SOCIAL SCIENCES RESEARCH CENTER

THE UNIVERSITY OF CHICAGO DIVISION OF THE SOCIAL SCIENCES

Social Science Computational Resources Series

February 4, 2020

Brooke Luetgert, PhD

THE UNIVERSITY OF CHICAGO | Office of Research and National Laboratories Research Computing Center

About this Series

This five part series is designed to expose participants to working with data in various software environments. While many tools will "get the job done", knowing the tips and tricks of multiple alternatives helps the user to work more efficiently and confidently through data analysis and presentation. Our focus is on good data practice, documentation and reproducibility across software platforms.

This is an introductory series for absolute beginners!

Series Schedule

We will meet every other week in 1155 60th St. Room 344 from 2-4 PM starting on February 4th

1. Feb 4- Find, Clean and Assemble with OpenRefine
2. Feb 18- Deep Dive into Stata
3. Mar 3- The R alternative
4. Mar 17- Python for Data Science
5. Mar 31- SSRC/RCC Mini-Hackathon

Contact Details

Brooke Luetgert, Computational Scientist at RCC

Email: luetgert@uchicago.edu
Office: TAAC 2, 5607 South Drexel
Telephone: (773)-834-5313

RCC Help Desk: Reg. 216, Mon-Fri 9AM-5PM

Materials on GitHub- use search bar, enter user:luetgert
Our folder is **luetgert/SSD_RCC_WorkshopSeries**

Plan for Today

Our focus will be two-fold:

1)We want to download and discover OpenRefine as a tool to sort, summarize and clean our data.

2)We also want to discuss online data resources for the social sciences and practice downloading sample data, uploading into OpenRefine for exploration and exporting our changes.

Our goal is to prepare data for further analysis in Stata, R or Python in the upcoming workshop sessions.

System Setup

Make sure that you have Firefox or Chrome browsers installed and set as your default browser. **Internet Explorer will not run OpenRefine properly.**

Download data- our sample data come from the Data Carpentry Social Sciences workshop on Studying African Farmer-Led Irrigation (SAFI) database. These are data from farmers interviewed in Mozambique and Tanzania in late 2016. They were asked about their household construction, number of members in the household, water usage, most valuable items owned and livestock.

https://ndownloader.figshare.com/files/11502815

Download OpenRefine

Go to **openrefine.org**

In the far left column, choose **download**. Now select your operating system from the list under Version 3.3. This will download the installer to your computer. Go ahead and click on the diamond icon to complete installation. (You may need to control-click the icon and then select open to override the anti-virus protection.)

This is open source software. It was previously known as Google Refine.

Data Types

Most common data file formats (reflected in file endings):

TSV tab separated values

CSV comma separated values

.xls and **.xlsx** Excel files

JSON JavaScript Object Notation

XML Extensible Markup Language

*All can be imported into OpenRefine. CSV is most prominent for distribution, universally importable.

Data Sources

American FactFinder- census, housing, geographic US

World Bank- development, macro-economic global

OECD- trade, environment, agriculture, econ.

UNdata- general health indicators, education, development

Eurostat- EU statistics economy, trade, demographics

Data-Planet- statistics from US Federal Agencies

Data.gov- machine readable data from US Federal Govern.

General Social Survey- opinion data from US

World Values Survey- pooled data from 65+ countries

Roper Center- US public opinion, approval rates

Data Sources (cont.)

Integrated Public Use Microdata Series (IPUMS)

Inter-University Consortium for Political and Social Science Research (ICPSR)

iPoll Databank- most Roper data plus Pew, Gallup and more

IQSS Dataverse- Harvard repository of raw social science

Enerdata- global energy stats, CO2 emissions, consumption

European Bank for Reconstruction and Development (EBRD)

Kaggle- devoted to machine learning- 19,000 data sets and 200,000 public notebooks with shared code

*Disclaimer- this is not a definitive list, but a starting point!

Now on to OpenRefine

OpenRefine is **not** a perfect product. It is open source, universally downloadable and very easy to work with. Use the Chrome browser for best results.

- It will autosave your work. Move slowly. If it crashes, all of your steps have been saved.
- It will not change your source data.
- It will document all changes made to the data.
- A ‘facet’ groups all the like values that appear in a column
- ‘Clustering’ will help detect typing errors
- You can choose a different export type than import type.