Data Sources 1

Ryan Womack

2025-04-09

Data Sources 1

major US data resources, locating data, with focus on diversity and data for minoritized and underrepresented groups

1 Overview

What do we think about when we look for data? Data is incredibly varied, and each researcher has their own needs. Finding a match between the researcher and a dataset is something of an art at times. Recognizing this, this workshop provides some general guidelines for thinking about and searching for data, and mentions major sources that most researchers will want to be familiar with. However, it is only a quick overview that scratches the surface of a complex topic. The goal is to provide a framework for searching for and locating data, as well as pointers to more detailed guides.

1.1 Scope

This workshop is centered on social science data sources, which can be put to multiple uses. We will make some nods to science data and humanities data, but the specialized needs of researchers in these fields are mostly beyond the scope of this brief introduction. Also this workshop discusses sources of data that are useful for research about the United States (although some sources also have international data). Federal government, New Jersey (but not other states), commercial/subscription, and nonprofit sources of data are covered. Note that Data Sources 2 (to be released in the future) will focus on international data.

2 Searching for data - the process

Researchers often have a very specific need in mind, but it is always important to reflect on the "W" 's of data collection.

- **Who** would collect such data?
- **What** data would be collected?

- **Why** would they collect it?
- **When** was it collected?
- **Where** was it collected?

The answers to these questions can guide us to appropriate data sources.

2.1 Case studies

Let's think about applying these questions to a few different research scenarios.

- **Business** The researcher is interested in financial details and actions of a specific company. Who collects this data? Public companies are required (the "why") to disclose financial details in annual and quarterly filings (the "when") to the *Securities and Exchange Commission*. This publicly available data becomes part of many data compiliations (the "where"). The "what" that is collected is the data in audited balance sheets and cash flow statements, and can be quite detailed. However, for other information, such as sales figures of particular products, costs associated with divisions of the company, and other details, the "why" is quite different. Releasing this information is not required, and so each company may or may not disclose such information (usually releasing only what makes them look good, such as when sales are growing.) Private companies also are not required to say much about their operations. Government sources will not help for this second category of information. But commercial business databases and business news sources do make efforts to investigate and compile this information to the extent possible.
- **Social behavior** A researcher is interested in the linkage between teenage illegal drug use and health over time. Who collects this data? Because it deals both with minors, a sensitive population, a sensitive topic, and sensitive biological data, it is not the kind of data that people can be compelled to provide. Only surveys that are carefully designed to provide confidentiality, encourage participant participation, and guarantee that researchers will not misuse the data can approach such questions successfully. One example of this is the *The National Longitudinal Study of Adolescent to Adult Health (Add Health)*, whose use mostly falls into the category of *restricted data*, which will be discussed further below. The "who" for such data may be funders who are interested in the outcomes of such behavior ("why"). The "when", "where", and "what" may be limited by the circumstances of collection.
- **Health statistics** A researcher is interested in cancer rates and obesity rates. Thinking about who or when this is collected may help understand the kind of data available. Hospitals track events that occur there. A cancer diagnosis or operation will be reported systematically. See *CDC Wonder*. Other things that affect the general population, such as obesity, can only be inferred by sampling. See *CDC Obesity* based on the survey, *Behavioral Risk Factor Surveillance System* where participants self-report their weight.

2.2 Strategies

These questions may lead you to specific data sources or providers. If you are not finding the data you're looking for, it is usually worthwhile to frame your search more generally. Internet searches are fine, but can be more relevant if you take into account the kinds of sources available.

Be aware that data sources change names and focus over time, or may be discontinued. So something that looks like a really great source might not be as current as you'd like. Also, as we'll highlight below, some minoritized populations might not have received as much coverage in data sources as other groups. Extra effort may be required to uncover data in those cases.

Finally, it is important to recognize that for original research, data sources may not be obvious. When searching isn't fruitful, looking at research articles that have investigated similar topics is often the best avenue to take. Other authors may have discovered a useful data source that you can also tap into. On the other hand, if other authors are collecting their own data to answer research questions like yours, that is a good sign that no general data source is out there for that kind of question.

3 Sources

As a reminder, this session focuses on US data.

3.1 Federal Government

The federal government is a massive provider of data, having the ability to comprehensively collect data across the country at scale, and to compel compliance in ways that others cannot. The federal government is also the custodian of administrative records such as those generated by the *Internal Revenue Service*, the *Social Security Administration*, or *Immigration and Customs Enforcement*. Most government agencies will have a section on their website titled "statistics", "data", or "reports" where you can browse the kinds of general data they release to the public.

Although, at the time of this writing, many data sources (e.g., *National Oceanic and Atmospheric Administration*, *National Center for Education Statistics* have seen removals and cuts under the current administration. The *Data Rescue Project* is tracking these events, coordinating archiving and preserving data, and providing links to backups. *DataLumos* from ICPSR is a major home for archived government data.

Some highlights of US government data

- **Census**. The Bureau of the Census is the largest collector of demographic information. The original Decennial Census has seen most of its questions moved to the American Community Survey, an annual sample that provides more up-to-date information, while the Decennial Census retains its role as a complete count of the population. The Census produces numerous other surveys on business and other topics, as well as microdata.
- **Bureau of Labor Statistics** - The BLS collects employment, inflation, and other information related to work, including the *American Time Use Study*.
- **FRED** and **ALFRED** - The Federal Reserve Bank collects and archives major economic data series in their comprehensive FRED (Federal Reserve Economic Data) site, along with ALFRED (Archival FRED), which archives original releases of data. The original releases can later be updated and corrected (in FRED), but for economists and others studying how markets originally reacted to information, ALFRED is important.

