Strategies for Finding Data and Statistics

GR ONL Success Series

Drexel University Libraries

Sarah Hughes, MSLIS, MA,
sh3675@drexel.edu
Spring 2026

Drexel University Libraries

Agenda

- Data vs. Statistics
- Step by step: Finding data & statistics
- Overview of web resources and library databases with data & statistics
- Citing Data and Statistics
- Wrap up and questions

Data and Statistics

The words *data* and *statistics* are often used interchangeably, and they can overlap, but they are different.

Data are records of individual facts and are the raw information from which statistics are derived.

Statistics are conclusions, interpretations, or summaries you derive from analyzing data.

Data and Statistics

- **Data** can describe numeric, textual, photographic, audio, or visual formats that have been produced or to support research. This includes lots of things, but many of the databases and repositories have a specific focus on a *type of data* (numeric, textual, etc.) or a *topic* (e.g., demographics, economics, genomics, etc.).
- **Statistics** usually refers to numeric summaries of data presented in tables. These include counts, percentages, ratios, and other results of data analysis.
- We will largely cover the main types of resources that contain:
  - Databases & Statistical Indexes
  - Data Repositories
  - Institutional Repositories
  - Web Resources

1. How to go about finding data - use the research literature

The top strategy for finding data that already exists is to look in the academic literature.

1. Visit a Drexel Library Database/Resource (or Google Scholar). Conduct a search using **keywords** that match your research question.
2. Jump to the **Methods** section of the article. Every empirical research paper will have a Methods section. (*Not that not all peer-reviewed literature is empirical. It is possible you will find articles that lack a methods section!)* Every Methods section will describe the data in some way. Note:
   - What **data** do they use? Did they **cite** their source? (Finding the data citation can make it much easier to track down exactly where to find the data.)
   - You can also track: Author/year of publication; Claim; Data; Dependent variable/estimate technique; Significant findings
3. Iterate. Try this technique with different keywords to find more articles.

1. Use the Research Literature

Start taking notes on potential data sources. Make a habit of jotting down the data used in each study you read to make it faster when you come back later in your search for data.

Can help you see and articulate how your research contribution is unique. Consider taking these notes in a table following for easy reference:

- Author(s) and Year of Publication
- Claim
- Data
- Dependent Variable/Estimation Technique
- Significant Findings

2. Ask yourself what you really want?

Ask yourself about your research question. What is your topic and what kind of claim do you want to make? Now consider, what kind of evidence would you need to support those claims? You can use that to describe your ideal dataset.

In a perfect world, what do you want for your:

- **Unit of analysis** - individuals, companies, counties, countries;
- **Geography** - Virginia, the American South, USA, Europe, all countries;
- **Time period** - 2000 to 2023, most recent year only; and
- **Frequency** - quarterly, annually, every 10 years for 50 years?

Keep in mind that just because you can imagine it, **does not mean it exists**. This is especially something to consider when you are under a deadline to finish a project.

2. Ask yourself what you really want?

Common limitations to finding your perfect dataset:

- The data do not exist. No one has ever spent the time or money to collect the data you have in mind.
- The data exist, but you can't access it. Sometimes datasets are restricted, and you can apply to access them.
- Sometimes data cost money to access.
- Intellectual property rights and Terms of Use can be a hindrance when trying to collect data yourself (i.e., it is not legal to web scrape every website you come across).
- Data might exist but are not in a machine-readable format. This might be true for historical forms of data.

3. How would the data have been collected?

- Consider who might have collected the data? Data are expensive and time consuming to collect. They don't just appear out of thin air. Common sources of data are:
  - Researchers.
  - Government agencies (e.g., Census, BLS, BEA).
  - NGOs and IGOs (e.g., UN, World Bank).
  - Think tanks, research organizations, private companies (e.g., Pew Research, Gallup, Bloomberg).

4. Use Search Tools

Library Databases

Web searching

Library Guides

5. Evaluate potential data

Ask yourself questions about the data that you find.

1. Find overview information

- Who created the data? Why? What is the scope? What is the geography and time period?

