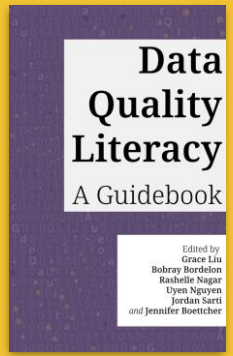
Data: Access, Quality & Sources for The National Academies of Sciences, Engineering, & Medicine. Division of Behavioral & Social Sciences and Education

September 22, 2025

Today’s agenda

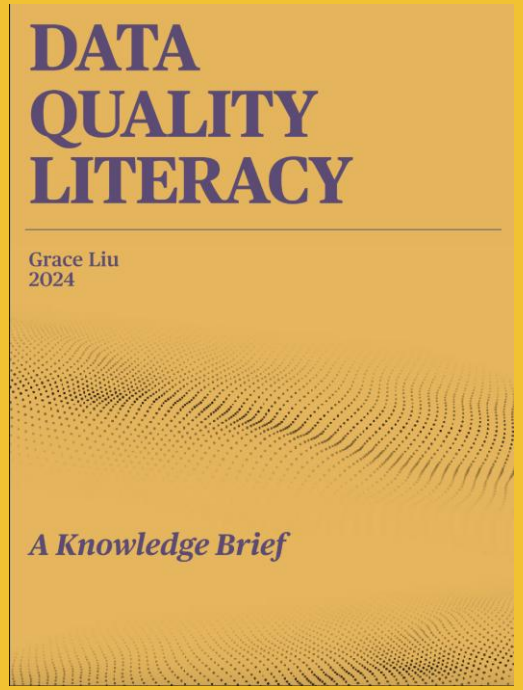
- Data quality
- Major sources
- Ethics



- Institute of Museum and Library Services (sponsored)
- National Forum Webinars Series
- Data Quality Literacy: A Guidebook
  - Competencies for Evaluating Data Quality
  - Data Documentation
  - Governmental Data (Administrative, Federal, International
  - Commercial Data
- Knowledge Briefs
- Project Website: https://www.dataqualityliteracy.org/

Knowledge Brief Overview

01 Data Reference Interview
02 Evaluating Data Documentation
03 Evaluating Dataset for Research Needs
04 Using and Evaluating U.S. Federal Statistics
05 Understanding Administrative Data
06 Evaluating Administrative Data Quality
07 Understanding Commercial Data
08 Evaluating Commercial Data Quality
09 Commercial Data Quality: Conversation with the Vendors
10 Commercial Data Quality: Conversation with Researchers
11 Evaluating International Government Data Quality
12 Understanding Survey Data and Public Poll
13 Evaluating Survey Data Quality



Evaluating Existing Datasets Involves Asking:

- Is the dataset accessible?
- Available within a reasonable timeframe?
- Include measures relevant to the question?
- Why is the data collected?
- Who does the sample represent, and who is missing from the data? Why are they missing?
- Can I compare this data to data from another source?
- Are the means, minimums, & maximums sensible? Why are some elements missing data?
- How large does the effect need to be to detect it in this sample size?

Evaluating Data Documentation:

Generally, data documentation is of higher quality if:

- The producer is an authority;
- transparent and well-documented by its creator;
- validated

Good Data Documentation Tells the Prospective User:

- Why the data is collected;
- How the data is collected, structured, and managed;
- Unit of analysis/observation (individual, household, business establishment);
- Geographic coverage and its granularity;
- Temporal Coverage (e.g., time periods covered, date of collection, or dates for different waves of collection);
- What concepts are being measured (e.g., data dictionary)?

Federal Statistics

Pros: Run by data professionals, public domain (no restrictions & free), large sample size, transparent documentation, APIs, geographic diversity, open to change

Cons: Decentralized collection & distribution, older data, few value added features, response rate decreasing

Help: fill out government surveys

Federal Statistics: Things to look out for

- **Changes in the Dataset Schema & Structure.** Data collection can also be discontinued due to funding cuts or changes in mandates. This change can affect sample size, data processing, & the comparability of data across time.
- **Inconsistent Use of Variable Codes.** Do not assume variable codes are the same between data collections. Example: 2 agencies collects income data, but the definitions of income are not the same. Caution when merging data when the variable is not standardized.
- **Changes in Survey Questions or Variable Codes or Labels between Surveys**. For example, a new survey collection instrument is used in one year but not the next. Then, the trend can become discontinuous.
- **Latency or Source Release Schedule Changes.** For example, the Internal Revenue Service (IRS) can be late in publishing an updated statistical release.

