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

January 2010

Released as Freebase Gridworks by Metaweb Technologies

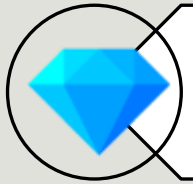
July 2010

Renamed to Google Refine after Metaweb was acquired by Google

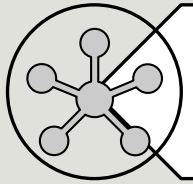
October 2012

Transitioned to a community-driven project and renamed again to OpenRefine

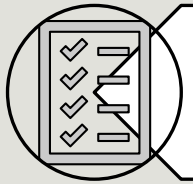
OpenRefine’s Functionality



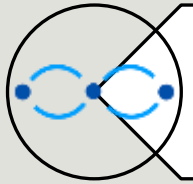
Faceting



Clustering and editing



Reconciling



History

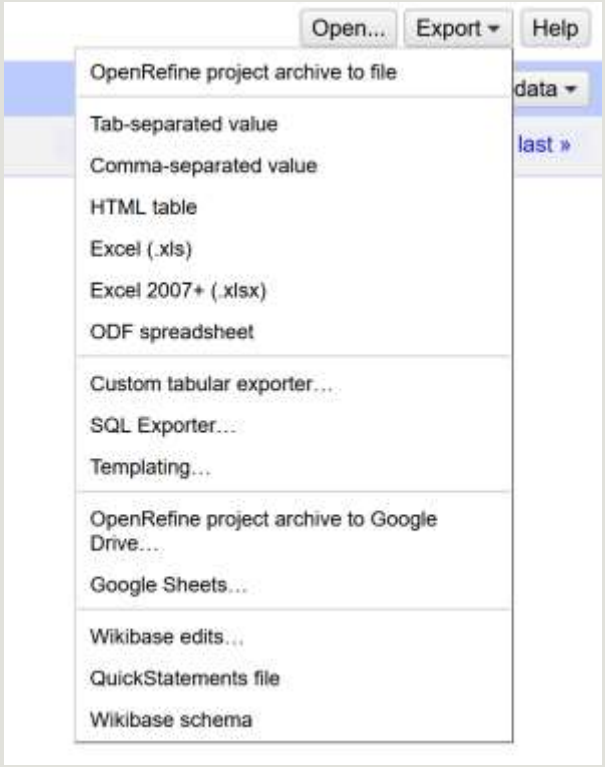
Cleaning data

Enriching data (data augmentation)

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