2. Find technical documentation

- Look for and download or document technical documentation about the dataset, including information on how it was created (e.g., survey, administrative reporting, direct measure), variable definitions, indications of what was included or excluded. Survey instruments are also helpful.

3. Identify the Download Options and Access Restrictions

- Who gets to use the data? Contact a librarian if you are unsure if you can access it. What formats of download are available - CSV, text, Excel?

Searching at the Variable Level

Most data catalogs describe datasets at the "study level." This means that when you search by subject or keyword, your search terms are matched against words used to describe the whole study broadly, not all the specific topics addressed within it. This practice is similar to the way books are described in the library's catalog.

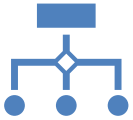
However, it is often the case that you would like to search at the "variable level." For example, you might like to search for all surveys that ask a question about happiness. What is a researcher to do? Because of improvements in data description, some datasets are searchable by variable.

Searching at the Variable Level

- ICPSR Variables Database
  - Search across approximately 1300 studies within the ICPSR archive at the variable level.
- UK Data Archive
  - Free online resource.
  - Change the dropdown option from "All of study description" to "Variables."

Searching for Raw Data

The term "raw data" is being used increasingly to mean many things, but in the world of social science research data, it means something much more specific.



If you are looking for microdata (data at the individual level, analyzable at the level at which it was collected, e.g., persons, households, etc.) formatted for use with statistical software, then you're in the right place.

Data in ASCII format, or formatted for use with statistical packages (e.g., SPSS, Stata, R), can often be found most efficiently by searching collections and archives of data.

Searching for Raw Data

Metadata Search

- Search across a large and growing collection of datasets in multiple disciplines that have been registered with the DataCite organization.

Dryad Digital Repository

- Nearly 10,000 publicly available scientific datasets.

Dataverse

- Search across the data collections of over 200 institutions, researchers, research centers and collections. Click the 'i' for more information and watch the format listed to determine whether the data are "raw" or not.

Replication Data

- A staggering amount of datasets can be found out on the free web, but can be tricky to find.

UK Data Archive

- A premier data collection of datasets from the United Kingdom and beyond. Carleton does not have access to much of this data, but it is a great place to search for international data sources. Consult with a librarian about access.

Data and Information Services Center at UW-Madison

- Another large data collection.

Finding Statistical Resources in Dragon Search

Web sites with data almost certainly have the most current numbers, but don't always include data collected in previous years.

Use Dragon Search to locate books or documents with statistics and data beyond what is available from databases online.

Books often contain appendixes with tables. Government and non-government agency reports and statistical compendia are likely to contain -- or consist nearly entirely of -- statistical tables and charts.

Finding Statistical Resources

Search

- Most statistical publications are cataloged with a subject heading that contains the word "**statistics.**"
- Perform an Advanced Search with one dropdown set to "Subject contains" and type the word "statistics."

Browse

- To find older statistics in print, browse the library stacks. This is a useful way of getting around the problem of needing to guess the right words for an effective catalog search.

Chase Citations

- Use the books and articles you already have in hand and skim the bibliographies and "Data" sections of the text for the names of studies, datasets, or collection agencies. Search for these names in the catalog.
  - For example, you might see the IMF's "Direction of Trade Statistics" cited frequently in the literature. Their web site provides access to the current data.

Factors to Consider When Evaluating Statistics - Evaluating Data and Statistics

Source

- Who collected it?
- Was it an individual or organization or agency?
- The data source and the reporter or citer are not always the same. For example, advocacy organizations often publish data that were produced by some other organization. When feasible, it is best to go to the original source (or at least know and evaluate the source).
- If the data are repackaged, is there proper documentation to lead you to the primary source? Would it be useful to get more information from the primary source? Could there be anything missing from the secondary version?

Factors to Consider When Evaluating Statistics - Evaluating Data and Statistics

Authority

- How widely known or cited is the producer? Who else uses these data?
- Is the measure or producer contested?
- What are the credentials of the data producer?
- If an individual, are they an expert on the subject?
- If an individual, what organizations are they associated with? Could that association affect the work?

