OPENREFINE. DATA REFINEMENT AND STANDARDIZATION.

Raw data is often messy, filled with inconsistencies, duplicates, and formatting issues. If left unrefined, these errors can lead to unreliable statistics and wrong conclusions. Data refinement is an essential step.

The best tool we can use for this purpose is OpenRefine. OpenRefine is a powerful open-source tool designed for cleaning, transforming, and organizing messy data. It allows exploration, structurization, and analysis of datasets without directly modifying the original file.
Key features:
- Detect and fix inconsistencies (e.g., different date formats, cases, extra spaces, typos).
- Convert text to a structured format (e.g., split names, standardize capitalization).
- Easily filter, group, and explore large datasets.

To create a new project:
Create project -> choose file -> enter project name -> create

To open an existing project:
Open project -> select a project

1. Mass edit

1.1
Remove Extra Whitespace

Unwanted spaces at the beginning, end, or between words can cause matching, filtering, and clustering errors.

example:
" New York" ≠ "New York"
"New York" ≠ "New York "

When analyzing data with inconsistent values like these, statistics can be inaccurate because the system treats them as different entries, so it's important to clean them first.

<u>To remove white spaces:</u>
Click the dropdown arrow on the column header
Edit cells -> Common transforms -> Trim leading and trailing whitespace
Edit cells -> Common transforms -> Collapse consecutive whitespace

These two actions will remove spaces at the beginning and end of the text and replace multiple spaces between words with a single space.

1.2
Capitalization

Inconsistent capitalization can also cause data inconsistency.
example:
"Vasil Stus" ≠ "vasil stus"

Even though they refer to the same person, the system treats them as different values.

To standardize the case:
Click the dropdown arrow on the column header
Edit cells -> Common transforms -> to titlecase (capitalizes the first letter of each word) / to uppercase /to lowercase

1.3
Setting values to null

Sometimes, data may look empty but still contain hidden characters, like spaces, line breaks, and NaN values, all due to different formatting. These little residues can also affect data filtering, analysis, and cleaning.

To fix this, you can explicitly set such values to null:
Click the dropdown arrow on the column header
Edit cells -> Common transforms -> to null

This ensures that values are, in fact, empty.

1.4
Facet function. Text facet

Facet function lets you group data to easily clean or analyze it. Text facet is better for columns with categorical or repeated text values, as it shows a list of unique text values with counts. It also allows you to know the number of unique values.

Click the dropdown arrow on the column header -> facet -> text facet

Sort by name: Alphabetical order.
Sort by count: From most to least frequent value (default is descending).

To edit a value:
Click on the value you want to change
Click Edit

Enter the new value
Click OK — all rows with that value will be updated at once

1.5
Cluster

The Cluster function in OpenRefine is a great tool used to automatically find and merge similar but not identical values in a dataset.

example:
42 E. 7th.. St., Nyw York З. N. У
42 East 7th Street, New York
42 E. 7 St., Nyw York З. N. У

As we can see, that is in fact one address, but because it's written in 3 different ways, the system treats and counts them as separate values.

<u>To edit such entries, you can use Cluster:</u>
Click the cluster button at the top right
Select a clustering method
Click Cluster

Key Collision Methods are good for catching simple spelling or formatting variations, they are fast and effective for names, places, etc.
Nearest Neighbor Methods use string similarity and are better for more complex variations.

When you use Cluster, OpenRefine automatically suggests a unified value for each group of similar entries. But you can review, change, or accept these suggestions before applying them.

Suggested value (right side):
This is the value OpenRefine thinks should replace all others in the group
Checkbox (left side):
Check this box to approve and apply the change for that cluster
Editable field (right side):
If the suggestion isn't quite right, you can type in a better value

<u>Applying Changes:</u>
! Review all suggested clusters
Edit any suggested value if needed
Check the box for each group you want to merge
Click "Merge selected and close" (or "Merge and re-cluster" if you want to keep refining)

42 E. 7th.. St., Nyw York З. N. У. |
42 East 7th Street, New York | -> 42 East 7th Street, New York
42 E. 7 St., Nyw York З. N. У. |

Test different methods and settings for the best results.

