Introduction to OpenRefine

Ariel Deardorff, Data Services Librian, UCSF Library

2020-11-03

Contents

1 Welcome **5**
- 1.1 Pre-Class Prep . . . 5
- 1.2 Learning Objectives . . . 5
- 1.3 Class Data . . . 5
- 1.4 About the Instructor . . . 5

2 Getting to Know our Data and Tools **7**
- 2.1 What is OpenRefine? . . . 7
- 2.2 Exploring the Data . . . 7
- 2.3 Uploading Data Into OpenRefine . . . 7
- 2.4 The OpenRefine Interface . . . 7

3 Basic Data Cleaning with OpenRefine **9**
- 3.1 Splitting Columns . . . 9
- 3.2 Undo/Redo . . . 9
- 3.3 Faceting Data . . . 9
- 3.4 Trimming Whitespace and Changing to Titlecase . . . 9
- 3.5 Clustering Data . . . 9
- 3.6 Sorting . . . 9

4 Advanced OpenRefine **11**
- 4.1 Pulling data from an API . . . 11
- 4.2 Extracting the Data . . . 11

5 Exporting Cleaned Data and Scripts **13**

6 Recommended Resources and Tutorials **15**

Chapter 1

Welcome

Placeholder

1.1 Pre-Class Prep

1.2 Learning Objectives

1.3 Class Data

1.4 About the Instructor

Chapter 2

Getting to Know our Data and Tools

Placeholder

2.1 What is OpenRefine?

2.2 Exploring the Data

2.3 Uploading Data Into OpenRefine

2.4 The OpenRefine Interface

Chapter 3

Basic Data Cleaning with OpenRefine

Placeholder

3.1 Splitting Columns

3.2 Undo/Redo

3.3 Faceting Data

3.4 Trimming Whitespace and Changing to Titlecase

3.5 Clustering Data

3.6 Sorting

Chapter 4

Advanced OpenRefine

Placeholder

4.1 Pulling data from an API

4.2 Extracting the Data

Chapter 5

Exporting Cleaned Data and Scripts

The OpenRefine project has been saving as you go - but not altering your underlying data (this is a good thing!)

- To export the cleaned version of our data we need to click `Export` (top left corner) then choose the type of file (CSV or Excel)
  - We recommend that you rename your new file with a good title like "dataxyz_cleaned" so you know that it is the cleaned version.
- What if we are going to get more of this same kind of data over and over and want to apply the same changes? We could export the script and then re-run it on new data.
  - In the `Undo/Redo` section click `Extract` - select the steps you want to save
  - Then copy and paste into a text editor (like Text Wrangler/Text Edit) and save as a txt file (probably in the same folder as your data or with other scripts you use)
- Open a new project with your new data - go to `Undo/redo` > `Apply` enter the plain text from your file > `Perform Operations` and blammo it will re-run all the same data cleaning steps on your new dataset.

Chapter 6

Recommended Resources and Tutorials

Here are some places to learn more about OpenRefine:

- OpenRefine Website
- Library Carpentry - OpenRefine (heavily used for the basis of this lesson)
- Data Carpentry – OpenRefine for Social Science Data
- Grateful Dead OpenRefine Tutorial
- Programming Historian - Fetching and Parsing Data with OpenRefine
- Fetching URLS from OpenRefine
- Reconciling Data with OpenRefine