Factors to Consider When Evaluating Statistics - Evaluating Data and Statistics

Objectivity & Purpose

- Who sponsored the production of these data?
- What was the purpose of the collection/study?
- Who was the intended audience for or users of the data?
- Was it collected as part of the mission of an organization? Or for advocacy? Or for business purposes?

Factors to Consider When Evaluating Statistics - Evaluating Data and Statistics

Currency

- When were the data collected? Not always close to when they were released or published -- there is often a time lag between collection and reporting because of the time required to analyze the data.
- Are these the newest figures? Sometimes the newest available figures are a few years old. That is okay, as long as you can verify that there isn't something newer.

Factors to Consider When Evaluating Statistics - Evaluating Data and Statistics

<u>Collection Methods & Completeness</u>

- How are the data collected? Count, measurement or estimation?
- Even a reputable source and collection method can introduce bias. Crime data come from many sources, from victim reports to arrest records.
- If a survey, what was the total population -- how does that compare to the size of the population it is supposed to represent?
- If a survey, what methods used to select the population included, how was the total population sampled?
- If a survey, what was the response rate?
- What populations included? Excluded?

Factors to Consider When Evaluating Statistics - Researching Your Data

To fully understand your data, what it can tell you, and how much it will strengthen your argument, look in the following places.

- Data Documentation
- The Research Literature
- Specialized Bibliographies

Factors to Consider When Evaluating Statistics - Data Documentation

Find the website of the institution that creates, disseminates, or hosts the data you are using and look for documentation in the form of:

- User Guides
- Codebook
- Questionnaire or "survey instrument"
- Statistical overviews can be helpful (ex. "How Australians Use Their Time)

These types of publications will give you the following information about your data (and much more):

- Purpose / overview / background of survey
- Methodology
  - Scope and coverage (who's included, who's not?)
  - Sample design
  - Data collection methods
  - Response rates
- Data quality

Research Literature

Use the following **indexes** to search for research articles that use the data you're researching. Three main kinds of searches will help you:

- Search for the name of the data in the abstracts
- Do a citation search for the data
- Search on the topic and look at the data used by the research you find. How does their data compare to yours?

Web of Science

A vast index of literature from the sciences, social sciences, and humanities. Also offers cited reference searching to see who has cited a particular document. 1898 - current.

Google Scholar

Use Google's special academic search to locate scholarly articles from across disciplines. Set Library Links to Drexel Library for full-text access. Use the "cited by" link to see who has cited a particular work you're interested in.

General Compendia

- Statistical compendia are the first place to look for statistics and sources of data when you don't know who collects the data and you suspect they may be collected administratively (e.g., by a government agency) or regularly (e.g., an annual survey).
- Even if the compendium does not have the exact data you need, use it as a "guide to data sources" and follow the source notes to the originating agency, where you are likely to find more similar data and statistics.

General Compendia

Statistical Abstract of the United States

First place to look for government statistics. Browse by topic or search by keyword and find tables on nearly any topic. Follow the source information to learn where to look next for similar statistics or data. Also a great source to learn about the major data-collecting agencies and programs in the U.S. government.

Statista

A vast statistics portal integrating data and facts onto a single platform, categorized into market sectors.

Nationmaster

This is a free web, ad-supported site with comparative national statistics. Very useful for brainstorming variables and determining sources.

Compendia of Foreign Countries

Most countries have agencies responsible for collecting national statistics. These agencies usually produce compendia like the Statistical Abstract of the United States. A good way to begin finding data for a foreign country is to identify the national statistical agency, then look on their site for their annual statistical yearbook or report. Also useful are the yearbooks of central banks.

International Statistical Agencies

- A very straightforward list from the Census, listed at the bottom of this page.

National Statistical Offices

- Links to each country's national bureau of statistics, maintained by the UN Statistics Division.