1.6
Fill Down

Sometimes, only the first row in a column is filled, and the rest are blank, even though they should all have the same value. Instead of typing it repeatedly, you can use the Fill Down feature.

example:

before	after
Material Type	Material Type
Book	Book
(blank)	Book
(blank)	Book

To fill down the rows:
Click the dropdown arrow on the column header
Edit cells -> Fill down

1.7
Replace
Dropdown -> Edit cells -> Replace

2. Creating new columns and rows from existing values

2.1
Multivalued Cell

A multivalued cell contains more than one entity in a single cell, usually separated by a delimiter like a comma, semicolon, pipe, etc.

example:

author	author
Valys Stus; Bohdan-Ihor Antonych; Ihor Kalynets	Valys Stus
	Bohdan-Ihor Antonych
	Ihor Kalynets

To clean, cluster, or analyze such entries properly, it's often helpful to split them into separate values.

How to Split Multivalued Cells:
Click the dropdown on the relevant column
Edit cells -> Split multi-valued cells -> Choose the appropriate separator (e.g. ;)

OpenRefine will convert the cell into multiple rows, one per value, while keeping the rest of the row data duplicated for each. This lets you normalize values individually and apply clustering or transformations per entry

! After standardizing individual entries, you can merge them back into a single cell.
Click the dropdown on the relevant column
Edit cells -> Join multi-valued cells -> Choose the appropriate separator (e.g. ;)
This is useful when you've cleaned or clustered values individually and want to restore the original format.

2.2
Multivalued Column

Sometimes a column contains more than one distinct type of data, not just multiple values of the same kind, but actually values that belong in different columns. We can consider this type of column multivalued.

example:

author
Василь Стус | Vasyl Stus
Игорь Калинец | Ihor Kalynets
Богдан-Игорь Антонич | Bohdan-Ihor Antonych

author_original	author_english
Василь Стус	Valys Stus
Игорь Калинец	Ihor Kalynets
Богдан-Игорь Антонич	Bohdan-Ihor Antonych

In this case, the column includes both the Ukrainian name and its transliteration, separated by a | symbol. We want those values to be set in different columns.

To split a multivalued column:
Click the dropdown on the mixed column
Edit column -> Split into several columns -> Enter the delimiter used in the data -> OpenRefine will create two new columns, one for each part -> Give a new column a proper name

After this step, you can edit, cluster, and analyze each column and its values separately.

! Just as with multivalued cells, you can also join the columns back if needed.
Click the dropdown on the relevant column.

Edit cells -> Join columns -> Select columns from the left field -> Choose the appropriate separator (e.g. ;) and enter it in the first field on the right

You can use this method to join any columns (not just ones you previously split) as long as combining their content makes sense for your use case.

2.3
Creating a column based on another column

We might want to create a column based on an existing one to keep the original values and add a column with modified ones.

example:
All publishing houses were added as they were mentioned in the book. We'd like to have them standardized for further analysis but want to keep the original record.

To add a column based on another one:
Click the dropdown on the mixed column
Edit column -> Add column based on this column -> Enter new column name in the field at the top -> Press Ok

3. Undo/Redo

One of the most important features in OpenRefine is the ability to undo and redo any step. When working with large or messy data, it's easy to make mistakes.

To Undo:
Click the Undo/Redo tab in the upper left panel
You'll see a list of all actions taken on your project
Click on any earlier step to revert to that state

It's like a visual Ctrl+Z for your whole project, but even better, because you can jump to any point in your workflow.

If you've undone something and want to bring it back:
Click the Undo/Redo tab in the upper left panel
Click forward in the list to redo the steps you previously reverted

This makes experimenting much safer and encourages you to try different cleaning methods without fear of losing work.

4. Data Export

4.1
Entire Dataset Export

Once you've cleaned your data, OpenRefine lets you export it in various formats.

Common Export Formats:
CSV – Comma-separated values
TSV – Tab-separated values
XLSX – Excel format

To export your data:
Click the Export button (top-right corner)
Choose your preferred format
The download will start automatically

4.2
Simplified Dataset Export

You also have export options if you don't want to export the entire dataset.

example:
You want to review publishing houses and the authors they published to examine their relationships and analyze how authors were distributed across different publishers. You don't need every column, just the ones related to publishers and authors.

To export related fields only:
Click the Export button in the top-right corner
Select "Custom tabular export" from the dropdown menu
In the export window:
- Select the columns you want to include
- Use filters or facets beforehand to limit rows (optional)

Click the Download tab.
- Choose your desired format (CSV, TSV, Excel, etc.)
- Click Preview to check a sample of your output (optional)
- Press Download to save the file.

This step allows you to create a simplified version of your dataset that includes only selected fields.

These features are just a glimpse of what OpenRefine can offer, but they will already help make your data cleaning work faster and more efficient.