SCIENCE MEETS LIFE

Introduction to Data Cleaning with OpenRefine

VIB BIOINFORMATICS CORE

8 May 2019

Alexander Botzki

Our areas of bioinformatics training comprise various topics of life sciences

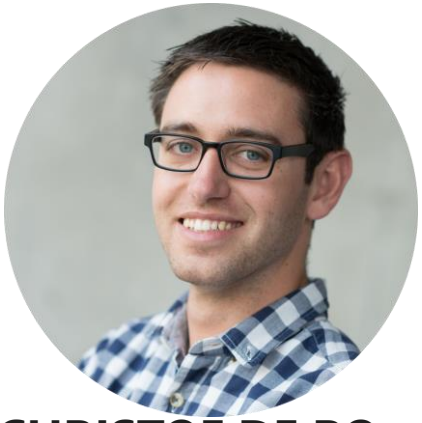
VIB BIOINFORMATICS CORE

SCIENCE MEETS LIFE

Our trainers for VIB and ELIXIR Belgium

JANICK MATHYS

TUUR MUYLDERMANS



CHRISTOF DE BO

VIB BIOINFORMATICS CORE

elixir BELGIUM

SCIENCE MEETS LIFE

What is OpenRefine?

A free, open source, power tool for working with messy data

Automatically tracks steps in working with data

http://openrefine.org

Large community / google group

How does OpenRefine compare with other tools?

- Compared to spreadsheets
  Basic unit of interaction is column (versus cells)
  Pro: easier to import data, explore, manipulate and export again

- Compared to scripting
  Pro: you see your data, while it is being transformed
  Con: for medium size data sets (1,000,000 rows)

- Compared to databases
  Pro: you see your data, while it is being queried
  Con: for simple data structures

What are typical use cases for OpenRefine?

- Explore unknown, new data files
- Manipulate/clean data to prepare for other tools (GREL)
- split in more granular parts
- match local data with other datasets
- get data from web services to enhance the dataset
- use as workflow tools to replay

We start OpenRefine and explore the interface together

- Create a project from the sample data
  https://ndownloader.figshare.com/files/7823341

- Exploring data by applying filters
  Faceting

  Explore the scientificName column

- What type of errors do we spot?

VIB BIOINFORMATICS CORE

SCIENCE MEETS LIFE

We learn more about facets and their use

- Types
  Numeric facets
  Timeline facets
  Custom facets
  Scatterplot facets
- Custom facets
  e.g. by blank, text length, word, duplicates

- Group common values, filter, and bulk edit them

Exercise 1

- Using faceting, find out how many years are represented in the census.
- Is the column formatted as Number, Date, or Text? How does changing the format change the faceting display?
- Which years have the most and least observations?

We use clustering to bulk edit values

Gödel and Godel or New York and new york

- Explore the scientificName column
- Cluster: key collision method and metaphone3

VIB BIOINFORMATICS CORE

SCIENCE MEETS LIFE

Exercise 2

- Split the column scientific name into genus and species
- Explore the Undo / Redo operations
- Which transform solves the issue?

How can we use a subset of the data

Text filter > bai

Exercise 3

- What scientific names (genus and species) are selected by searching for 'bai'?
- How would you restrict this to one of the species selected?
- How many rows are matched for each of the species?
- Use include / exclude to select entries from one of the species.
- Add custom text facet to arrive at the same selection like the text filter 'bai'
  e.g. value.toLowercase().contains("bai")

Sorting

Text, number, dates or Booleans

Exercise 4

- Continue to work with both species.
- Sort by month. How can you ensure that months are in order?
- Sort the data by plot. What years were observations recorded for the filtered set?
- How do you sort your data by month?
- How do you sort the subset in chronological order.

We examine numbers in OpenRefine

Edit cells > Common transforms on recordID

Exercise 5

- Use the full set again.
- Transform the columns recordID, dy, mo, period, plot_id to numbers.
- Edit several cells of the column 'dy' to contain text or empty cells. Explore the numeric facet on that column.
- Create a scatterplots by plotting pairs of numeric columns.
- Why does the scatterplot recordID vs period have the pattern is does?

Exercise 6

- Create a project with the bed file: syst-nocallsCG69.bed https://bit.ly/2Q9gA05

Try to determine

- a) the number of no-call regions that are larger than 1040 bases long in chromosome 21
- b) the length of the longest region in chromosome 1

Here you can find additional resources about OpenRefine

- https://datacarpentry.org/OpenRefine-ecology-lesson/06-resources/index.html

Questions?

Comments?