Data Cleaning in OpenRefine

Marley Kalt, Data Management Consultant

September 27, 2022

JOHNS HOPKINS LIBRARIES

Data Services

JHU DATA SERVICES

Before we start, a bit about ZOOM

- Mute audio and video
- Ask questions!
  - Use the public/private chat
- This webinar will not be recorded
- You will receive today's materials by email and on GitHub

Today's software

Download here: https://openrefine.org/download.html

Recommended web browsers: Chrome, Edge, Safari

Today's materials

Materials available on GitHub:

https://github.com/jhu-data-services/data-cleaning-openrefine

Repository contains:

- These slides
- Workshop data
- Step-by-step workshop guide
- Resources

JHU DATA SERVICES

HELPING YOU NAVIGATE DATA

WE HELP FACULTY, RESEARCHERS AND STUDENTS

FIND

USE

MANAGE

VISUALIZE

SHARE

FIND OUT MORE

GO TO dataservices.library.jhu.edu

EMAIL dataservices@jhu.edu

SHARE AT archive.data.jhu.edu

JOHNS HOPKINS LIBRARIES | Data Services

Agenda

- What is “clean” data?
- Introduction to OpenRefine
- Data cleaning: NUFORC dataset
- Resources

OpenRefine

Learning Objectives

- Understand the importance of cleaning and standardizing data
- Carry out at least three transformations to standardize a messy dataset
- Become familiar with the reproducible aspects of OpenRefine and how to apply transformations to a new project

Data Cleaning

What is “clean” data?

Clean, or “tidy” data is structured in a way that makes it easier to analyze

“80% of data analysis is spent on the process of cleaning and preparing the data” (Wickham, 2014)

What is “data cleaning”?

- Process of re-structuring datasets in a standardized way
- Removing incorrect information
- Fixing inconsistencies, missing values, misspellings, etc.
- Preparing for data analysis or visualization

What is “data cleaning”?

Common symptoms of messy data include:

- Special characters (e.g. commas in numeric values)
- Numeric values stored as text/character data types
- Duplicate rows
- Misspellings
- Inaccuracies
- Leading or trailing white space
- Missing data
- Zeros instead of null values

Clean data?

Patient #	Height	Weight	Ex. Dur	HR	Location
154398			100	70	MD21218
582394			32	120	MD21044
814293	187	87	22	117	MD20770
392014	176	77	14	87	MD21202
178294	152	67	54	90	MD21218
239482	149	45	40		MD21001
403291	167	1000		96	MD21010
290300		97	33	70	MD21014
770543	154	62	43	65	MD21022
125765	160	50	88	98	MD21218

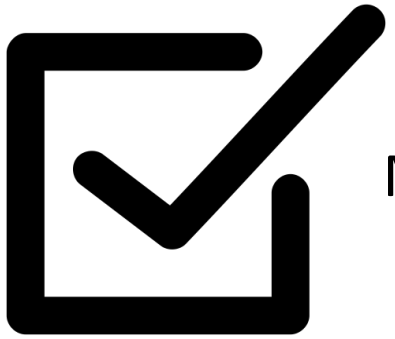
Does this patient code refer to PHI?

State and zip code in same column

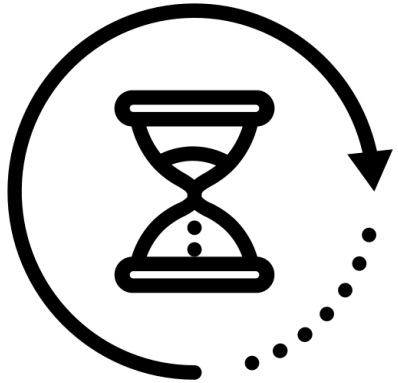
Missing value

Is this value correct?

Why clean your data?

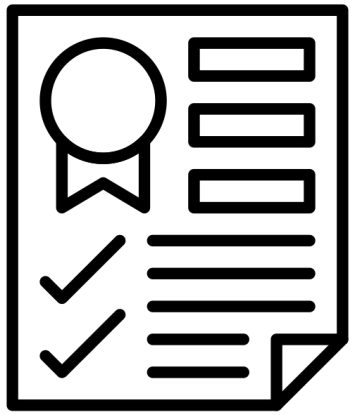


Makes data reliable and reusable



Saves time!

