with the datasets. To share a project you need to export and reimport it. You are encouraged to use Git and GitHub for this purpose.

Submission:

The team will submit the documents as a whole. The documents to submit are:

1. Short report (max 4 pages) containing (at least) the following information: 1. brief description of the dataset 2. description of the steps taken and choices operated 3. description of the final dataset (differences with initial) 4. description of the workflow 5. Evaluation of the tool 6. examples of GREL used and of their utility
2. the final version of the dataset, as a .csv file

On the report indicate clearly the members of the team (first name, last name, s number, email). The submission will happen via Toledo. The project is due by 24 November, 23h59. The project counts for 30% of the final grade, and every group will receive a common grade.

Resources for OpenRefine:

- Seth van Hooland, Ruben Verborgh, and Max De Wilde, "Cleaning Data with OpenRefine," Programming Historian 2 (2013), https://doi.org/10.46430/phen0023.
- Clean Data with OpenRefine | Hands-On Data Visualization (handsondataviz.org)
- https://docs.openrefine.org/manual/expressions#regular-expressions
- https://docs.openrefine.org/manual/grelfunctions
- Introduction | Wrangling and Versioning with OpenRefine and GitHub (datacuration.github.io)