Compendia of Government and Nongovernmental Agencies

Like national statistical agencies, government agency and NGO and IGO web sites can be large and occasionally difficult to navigate, especially for the first time visitor. Also like statistical agencies, most government agencies and NGOs and IGOs have annual reports, yearbooks, or compendia. Look for these annual publications on the web site as a guide to available data. Below are just a few examples.

NCES Digest of Education Statistics

- Compilation of U.S. statistical information from public and private sources, covering pre-K through higher ed.

Social Sciences: Data Archives and Data Catalogs

Roper Center Survey

The Roper Center data archive at the University of Connecticut includes datasets for secondary analysis from over 10,000 studies of public opinion. Includes surveys by Gallup, NORC, ABC News and Washington Post and more.

Data and Program Library Service

This data library at the University of Wisconsin-Madison has a searchable catalog of data as well as a browsable archive of locally collected data.

UK Data Archive

A premier data collection of datasets from the United Kingdom and beyond.

Odum Institute for Research in Social Science

Extensive collections of Census data, public opinion data, health statistics, and other topics.

Sciences: Data Archives and Data Catalogs

Dryad Digital Repository

Nearly 10,000 publicly available scientific datasets.

re3data.org

A global registry of research data repositories that one can browse or search and filter. In 2015 it was combined with DataBib.

Data Repositories Wiki

A list of open data repositories for many different disciplines - primarily STEM, but others too.

National Science Foundation Data

The NSF provides data for use in STEM fields.

Informally Shared Data Repositories

- Also referred to as generalist repositories, these repositories accept data from all kinds of dataset authors, or in all kinds of formats, or on any topic.
- They are generally not curated, so quality can vary widely. However, because they have few barriers to deposit, data can be shared here quickly.

Informally Shared Data Repositories

Dataverse

An open repository hosted at Harvard in which authors can deposit their replication datasets. Includes datasets from important journals and research projects, such as American Journal of Political Science and survey organizations like Gallup.

Dryad Digital Repository

Nearly 10,000 publicly available scientific datasets

FigShare

A place where anyone can post research data and publications. There is no editorial or curatorial process to ensure quality, so evaluate what you find carefully.

IEEEDataPort Freely available research datasets shared with the global technical community. Code often accompanies datasets.

Informally Shared Data Repositories

Kaggle Tutorial
Learn how to find a dataset using specific search terms, read and understand what's in your dataset, and check the validity of your dataset.

Mendeley Data
A free repository for sharing data of all kinds.

OSF (Open Science Framework)
A free platform for sharing research results, where datasets are often shared as components of projects.

Tableau
One of the more robust options for uploading, visualizing & exploring, downloading and publishing your data. Charts and graphs built on your data will stay live, so if you update the data, the visualizations update accordingly.

Zenodo
An open data repository, which contains all forms of research output including datasets. Originally formed in the EU for open science, but open to users worldwide.

Science Data Bank A public, general-purpose data repository dedicated to open science, operated by the Chinese Academy of Sciences.

Data Producers

Numeric data are collected and disseminated by a variety of organizations, from think tanks to the national statistical offices of governments.

Thinking about these questions can be helpful in your data search:

- Are these data likely to have been collected?
- Who might have been interested in collecting these data?
- Would these individuals, organizations, or government agencies have reason to share these data? Where? Are these data available to me and where?

Intergovernmental organizations such as the World Bank include data collection and dissemination in their mission. Think tanks and nongovernmental organizations might collect and share data to further their communication and advocacy objectives. Governments collect a wide range of data on everything from their nation's trade to censuses of populations.

Keep in mind that not everything worth being counted is in fact yet counted, and data might be collected but kept proprietary or sold at prohibitive cost. Data that are shared might still be hard to access.

Government Agencies

Even with Google, it can be very hard to locate government websites.

Vanderbilt's Statistics -- National Agencies and Compendia

Governments on the WWW

International Government Information at UC Berkeley

Northwestern'sList Foreign Government

- Links to major agencies within countries around the world.

U of MI's Foreign Government Data Sources

U.S. Census: International Statistical Agencies

Intergovernmental Organizations (IGOs)

Intergovernmental organizations are major data producers.

