Cleaning Data with OpenRefine

Introduction to Data Management Practices course

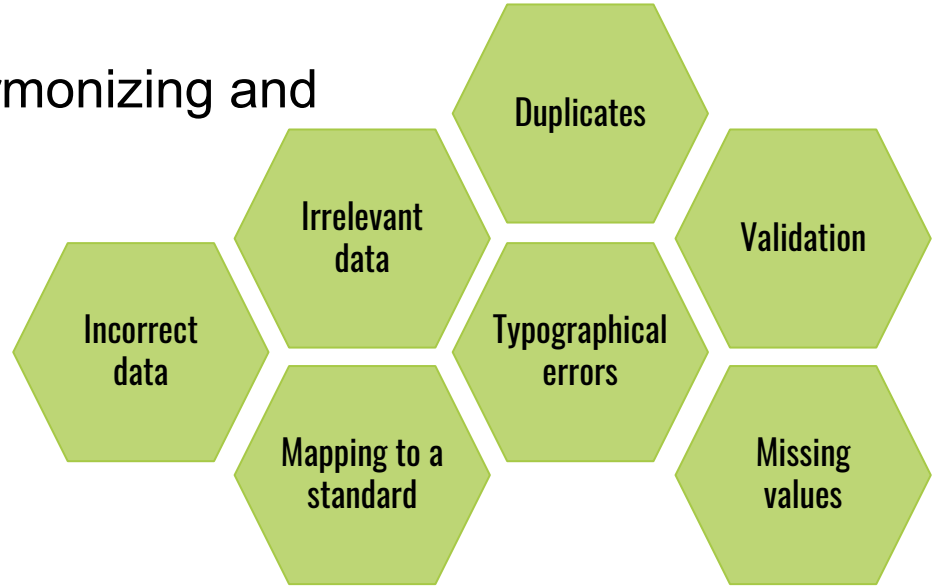
NBIS DM Team

data@nbis.se

https://nbisweden.github.io/module-openrefine-dm-practices/index.html

Data Cleaning

The process of harmonizing and standardizing data



Typical workflow

Inspecting	Cleansing	Verifying	Reporting/Documenting
Detect unexpected, incorrect, and inconsistent data.	Fix or remove the anomalies discovered.	After cleaning, the results are inspected to verify correctness.	A report about the changes made and the quality of the currently stored data is recorded.

OpenRefine

OpenRefine

A powerful open source tool that can be used for data cleaning

- Free
- Does not change your original data file
- Keeps your data private on your own computer until you choose to share it
- Automatically tracks any step you take allowing you to easily document and reuse the cleaning process
- Works with fairly large datasets

1. Working with OpenRefine

- *How can we bring our data into OpenRefine?*
- *How can we sort and summarize our data?*
- *How can we find and correct errors in our raw data?*

Sorting and summarizing our data using **Facets**:

- Groups all the like values that appear in a column
- Allow you to filter the data by these values and edit in bulk

Facets are a useful way to explore your data and seeing the overview picture

https://nbisweden.github.io/module-openrefine-dm-practices/02-working-with-openrefine/index.html

1. Working with OpenRefine

Finding and correcting errors using **Clustering**:

- Identifying and grouping different values that are alternative representations of the same thing.
  - “New York” and “new york” - same concept different capitalization
  - “Gödel” and “Godel” probably refer to the same person

- Allow you to filter the data by these values and edit in bulk

Clustering is very powerful for cleaning up misspelled or mistyped entries or when applying a standard retrospectively.

https://nbisweden.github.io/module-openrefine-dm-practices/02-working-with-openrefine/index.html

2. Filtering and Sorting

- *How can we select only a subset of our data to work with?*
- *How can we sort our data?*

When a dataset has many entries, **filtering** can be used to create a subset of the data that is relevant for the specific task at hand.

Data **sorting** arranges the data into some meaningful order to make it easier to understand, analyze or visualize.

https://nbisweden.github.io/module-openrefine-dm-practices/03-filter-sort/index.html

3. Numbers in OpenRefine

- *How can we convert a column from one data type to another?*
- *How can we visualize relationships among columns?*

Each value in a cell in OpenRefine is assigned one of the following **data types**:

- string/text - ***default upon import***
- number
- date (YYYY-MM-DDTHH:MM:SSZ)
- boolean (“true” or “false”)

Note: text values can be sorted as numbers without changing the data type

https://nbisweden.github.io/module-openrefine-dm-practices/04-numbers/index.html

Key points so far..

- OpenRefine can import a variety of file types.
- OpenRefine can be used to explore data using facets.
- Clustering in OpenRefine can help to identify different values that might mean the same thing.
- OpenRefine can transform the structures and values of a column.
- OpenRefine provides a way to sort and filter data without affecting the raw data.
- OpenRefine provides ways to get overviews of numerical data.

Reusability and Reproducibility

- OpenRefine tracks and documents all the modifications done to the data
- OpenRefine allows you to export the documentation in order to apply the same modifications to another dataset with the same structure

Why is this important?

- It makes your own work more efficient

- It provides documentation for yourself and others to understand how the data has been modified

- It provides everything necessary to reproduce your cleaned data

Using Scripts and exporting data

- *How can we document the data-cleaning steps we've applied to our data?*
- *How can we apply these steps to additional data sets?*
- *How can we save and export our cleaned data from OpenRefine?*

A script is a recipe with stepwise instructions for machines.