Facilitates further analysis or visualization, especially in specialty software



Your analysis is only as good as your data

OpenRefine

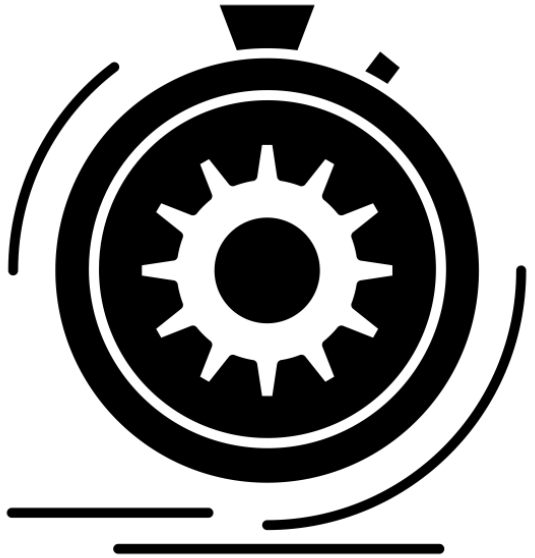
What is OpenRefine?

https://openrefine.org/

- Free, open source tool for cleaning messy data (previously developed by Google, as GoogleRefine)
- Application opens in a web browser, runs on a local server
- Perform actions using graphical user interface (GUI) or writing expressions in General Refine Expression Language (GREL)
- Stores all actions and transformations for a project, can be replicated in new projects

What can OpenRefine do?

- Clean data for further analysis or visualization
- Good with text data
- Add data from external sources – reconciliation and APIs
- Save your transformations to apply to new dataset – reproducibility!
- Works for medium to large datasets – 100,000s of rows



OpenRefine and Data Security

- OpenRefine is installed locally and stores data locally on your computer
- Does not send data outside of local environment (exception: Reconciliation)
- **Reminder:** it is the researcher's responsibility to keep data secure

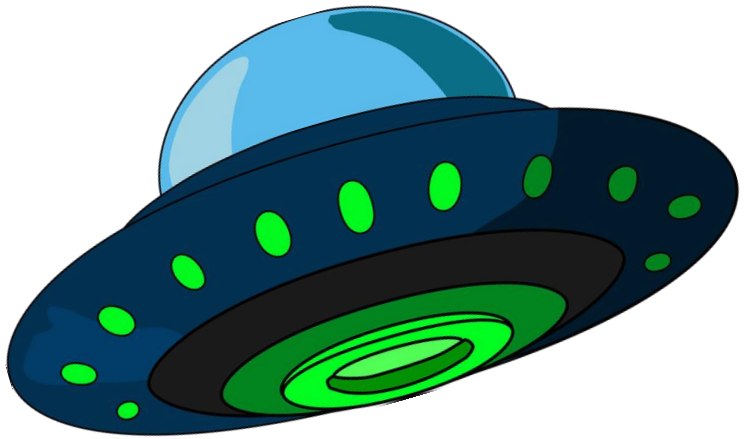
OpenRefine and Accessibility

- Main interface components should be compatible with text-to-speech software
- Many actions (clustering, reconciliation, etc.) may not be compatible with assistive technology
- Developer note in OpenRefine FAQ (scroll down to Accessibility heading)

Dataset: National UFO Reporting Center (NUFORC)

How we will clean today's data:

- Standardize values through editing, clustering, and writing GREL expressions
- Remove duplicate rows
- Split one column into two columns
- Reconcile data against external sources
- Add data from an external source



See the **Workshop Guide** for step-by-step instructions

Resources

Data Cleaning

- Towards Data Science: The Ultimate Guide to Data Cleaning
- Tidy Data by Hadley Wickham, Journal of Statistical Software

  DOI: 10.18637/jss.v059.i10
- The Programming Historian: Understanding Regular Expressions

  DOI: 10.46430/phen0033
- Regular Expression Cheat Sheet: https://regexcheatsheet.com/

Using OpenRefine

- OpenRefine Official Documentation
- The Programming Historian: Cleaning Data with OpenRefine

  DOI: 10.46430/phen0023
- Tutorial: Cleaning Data with OpenRefine
- University of Illinois Libguide: OpenRefine

Upcoming Data Services Workshops

- Sept 28: Data Cleaning in R (some R experience required)
- Oct 4: Bringing Your Research Alive with StoryMaps
- And more!

Register here: https://dataservices.library.jhu.edu/training-workshops/calendar/

JOHNS HOPKINS
LIBRARIES

Data Services

Contact JHU Data Services

GO TO

dataservices.library.jhu.edu

EMAIL

dataservices@jhu.edu

SHARE DATA AT

archive.data.jhu.edu

Helping you

FIND

USE

MANAGE

VISUALIZE

SHARE

DATA

TAKE OUR **SURVEY**

https://www.surveymonkey.com/r/openrefine

JOHNS HOPKINS LIBRARIES | Data Services