IFS: International Financial Statistics

Thousands of time series covering more than 200 countries and areas, published by the IMF (International Monetary Fund). Covers 1945 - current.

OECD Statistics Portal

Starting point for links to OECD statistics and data. We don't subscribe to all OECD data but can obtain publications via interlibrary loan if needed.

UNData

A single gateway to the UN's 24 statistical databases and 60+ million records).

WDI: World Development Indicators

Statistical data for development indicators and time series data for over 200 countries and 18 country groups. Data includes social, economic, financial, natural resources, and environmental indicators. Covers 1960 - present.

Nongovernmental Organizations (NGOs)

Start searching for information on NGOs by:

- Reviewing some excellent library research guides.
- Then turning to the larger directories and search engines.

Berkeley's guide to researching NGOs

Duke NGO Research Guide

UN NGO Database

There are currently 3000+ NGOs with consultative status to the UN Economic and Social Council, searchable in this database.

Google custom NGO search

Search of Non-Governmental Organizations (NGOs) websites. Sites were chosen based on their consultative status with the United Nations Economic and Social Council (ECOSOC) and also collated from University of Minnesota Human Rights Library, Duke University Libraries' NGO Research Guide, and the World Association of Non-Governmental Organizations (WANGO).

U of MN Human Rights Library

Think Tanks and Research Organizations

Foreign Affairs Online

Think Tank Search Engine

- Using a specialized Google search engine, directly search the web sites of hundreds of think tanks.

Michigan State University's Directory of Think Tanks

National Institute for Research Advancement's (NIRA) Think Tank Directory

Think Tank Funding Tracker (Quincy Institute)

- Publicly available repository of foreign government, U.S. government, and Pentagon contractor funding of the U.S.'s top 50 foreign policy think tanks going back to 2019.

University of Michigan's GovDocs Think Tank Directory

Citing Data in Your Research

There is no universally recognized way to cite data. It is best to follow the guidelines of the style you are using to the extent that it addresses data.

A dataset citation includes many of the same components of a traditional citation.

- **Author(s)**: Who created the data? an organization, individual, group of individuals
- **Title:** Title of the dataset or name of the study
- **Date of publication**
- **Publisher and/or Distributor:** who has made the data available
- **Edition/Version**
- **Electronic Location or Identifier**: a URL or DOI where data was found

Some other elements to consider adding to a citation:

- **Date of access**: When did you access or download the data?
- **Format:** What type of file or resource is it?
- **Scale**: Use when citing a map. Not usually used for datasets.

Citing Data in Your Research

How to Cite Data A useful guide with examples for published statistics and datasets in different styles such as APA and ASA.

IASSIST Data Citation Guide From the International Association for Social Science Information Services and Technology (IASSIST), an association of data experts. Includes examples.

"Citing Data" A great guide from MIT's Social Science Data Services

"Citing Online Numerical Data" A guide with a helpful diagram from Lafayette's library.

Why and How Should I Cite Data ICPSR's guidelines and examples.

WashU's GIS Citation Guide

Disciplinary Standards for Citing Data

American Sociological Association - Checklist for Preparation of Manuscripts
Follow guidance for software, tables, and figures.

American Economic Review: Data Availability Policy "It is the policy of the American Economic Review to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide to the Review, prior to publication, the data, programs, and other details of the computations sufficient to permit replication."

APSA Guide to Professional Ethics, Rights and Freedoms "6.1 Data access: Researchers making evidence-based knowledge claims should reference the data they used to make those claims. If these are data they themselves generated or collected, researchers should provide access to those data or explain why they cannot.

American Sociological Review Notice to Contibutors See esp. section 5 "References" under Manuscript Submission, Manuscript Preparation.

Questions?
Thank you!

sarah.hughes@drexel.edu

https://libcal.library.drexel.edu/appointments/sarahhughes

Acknowledgements
Much of this content is adapted from:
Carleton College Gould Library. "Data, Datasets, and Statistical Resources - When you're not sure where to start."