OpenRefine uses the data format JSON to generate scripts.

https://nbisweden.github.io/module-openrefine-dm-practices/05-scripts/index.html
https://nbisweden.github.io/module-openrefine-dm-practices/06-saving/index.html

Working with a subset of data

Scenario:

- **Sam** is going to submit **sequencing data** to the repository ENA and the sample metadata is stored in the common spreadsheet we have been working with.

- Sam needs to transform and extract a subset of the data in the common spreadsheet to prepare a sample metadata file compatible with the ENA and need to consider the following questions

  - Which of the existing columns are relevant for the submission?
  - Are they named correctly?
  - Are there additional columns that need to be added?

ENA sample metadata

Mandatory metadata for all ENA samples:

Basic details:

- **sample_alias** - *The unique name is a submitter provided unique identifier.*
- **sample_title** - *The sample title is a short, preferably a single sentence, description of the sample.*

Organism details:

- **tax_id** - *The NCBI taxonomy id*
- **scientific_name** - *based on tax_id*

Question: Can any of the existing columns be used to provide the mandatory metadata?

animal ID	researcher	experiment refer	sample	genotype	tax_id	date	mouse line	strain	age	developmental s	sex	organism part	experiment type
834217	Kim	up_235_1	A	Kdr Y949F/Y949F	10900	2020-02-18	Alk3	BALB/cJ	4	adult	male	lung	sequencing assay
836507	Sam	up_201_4	D_hom	Kdr Y949F/Y949F	10900	2020-02-23	Kdr	C57BL/6	9	adult	male	lung	sequencing assay
842068	Sam	Feb2720_IHC	C	KdrY949F/Y949F	10900	2020-02-27	Kdr	C57BL/6	P9	pup	female	lung	IHC
843132	Sam	Mar0418_IHC	D	KdrY949F/Y949F	10900	2018-03-04	Kdr	C57BL/6	P9	pup	female	lung	IHC
845290	Kim	up_235_2	B	Kdr Y949F/Y949F	10900	2019-03-07	Alk3	BALB/cJ	8	adult	male	lung	sequencing assay

ENA sample metadata

Current variable name	ENA Variable name	Measurement unit	Allowed values
animal ID			
date			format: YYYY-MM-DD, >=proj_start_date & <=today
mouse line	sub_strain		
strain	strain		NCIT ontology: C56BL/6 Mouse (NCIT:C14424), BALB/cJ Mouse (NCIT:C14657)
age		days,weeks (?)	
developmental stage	dev_stage		BTO ontology: pup (BTO:0004377), adult (BTO:0001043), embryo (BTO:0000379)
sex	sex		male, female, unknown
organism part	tissue_type		MA ontology: lung (MA:0000415), brain (MA:0000168)
genotype			
experiment type			
experiment reference			
researcher			

Checklist-derived metadata:

- strain
- sub_strain
- dev_stage
- sex
- tissue_type

To specify the ontology terms we will add **custom fields:**

- strain_ID
- dev_stage_ID
- tissue_type_ID

ENA exercise

1. Create a new project in OpenRefine named **ENA sample metadata** by loading the same data as before (samples_openrefine_lesson.csv)
2. Open the file ***ENA_sample_metadata_script.txt*** found in the project folder. Copy the JSON script and apply it to the project.
3. Export the cleaned data as a tab separated file (.tsv)
4. Open the file in a text editor and add the following two lines at the beginning of the file:
   #checklist_accession ERC000011
   #unique_name_prefix

   NB! Make sure that you have a tab between #checklist_accession and ERC000011
5. Save the file in your course folder and use in the next lesson.

https://nbisweden.github.io/module-openrefine-dm-practices/exercises_april2021/index.html

Resulting .tsv file

ENA-sample-metadata.tsv

```
#checklist_accession    ERC000011
#unique_name_prefix
sample_alias    tax_id  scientific_name sample_title    dev_stage       tissue_type     sex
        sub_strain      strain  strain_ID       dev_stage_ID    tissue_type_ID
up_201_4        10900   Mus musculus    D_hom Lung tissue from adult Kdr(Y949F/Y949F) mouse.
        adult   lung    male    Kdr     C57BL/6 NCIT:C14424     BTO:0001043     MA:0000415
up_201_6        10900   Mus musculus    F_hom Lung tissue from adult Kdr(Y949F/Y949F) mouse.
        adult   lung    male    Kdr     C57BL/6 NCIT:C14424     BTO:0001043     MA:0000415
up_201_5        10900   Mus musculus    E_hom Lung tissue from adult Kdr(Y949F/Y949F) mouse.
        adult   lung    female  Kdr     C57BL/6 NCIT:C14424     BTO:0001043     MA:0000415
up_201_2        10900   Mus musculus    B_wt Lung tissue from adult wildtype mouse.     adult
        lung    Male    Kdr     C57BL/6 NCIT:C14424     BTO:0001043     MA:0000415
up_201_1        10900   Mus musculus    A_wt Lung tissue from adult wildtype mouse.     adult
        lung    female  Kdr     C57BL/6 NCIT:C14424     BTO:0001043     MA:0000415
up_201_3        10900   Mus musculus    C_wt Lung tissue from adult wildtype mouse.     adult
        lung    Male    Kdr     C57BL/6 NCIT:C14424     BTO:0001043     MA:0000415
```

Other resources

- *What other resources are available for working with OpenRefine?*

OpenRefine has its own web site with documentation and a book:

- OpenRefine web site
- OpenRefine Documentation for Users
- Using OpenRefine book by Ruben Verborgh, Max De Wilde and Aniket Sawant
- OpenRefine history from Wikipedia