When and Where

A Framework for Finding and Evaluating Social Science Data for Reuse

Ari Gofman, Social Science Data Librarian, Tisch Library, Tufts University, Ari.Gofman@tufts.edu

NUTRITION INFORMATION

This framework introduces undergraduate and graduate learners to questions to ask to successfully find and evaluate social science data for reuse. A graphic is presented on the board to support learners who prefer visual information, and more detail is provided verbally with multiple prepared opportunities for collaborative formative assessment. This modular lesson is usually combined with an interactive activity introducing a data source appropriate to a course's topics such as PolicyMap (a subscription tool), IPUMS (https://www.ipums.org), or Data.gov.

TARGET AUDIENCE AND NUMBER SERVED

This recipe serves advanced undergraduate students (juniors and seniors) and first-year graduate students working on a project that involves finding quantitative social science data for reuse, such as finding census data to use in a research paper. It also works well with medium-size classes (10–30 students) that can be separated into 3–4 breakout groups with different tutorials; each group is asked to teach back data sources using the *Finding Data Framework* covered in this chapter.

LEARNING OUTCOMES

Students will

- identify and select appropriate data for a given inquiry (microdata or aggregate data)
- identify how the creation process impacts the way the information can be used
- use brainstorming and other techniques when searching, including flexibility with proxying concepts for variables
- match information need with search strategy to represent the concepts in a research question

COOKING TIME

10–15 minutes for the basic activity (or as an asynchronous video); 45–75 minutes total for the theory presentation and additional resource exploration with teach-back activity.

DIETARY GUIDELINES

This framework provides an easy-to-understand theoretical guide to evaluating social science data for reuse and gives practical recommendations that improve the searching experience. It touches on the frames Searching as Strategic Exploration, and Information Creation as a Process from ACRL's *Framework for Information Literacy for Higher Education*.

INGREDIENTS

- Instructor presentation (can be PowerPoint or Google Slides).
- Student computers to search and access data sources.
- (optional) Worksheets for resource exploration and teach-back activity. Can be uploaded to Box or Google Drive for remote synchronous activity.

PREPARATION

Align the session with the broader academic course: discuss learning objectives and any relevant assignments with the professor. Discuss what sort of data students will be expected to find and the general types of analyses. If planning a longer session, collaborate to identify 3–4 resources for additional exploration. These can include exclusively data sources or a mix of data sources and secondary literature sources such as PsycInfo. If the session is in person, print enough worksheets for all students, plus two additional copies of each exercise for the librarian and professor. Walk through the instructions of each exercise to ensure the interface and links are effective. A sample worksheet is provided in the appendix.

INSTRUCTIONS

Welcome students to the session in whatever way you are accustomed to. For example, you could introduce yourself and make a pitch for scheduling a research appointment.

Outline the objective of the lesson. This module will introduce you to different types of data, identify important questions to ask when searching for and evaluating data, and provide an introduction to the resources you can find in your online research guide.
If the session is synchronous, ask students what they think of when they hear the word *data*. Most responses will be in the general theme of "numbers" or "Excel spreadsheets." After students have shared, offer a definition that data are units of information that can be analyzed and synthesized into new information. These sorts of quantitative, structured data make up a substantial part of research in the social sciences but are not the only form of data. Other forms of data include qualitative data in the form of words and text, such as interviews; images; video; audio; art; and more. However, this session will focus on structured quantitative data.

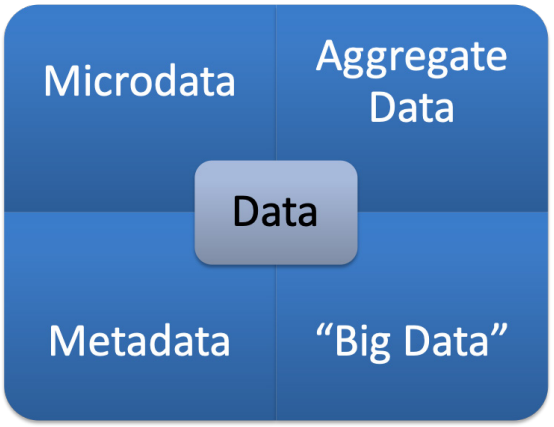

FIGURE 1
Data Types

Forms of Data
Proceed to introduce the framework for different forms of data offered in figure 1, which is one of many ways of breaking down types of data. We will use the US census as our example data set.

Microdata. Every ten years, the US Census Bureau attempts to count every individual in the United States, regardless of their immigration status. Every household receives a postcard asking them to respond to some basic demographic questions. Most governments do some version of a census to understand the makeup of the population for core functions such as taxation, military, education, welfare, and more. An individual person submits answers—which together in the United States will become approximately 330 million rows of data representing 330 million individuals. These are known as microdata—individual-level responses. They are very useful for statistics and running cross-tabulations. However, imagine Maria Gonzalez at 123 North Street, who just shared her income, address, the names and ages of her children, and even more private information. How would she feel about that information being publicly accessible? There can be significant privacy concerns in releasing microdata that include location, age, and other identifiable information, which means that most microdata are released only in an anonymized form.

Aggregate data. Most data are released in aggregate form. Instead of sharing all individual data, they are aggregated, or combined, into geographic or demographic categories. For example, the census may release a table identifying the percentage of people in Massachusetts who are in the 60–64 years age range. These data pose fewer risks to individuals and are often sufficient to answer many research questions.

