

Cleaning and enriching data with OpenRefine

ETHAN YOO, DATA SCIENCE GRADUATE SPECIALIST

FEBRUARY 7, 2023

ETHAN.YOO@RUTGERS.EDU

Outline

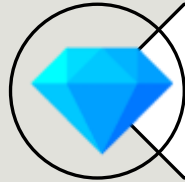
- OpenRefine and its core functions
- Cleaning data with OpenRefine
- Enriching data with OpenRefine
- Demonstration/walkthrough
- Additional resources

OpenRefine

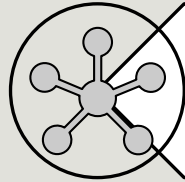
- “a Java-based power tool that allows you to load data, understand it, clean it up, reconcile it, and augment it with data coming from the web” ([OpenRefine](#))
- Free and open source software



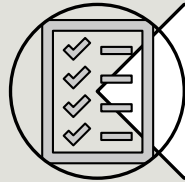
OpenRefine's Functionality



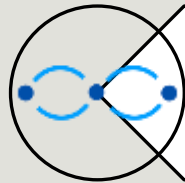
Faceting



Clustering and editing



Reconciling



History

Cleaning data



Enriching data (data augmentation)



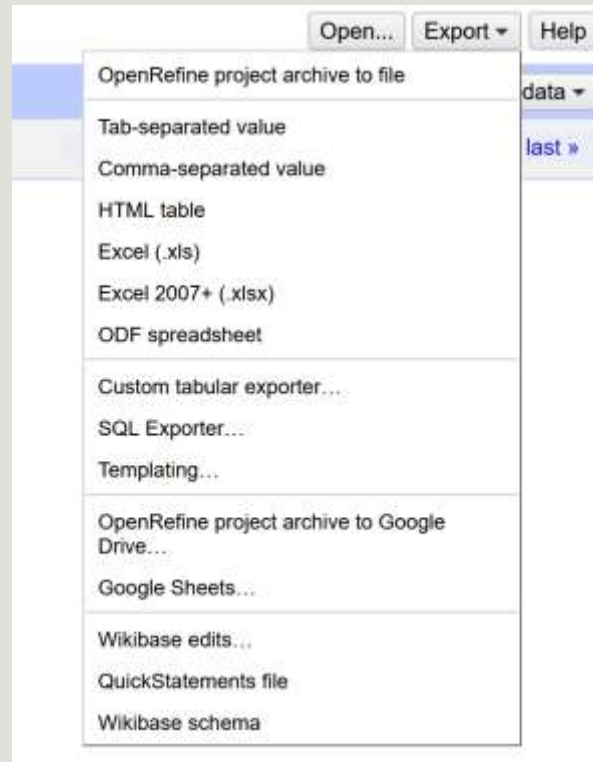
The screenshot shows a web-based data enrichment interface. On the left, there is a sidebar with a search bar and several expandable sections. The main area displays a table with multiple columns, including what appears to be a primary key, a name, a date, and several numerical or categorical fields. The table is populated with data rows, and the interface includes standard web controls like pagination and filtering options.

ID	Name	Date	Value 1	Value 2	Value 3
1	John Doe	2023-01-01	100	200	300
2	Jane Smith	2023-01-02	150	250	350
3	Bob Johnson	2023-01-03	200	300	400
4	Alice Brown	2023-01-04	250	350	450
5	Charlie Davis	2023-01-05	300	400	500
6	Diana Prince	2023-01-06	350	450	550
7	Frank Miller	2023-01-07	400	500	600
8	Grace Wilson	2023-01-08	450	550	650
9	Henry Taylor	2023-01-09	500	600	700
10	Ivy Clark	2023-01-10	550	650	750

Running OpenRefine for the first time

- Requires a browser but an Internet connection is not required for most actions
 - Open any web browser (e.g., Mozilla Firefox or Google Chrome)
 - Navigate to <http://127.0.0.1:3333> (or <http://localhost:3333>)
 - Create a project by importing data
 - **Supported:** TSV, CSV, and other delimited files, Excel files, JSON, XML
 - **Sources:** Local computer, web address, clipboard, relational database (SQLite, PostgreSQL, MySQL/MariaDB), Google Sheets (public spreadsheet or authenticated account)

Exporting or backing up work



Select “Export” in the upper-right corner and choose an option

Additional Resources

- **Data Carpentry**
 - [OpenRefine for Social Science Data](#)
 - [Data Cleaning with Open Refine for Ecologists](#)
- [**OpenRefine user manual**](#)

Microsoft Excel

- **Data**
 - Get & Transform Data
 - Get Data From Other Sources



Microsoft Excel

- **Data**
 - Get & Transform Data
 - Get Data From Other Sources

