# Cleaning Data with OpenRefine

*Introduction to Data Management Practices course*

NBIS DM Team
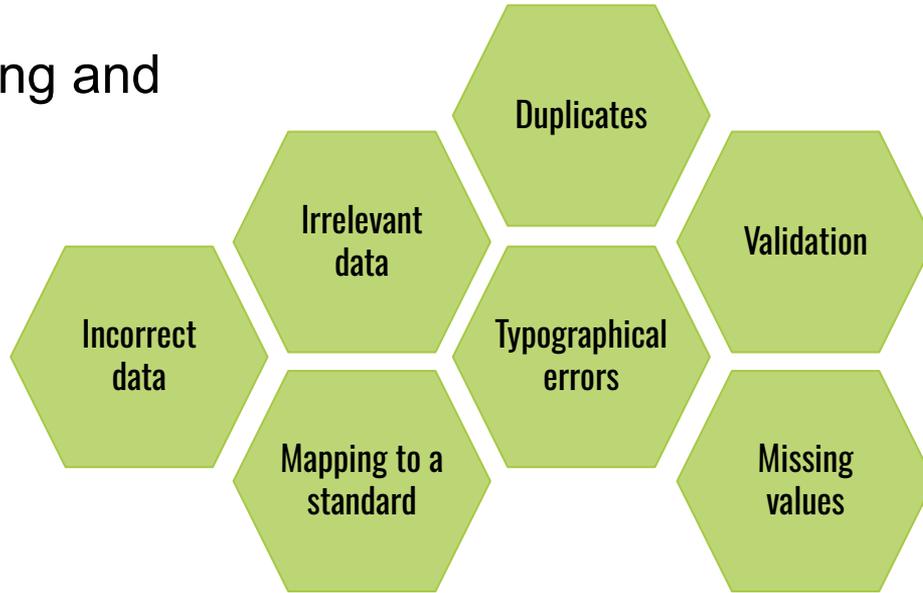
data-management@nbis.se

# Data Cleaning

The process of harmonizing and standardizing data

Duplicates

Irrelevant data

Validation

Incorrect data

Typographical errors

Mapping to a standard

Missing values

Typical workflow

| Inspecting | Cleaning | 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



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