Cleaning Data with OpenRefine

Introduction to Data Management Practices course

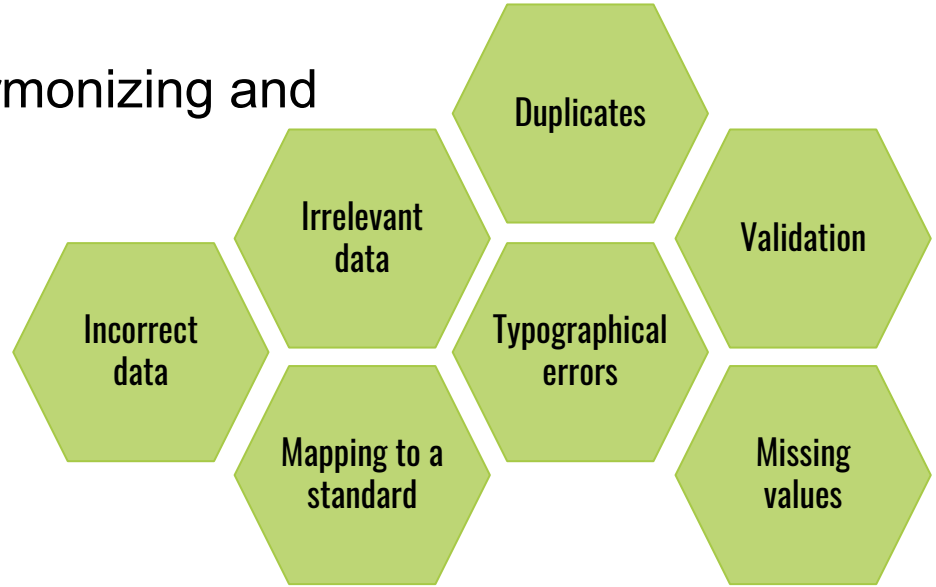
NBIS DM Team

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

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