

Cleaning and enriching data with OpenRefine

ETHAN YOO, DATA SCIENCE GRADUATE SPECIALIST

FEBRUARY 7, 2023

ETHAN.YOO@RUTGERS.EDU

Outline

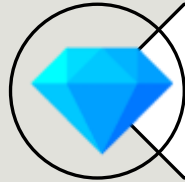
- OpenRefine and its core functions
- Cleaning data with OpenRefine
- Enriching data with OpenRefine
- Demonstration/walkthrough
- Additional resources

OpenRefine

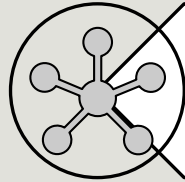
- “a Java-based power tool that allows you to load data, understand it, clean it up, reconcile it, and augment it with data coming from the web” ([OpenRefine](#))
- Free and open source software



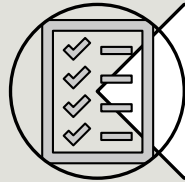
OpenRefine's Functionality



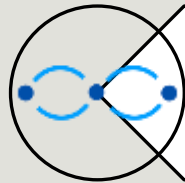
Faceting



Clustering and editing



Reconciling



History

Cleaning data



Enriching data (data augmentation)



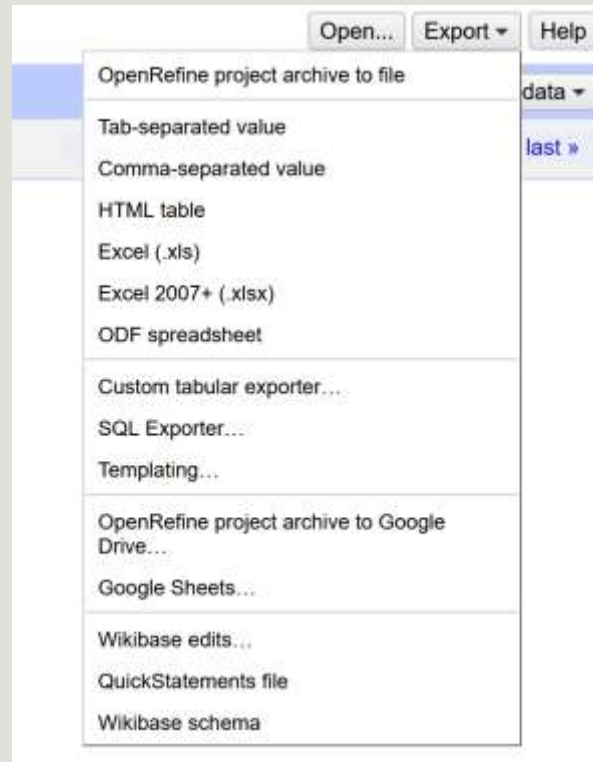
The image shows a screenshot of a data enrichment tool interface. On the left, there is a sidebar with a search bar and several expandable sections. The main area displays a table with multiple columns, including what appears to be a primary key, a text field, and several numeric or date fields. The table contains several rows of data, some of which are highlighted in yellow. The interface has a clean, modern design with a light blue header and a white background for the table.

id	name	age	gender	email	phone	address
1	John Doe	30	Male	john.doe@example.com	1234567890	123 Main St, New York, NY 10001
2	Jane Smith	25	Female	jane.smith@example.com	9876543210	456 Elm St, Los Angeles, CA 90001
3	Bob Johnson	40	Male	bob.johnson@example.com	5555555555	789 Oak St, Chicago, IL 60601
4	Alice Brown	28	Female	alice.brown@example.com	1111111111	101 Pine St, San Francisco, CA 94101
5	Charlie Davis	35	Male	charlie.davis@example.com	2222222222	202 Cedar St, Houston, TX 77001

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