Big data. The data we've talked about involve many observations, but you can engage with them using typical statistical techniques and readily accessible software and computational power. Big data are known for having high volume, variety, and velocity; we are not focusing on big data in this module.

Metadata. If you've ever downloaded a data set, you may have wondered: What do those variable names mean? What is the sample size, or sampling method? Have the data been cleaned? The answers to these questions and many more can be found in the metadata, commonly described as the data about the data. Metadata is similar to bibliographic information about a book. You search for and choose a book before you actually start reading it. First you look at its title, author, subject, publication date, publisher, genre, etc. Similarly, data sets have titles, authors, subjects, funding, geographic coverage, methodology, and more. This information can usually be found in a document called a codebook, or technical documentation.

After answering any questions about the previous section, continue to the next theoretical framework, finding data.

FIGURE 2
Finding Data Framework

Finding Data

By reordering the journalistic questions shown in figure 2, learners are guided by the instructor to ask six categories of questions.

1. *Where?* The biggest limiter in most social science data is geographic. What location are you researching? Is it the United States, Ethiopia, or Brookline Village? The second way to ask where refers to the unit size of location: an entire country, or broken down by state, county, census tract, etc.? Generally, the smaller the unit of geography, the more restricted the data will be.
2. *When?* The next question is temporal: When do you want data from? The year 1851, or the most recent available? You'll also consider the type of time you need. Do you want a snapshot (for example, population in 1851) or do you require longitudinal data that you can compare over time, such as how the unemployment rate changed each month from 2007 to 2012? Make sure that the same question is being asked of the same representative population. This generally means you want to find longitudinal data in a single data set to ensure comparability.
3. *Who?* Ask about people: Who do you want data about? The entire population? A particular age, gender, socioeconomic class, race, or other demographic group? Are any of the groups you want to research protected or vulnerable, such as children or people who are incarcerated? If so, you may find that access to those data is more restricted to protect people who are less able to advocate for themselves. You will need to be cognizant of your responsibility as a researcher to do no harm.
4. *What?* What variables could answer the research question? This is intentionally asked after the three previous questions because often there are multiple variables that could address your question. Emphasize flexibility and creativity in accordance with Searching as Strategic Exploration to proxy the concepts in a research question. If we want to evaluate farmer incomes, what variables can we look at? Answers

could include self-report survey, tax filings, or proxies such as rainfall. To find out what the variables actually mean, check the codebook, which describes how that question was asked and what it represents. Ask what type of data: Do you need microdata or aggregate data?

5. *Why?* Why would someone collect these data? This question can be used to identify stakeholders who may gather data, which you can then intentionally search for. Where will you be able to find them, and what biases do their goals introduce? If we want to get data about electricity access in Nigeria, who are the stakeholders? Answers include the power company, the government, nongovernmental organizations (NGOs), researchers/academics, and local community groups or newspapers. We discuss strategies for searching for each stakeholder's data in different sources. What are the goals or incentives of each group, and how might that influence or bias their results? For example, is Coca-Cola funding this research on sugar?
6. *How? How* refers to how these data were gathered and how that affects your analysis? I discuss known data quality issues through humor—by observing that everyone says they're taller, richer, and more attractive than they actually are—to introduce methodological considerations of self-reported data, among other discipline-relevant issues.

REVIEWS/ASSESSMENT STRATEGY

A graphic is presented on the board to support learners who prefer visual information, and more detail is provided verbally with multiple prepared opportunities for collaborative formative assessment, as discussed above.

For longer sessions, participants have the opportunity to apply the framework to an exploration of a particular data source using a guided worksheet in a small group and demonstrate their understanding by teaching back to the class.

ADAPTING THE RECIPE

Worksheets can be adapted to different data sources with the same basic structure and questions, with specific instructions on accessing the desired information.

APPENDIX: SAMPLE WORKSHEET

Resource: NYC Open Data Portal

Topic: Start and end locations of taxis in 2019

Adapted from a worksheet by Erica Schattle

Goal: To find out more about the contents of this resource and think about if and when it might be useful for your research.

Deliverable: After completing the steps and considering the questions outlined below, you will have *3 minutes* to present and demonstrate what you learned about this resource to the rest of the class.

Steps

- Navigate to https://data.cityofnewyork.us/. What is it? Who funds it? What does it include and not include?

- Begin searching for data on your topic.
- What geographic regions are covered in this resource? Time periods?

- Identify a data set of interest to you. Is the data set derived from another source, or is it primary? How can you tell?

APPENDIX: SAMPLE WORKSHEET (continued)

- In what file format(s) is this data set available?
- Download the data set for Excel (.csv or .xlsx file formats). Explain the steps you took to download the file(s).
- Create a citation for your data set (hint: https://www.datacite.org/cite-your-data.html has more information).
- Do other municipalities have similar data? Can you find open data portals for Mexico and Cambridge, MA? How detailed are their data? How do you find geospatial data? (Divide your group in half—one half looking at Cambridge and one half looking at Brazil).

Questions to Consider for Your Presentation

- A basic overview of the resource: What type of data does it include? What does it not include? Who runs it, and what are their goals?
- What geographic regions are covered in this resource? Time periods?
- Is the data set derived from another source, or is it primary? How can you tell?
- In what file format(s) is this data set available? Can you get a command file?
- How granular were the variables in the data set you downloaded? Are these microdata or aggregate data?