- **Crime Data Explorer** - represents the current information from the Federal Bureau of Investigation's *Uniform Crime Statistics*.
- **CDC Wonder** - the Center for Disease Control's data collection, a first stop for mortality, disease, and other health statistics.
- **Data.gov** - Open data from across the federal government, a large and sometimes chaotic archive of full datasets.

There are many, many more possibilities with federal data, but we'll stop with these highlights.

3.2 New Jersey - State Data

- **New Jersey State Data Center** - The New Jersey State Data Center (NJSDC) is a cooperative project of the State of New Jersey and the U.S. Bureau of the Census, serving data users in the public, private, and academic sectors.
- **NJOIT Open Data Center** - various freely available datasets from NJ government on this open data portal

Just as in the Federal government, various New Jersey departments release data and reports via their websites. You can start at *NJ.gov* to browse.

Other states operate similarly, so just apply the same principles to your search if you're looking for state-level data outside of New Jersey.

Other open data sites also exist, such as *NYC Open Data* for New York City and *OpenDataPhilly* for Philadelphia.

3.3 Nonprofit organizations

Data *archives*, like ICPSR, collect data from multiple sources. Other projects, like IPUMS, have specific aims and goals.

- **ICPSR** - Rutgers login - largest social science data archive in the world, with fully documented research datasets of all shapes and sizes, from focused and specific as created by individual researchers, to large-scale nationwide longitudinal data from government. Predominantly US data, but has international coverage. Also many *topical data collections*, learning and teaching material, and a bibliography of research that uses the data.
- **Roper** - Rutgers login - archive of public opinion surveys and polls, including downloadable full datasets
- **IPUMS** - IPUMS provides census and survey data from around the world integrated (harmonized) across time and space. IPUMS integration and documentation makes it easy to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community contexts. Data and services available free of charge (with registration). Focus on microdata. Excellent coverage of historical US Census microdata in research-friendly formats. NHGIS component provides GIS-coded Census data.

- **General Social Survey (GSS)** - a nationally representative survey of adults in the US conducted since 1972. The GSS collects data on contemporary American society in order to monitor and explain trends in opinions, attitudes and behaviors. The GSS has adapted questions from earlier surveys, thereby allowing researchers to conduct comparisons for up to 80 years.
- **American National Election Studies (ANES)** - produces high quality data from its own surveys on voting, public opinion, and political participation.
- **re3data** is an index to data repositories, ranging from discipline specific to general, and can be used as a discovery tool for potential archives of data.

3.4 Commercial (subscription)

These require a subscription to access. Links are to the Rutgers-subscribed versions, which require authentication via Rutgers NetID. If you are viewing this from another institution, check your library for access.

- **Sage Data** - a large collection of statistics from state, federal, international, and private sources in a standardized, searchable format, with download and graphing capability.
- **Social Explorer** and **PolicyMap** are two resources that provide easy access to demographic data in map and downloadable form for the US.
- **Statista** provides quick lookups of current statistics with a business focus.
- **Statistical Abstract of the US** for quick statistical tables. For pre-2000 tables, try **Historical Statistics of the US**.

There are many business databases that provide information about companies. For detailed financial information, *COMPUSTAT* (corporate financials) and *CRSP* (stock prices) are two of the most commonly used. At Rutgers these are available through the Rutgers Business School's subscription to *WRDS*.

4 Restricted data

A word about restricted data. In the social sciences, there are privacy concerns around releasing data about sensitive topics, personally identifiable information, and data about vulnerable populations (minors, incarcerated, mentally ill, etc.). This can result in the free, publicly available datasets being somewhat limited. Accessing the more detailed data, with information at the individual level (microdata) often requires additional steps, where the researcher must apply to use the data and demonstrate that they can handle it responsibly. ICPSR is one major provider of restricted data.

Restricted data may have to be accessed in a secure data enclave. The *Federal Statistical Research Data Centers* provide secure environments supporting qualified researchers using restricted-access data while protecting respondent confidentiality. These restricted-access data come from censuses and surveys of businesses and households, linked employer-employee data, and administrative records from federal and state agencies and other sources. The Census Bureau

maintains and enhances these confidential microdata through the FSRDC network to equip researchers with the resources they need to generate insights for the public good.

5 Data for minoritized and underrepresented groups

Many organizations and projects focus attention on data and issues related to minoritized and underrepresented groups, surfacing data that has lacked attention previously. Some examples follow:

- **Resource Center for Minority Data (ICPSR)** - highlights from the ICPSR collections across many dimensions of minority populations and experience.
- **Diversitydatakids** - data focused on equity for children of all races/ethnicities.
- **Civil Rights Data Collection** - on education and civil rights, since 1968.
- **The Sentencing Project** - advocates for effective and humane responses to crime that minimize imprisonment and criminalization of youth and adults by promoting racial, ethnic, economic, and gender justice. Data tools and resource library available.
- **Mapping Inequality** - redlining in New Deal America
- **Mapping Prejudice** - visualizing the hidden histories of race and privilege in the built environment
- **IASSIST Antiracism resources guide** - for more data sources, check here.

6 Further exploration

For further information, consult the following research guides from Rutgers University Libraries and elsewhere:

- Data by Subject - links to the Economics, Political Science, and Social Work data guides
- GIS data
- Qualitative Data
- Criminal Justice Guide
- Elections guide (Princeton) - from Jeremy Darrington, a comprehensive guide to finding elections data
- Humanities Data
- Linguistic Corpora