Data Management

Rachel Starry

Learning Goals

1. Understand the structure and appearance of datasets
2. Discover where and how to find data online
3. Apply learned concepts to clean raw datasets with OpenRefine

Total Estimated Time: 2 hours

Study

I. Thinking about Data (45 minutes)

What is a dataset? Where can we find raw data? How do we know whether the data is “good” or if it needs to be cleaned before it can be analyzed? These are some of the questions this module is designed to help you answer, before looking at examples of how to explore a dataset and transform it from its “raw” form into something that you can use for a Digital Scholarship project.

- What is a dataset? Start by looking over **the Wikipedia definition of a dataset**. Most datasets have the form of a table with columns that describe different variables (e.g. a person’s name, height, eye color, gender, etc.) and rows with individual records of the dataset (e.g. a row for information about Bob, a row for Alice, etc.). Can you think of some types of datasets have you encountered before, for a class, a project, or while keeping track of your own personal data? You might find it helpful to jot down your answers to these questions, as there will be time to share your thoughts at the end of this section.

- Where can we find raw data? The “External Links” section at the bottom of the Wikipedia page on datasets provides several good sources of data, which are not always "clean" enough to use for a project in their current form. One of the most common sources is the US Government: check out **Data.gov** - choose one of the topics to browse and *take a few (5-10) minutes to explore* some of the pages and datasets that are available for that topic to answer the following questions (jot down your answers to discuss as a group):

  - What file formats do the datasets appear in?
  - What do you think the difference is between a geospatial and non-geospatial dataset? Click here for an explanation of what defines geospatial data.
  - Is there much variety in terms of (a) the sources of governmental data? (b) the decades or years data comes from? (c) federal versus state data?

- Let's look at an example of a fairly clean dataset from Data.gov: the **US Renewable Energy Technical Potential**. Read about what sort of information this database includes, and then scroll down to see the data files for the database. The following exercise should take 15-20 minutes.
  - For this database, what file formats are available to view and download? When was the dataset created? Who created it?
  - Before you look at any of the datasets, *take a moment to think about what you expect* the differences to be between, for example, how the data might be reported in a PDF versus how it might be stored in an Excel notebook (based on what you already know about those file formats).
  - Click on the download link for the .xlsx file and open the file in Excel. How is the dataset organized? What values or categories do the rows and columns of the table correspond to?
  - Click on the download link for the .csv file. First open it using a text editor (such as Sublime Text) and then also open it using Excel. Is the data organized or stored differently in the .csv (comma separated values) file? What do you think the benefits might be of the .xlsx dataset versus the .csv dataset - for understanding the data, for manipulating the data, for visualizing the data? (You don't need to have definite answers for this question now; keep it in mind as it will be revisited at the end of the module.)
  - Finally, open the PDF report that accompanies this dataset. Skim the table of contents and briefly look over the body of the report (it is not necessary to read it in detail - skim the headings and try to get a sense of how the report is conveying the data). Does the report use the same organizational scheme as the dataset itself? How does it convey the data or insights generated from the data - using words, images, charts, etc.?

To wrap up this learning session, *take a few (5-10) minutes as a group* to share any insights you have gained about the way data is stored, presented, and made available online at a site like Data.gov.

II. Introduction: OpenRefine (10 minutes)

Why use OpenRefine (previously Google Refine) to study and manipulate raw data? It's an open-source tool that allows users to clean and organize datasets that are unstructured, inconsistent, or otherwise "dirty." Some of the advantages of using OpenRefine over other methods for cleaning data, such as Microsoft Excel or programming languages like R or Python, include that it is free and open-source, it is fairly easy to learn, and it can batch fix problems in the data (i.e. change many fields at once) more quickly than manually cleaning the data in a program like Excel.

- Look over **the OpenRefine website** to get a sense of what this software is and who is maintaining it. There are pages on how to download OpenRefine, on the history of the tool (important for understanding its transition from proprietary software owned by Google to free, open-source software supported by the user community), a FAQ page, etc.

- Watch **this short video** about what OpenRefine is and how it is useful for cleaning up data. It is not necessary to take detailed notes or understand exactly what is shown in the video; the goal is to see how the program can change data from a raw to a more refined state. You will practice this yourself during the following activity.

Sandbox

III. OpenRefine Tutorial: Data Cleaning & Exploration (45 minutes)

Goal: install OpenRefine and learn basic OpenRefine functionality for cleaning and exploring an open-source dataset

This YouTube video gives a brief overview of how to download, install, and use OpenRefine. Follow along as you watch the video, pausing if necessary at various steps.

The dataset used in the tutorial is no longer available, but a more recent version is. Find it here (it will take a few minutes to download).

Share

IV. Reflection (5 minutes)

Take a few minutes to write down, in your flog, 2-3 things you have learned or insights you had while completing this module. Think especially about what you found to be most challenging about completing this module, as you might decide to write about the your experience wrangling raw data for the first time in your flog.

V. Discussion (15 minutes)

- Share some of your personal learning accomplishments or insights you had about working with data while completing this module.
- What have you found most challenging as you have started learning how to find, explore, and clean raw data? Did you find OpenRefine easy to navigate and use? Do you think you are now familiar enough with OpenRefine to clean other datasets?
- What are some of the research questions or Digital Scholarship project goals that might be important to keep in mind when you (a) search for datasets online or begin building one yourself? (b) start exploring the data? (c) go about cleaning up a dataset?