

# Strategy for Finding Statistics and Data

Created by Grace Liu, Business Librarian, West Chester University; Advised by Bobray Bordelon, Economics & Finance Librarian/Data Services Librarian, Princeton University; Reviewed by Rashelle Nagar, Business Research & Collections Librarian, Stanford University

## Step 1: Assess your data needs

\*Build your awareness of the **potentials** and **challenges!**

### Topic

What are your research questions and potential **variables**? For abstract concepts (e.g. happiness, economic freedom), look for an **index** (e.g. happiness index) for variables to consider.

### Geography

Do you need **country-level**, **national**, or **subnational** (e.g. state, county, zip code) data? Subnational data may **not** always be **available**. Monetary cross-country comparison (e.g. real GDP) may need to be **adjusted** for the purchasing power parity/market-based exchange rate.

### Time Period

Do you need **time series** data? Historic data may have **gaps**; collection methods can **change** over time (e.g. CPI); monetary time series may need to be **adjusted** for inflation. A **time lag** between data collection and its release is typical.

### Frequency

Do you need **quarterly**, **monthly** or **annual** data? Some data may be available in daily increments (e.g. stock price); some data will only be in 5 years (e.g. Economic Census). The frequency you expected may **not** be **available**. For data collected multiple times a year, **seasonally-adjusted** data (e.g. retail, air travel) may be needed.

### Granularity

Do you need **microdata** (with unit-level data/individual responses) or **summary data** (e.g. data table)? Microdata may have **restrictions** in availability and use. Publicly available microdata may only be available through **specific sources/repositories**.

### Method

Would your data be collected via a **survey/interview** (e.g. public opinion), **direct tracking** (e.g. POS scanner), **administrative reporting** (e.g. crime incidents), etc.? Consider **consistency** and **comparability** when merging datasets. Collecting data is a **costly effort**; **asking why** is important for evaluating its quality!

## Step 2: Ask who cares about the data


\*Understand layers of data sources by varied stakeholders and their **pros & cons!**

 **Government Agencies** (e.g. U.S. Census Bureau, Bureau of Labor Statistics) collect data via **various surveys** and release it as **data tables**, **data files**, **data portals**, or **reports**. Since the data sample, categories, and definition **may be different** from your understanding, **read the data collection methodology very carefully**.


 **Researchers at Academic/Research Institutions/Think Tanks** (e.g. Harvard University, Urban Institute, Pew Research Center) conduct research, collect data, and publish **reports**, **working papers**, or **journal articles**. Original data may be **restricted** to the public and may be tied to specific research agenda that may be **different** from your own.


 **Trade/Industry Associations** (e.g. American Hospital Association, Risk Management Association) may collect data from their members. Factsheets and short reports are often **free**, but detailed data are often **not free**. Data may **not** be from random sampling, so may **not** be **statistically reliable**. Be aware of **potential biases** of the data towards the association's interests.

 **Data Archive/Repository** (e.g. ICPSR, Roper ipoll, UK Data Services) provides **easy access** to research data. Free self-archiving repositories (e.g. OpenICPSR; Harvard Dataverse, GitHub, Kaggle) often do **not appraise** dataset quality. It may have privacy, confidentiality, copyright **violations**; **incomplete** metadata; **missing** documentation; or the format you need **may not** be **available**.

 **International Organizations** (e.g. World Bank, IMF, WTO) collect statistics and data from member countries and often share them for **free**. Their **working papers/reports** are often more **timely** and methodologies are more **detailed** than journal articles. Data quality **depends** on member countries' data practices and the quality of the organization's assessment or evaluation frameworks.

 **Nonprofit Organizations** (e.g. Kaiser Family Foundation, Guidestar, Kauffman foundation) invest in their mission-related data collection. Their data are valuable in **filling some gaps** in current government data and may be **totally/partially free**. However, be critical of **potential biases** that promote or further their interests.

 **Private Data Vendors** compile public or private data into a database (e.g. Bloomberg, Statista, Data-Planet; IRI) and make scattered data more **available** and **accessible**. Databases are often **expensive** but the data can still be **contaminated** by missing values, errors, inconsistencies, standardization, rounding, or selection bias. Use it as a **pointer** to find original data and **make sure to verify data accuracy**.

 **Libraries** (e.g. FHG library at WCU) provide access to some paid statistical databases from data vendors or archives. Librarians create **library guides** to help users find statistics and data and develop **data literacy**. Library guides are a helpful tool to find many free and paid data sources, but may be **incomplete** or **not up-to-date**. So, always check guides from different libraries for publicly available datasets.

## Step 3: Search through different paths

\*Be flexible and persistent is the **key** to success!

### 1. Literature

Find scholarly articles or working papers at **Google Scholar**, **EconLit**, **Business Source Complete**, **NBER Working Papers**, **SSRN**, etc. to understand your topic and variables.

### 5. Data Portal

Search data portals (e.g. **Explore Census Data**; **Canada Open Data**) since their embedded data may not be discoverable using a Google search.

### 2. Database

Data Aggregators such as **Statista** and **ProQuest Statistical Abstract** are a good place to start and find pointers to original data sources.

### 6. Microdata

Use data repositories such as **ICPSR**, **IPUMS**, **UK Data Archive**, **World Bank Microdata**, etc., or search "microdata files" online.

### 3. Library Guide

Search **library guides** on your data topic. It will save you a lot of time since these pages list multiple sources in one place. Consult several guides to build your own data source list.

### 7. Restricted Data

"Restricted" (e.g. CDC Vital Statistics county-level data) doesn't mean "inaccessible." It can be accessed via a **request-approval** process.

### 4. Online Search

Try **Google Dataset Search** or use Google advanced search **site.org**, **gov**, or **edu** to specifically locate data from trade/nonprofit organizations, government, or educational institutions.

### 8. Ask for Help

Many people are here to help you - **librarians**, **statistical agency staff**, and repository **data experts**. Just don't hesitate to ask.