Class Outline: QA/QC 2 - Open Refine

Objectives and Competencies:

By the end of this lesson students will:

- Be able to import a data set into OpenRefine, make changes to the data set and its structure, and export the revised data set
- Learn how to automatically track changes made and export the record of changes
- Be able to apply these changes to a different data set

Pre-Class Preparation (Instructor):

- Remind via email about OpenRefine Installation
- Post Data sets

Bring to Class:

- Snacks
- Tent cards for student names

Pre-class Preparation (Students):

Online Lectures: None

Readings: None

Computer Resources

1. Install OpenRefine on your computer and verify it works by following the instructions here.
2. *Optional:* Read and watch about how OpenRefine works here. You can also review the basic workflow we will learn.

In-Class: Using Open Refine to clean data

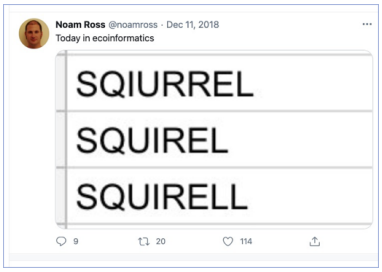

Figure 1: How many ways can you spell....

OpenRefine is a powerful, free, and open source tool that is used to work with and clean messy data. We will be working through some of OpenRefine's basic features, after which you will trying them onm your own on a new data set.

Intro to OR (10 min)

Working with OR (10 min)

Filtering and Sorting (45 min)

Break (10 min)

Examining Numbers (30 min)

Using Scripts, Exporting, and Saving (45 min)

Wrap-up, Questions (10 min)

Assignment (20 min)

Now it's your turn. Download this csv file and use OpenRefine to clean it up. After you create a Project, edit the data as follows:

1. Correct and standardize the names of the countries in which the rodents were captured.
2. The column `scientificName` contains two pieces of information (the Genus *and* species of each animal). Split this into two columns, rename them as `genus` and `species`, and then correct and standardize the data in each column as needed. NB: You may run into an obstacle when you try to rename the columns. How can you get around it?
3. Save the clean data as a CSV file on your desktop.
4. Extract and save your steps (i.e., 'operation history' as JSON. Save this as a text file.
5. *Bonus Brainteaser:* Many of the cells in the column for the Latin bonomial are blank. How might you go about filling them in based on the column with the abbreviation?
6. **Submission:** Submit your clean .csv and the JSON text file as `week6_hw` on Canvas.

Grading Rubric:

Data corrected and JSON file can be used on another data set: 50
Most data correction properly programmed; some require instructor follow-up: 40
Many of the corrections missing, JSON file unable to process new data : 30
Instructor follow-up required to implement most changes: 20

Sources for this lesson

1. Data Carpentry: Data Cleaning with OpenRefine for Social Scientists.

2. Data Carpentry: Data Cleaning with OpenRefine for Ecologists

Additional Tools and Resources

OpenRefine Home

- Open Refine Homepage. Includes the user's manual and links to more tutorials.

UF Library Workshops

- The UF Library teaches a number of excellent workshops, including one on using OpenRefine taught by Dr. Hao Ye. You can see the schedule (or request one) here.

Tutorials

- Environmental Data Initiative OpenRefine Tutorial
- Cleaning Data with OpenRefine Video Tutorials:
  - Video Tutorial No. 1
  - Video Tutorial No. 2
- JHU Library: Cleaning Data with OpenRefine
- The Programming Historian Website: Cleaning Data with OpenRefine.

GREL Cheatsheets

- Belinda Weaver's GREL Cheatsheet with examples
- OpenRefine GREL Manual
- A really good GREL Guide from the Univ Illinois
- Even better: code4lib Toronoto's OpenRefine cheatsheets, including for GREL commands.
- Datenschule's OR Cheatsheets

R Tools

- The rrefine package allows you to do some OpenRefine tasks from within R.