Federal Statistics: Things to look out for

- **Changes in Reference Materials.** For example, updates to North American Industry Classification System (NAICS) codes every 5 years can affect data comparability over time, potentially disrupting continuity in series.
- **Lack of Documentation.** Especially for administrative records, the data documentation may be insufficient for researchers to understand how to use the data.
- **Some Sources are not Readily Available.** Some data sources, like satisfaction surveys were collected but remain unavailable. In some cases, potentially valuable sources that could be leveraged are absent. Might have to use FOIA.

International Data (1 of 2)

- **Missing Geographies and Values**: Not every country reports every metric.
- **Temporal Limitations:** Harmonizing data across countries is time-consuming, often taking months or years. Challenges include discontinued indicators, regime changes, lack of transparency, and insufficient statistical capacity in some countries.
- **Adjustments:** Converting national currencies to US dollars and switching between constant (real) and current (nominal) prices are common tasks.
- **Changes in Methodologies:** Governments may revise their statistical methodologies due to changes in the economy or counting methods. New versions may not map well to older versions, and IGOs may not keep older data.

International Data (2 of 2)

- **Discrepancies**: International data can be published in different databases (e.g., in both the International Labor Organization statistics and UNData). When data in one system is updated or deleted, it may still exist in another.
- **Data Manipulation**: Data may be intentionally misreported by countries. Occasionally, data may be manipulated to achieve a goal or a target. For example, the World Bank's *Doing Business Report* was discontinued after accusations that it gave certain countries preferential treatment in the report's annual country rankings.
- **Concept-Measurement Gap:** There can be a gap between a concept defined in a manual & the capacity of a government or statistical authority to measure it accurately.

Evaluating Administrative Data Quality

Federal Administrative Data

- Examples: Internal Revenue Service (IRS) data, Social Security Administration (SSA) administrative records, U.S. Patent and Trademark Office patent applications, and Center for Medicare and Medicaid Services data.

State and Local Administrative Data

- Examples: Department of Motor Vehicles Driver's License data, Supplemental Nutrition Assistance Program and Temporary Assistance for Needy Families (SNAP/TANF) data, and Unemployment Insurance data.

Commercial Administrative Data

- Examples: Black Knight (master address data, mortgage data), Experian (credit bureau header data), InfoGroup (household member data), Circana (point of sale scanner data), J.D. Power (new vehicle transaction data), and D&B (business directory; credit and risk data).

Evaluating Administrative Data Quality

- When repurposing administrative data, population coverage & sampling biases (e.g., self-selection bias or survivorship bias) may be of particular concern.
- Meanings of particular data values in administrative data are likely to be different from the user's concept of interest. May not include broader variables of interest such as economic & demographic variables.
- Administrative records alone often cannot be used to address all analysis questions; for example, eligibility data doesn't provide information about nonparticipants.
- Micro-level administrative data is often difficult to access. Privacy & disclosure concerns are major constraints.
- Data cleaning & preparation can be complex, especially if trying to link administrative data with other data sources.

Evaluating Commercial Data Quality

Completeness
○ **Missing Values** (may make the database less valuable/unusable.)
○ **Header Data** (may lead to misclassification for time-series analysis.)
● **Accuracy**
○ **Data Errors** (typos, arithmetic errors, coding errors, date errors, classification errors, etc.)
○ **Biases** (sampling, under-coverage, survivorship bias, etc.)
● **Consistency**
○ **Inconsistency** (format, classification, breaks, multiple sources, international data, duplicates, etc; takes users extra efforts to clean the data.)
○ **Discrepancies** (between different sources; may lead to “database effect.”)

Data Quality Literacy: :
A Knowledge Brief

01 02 03 04 05 06 07 08 09 **10** 11 12 13

Commercial Data Quality: Conversation with Researchers

After one or more conversations with the vendors (Refer to Data Quality Literacy Series 09: Commercial Data Quality: Conversation with the Vendors), librarians can work with the researcher to examine the data quality more closely through the sample or trial. Here are some further questions to consider to assess its fitness for use:

Coverage

- **Data Fields and Variables:** Does the dataset include all relevant data fields, variables, identifiers, and classifications needed for the research?
- **Geographic and Temporal Scope:** Does the geographic and temporal coverage align with the research needs? Is the level of detail (granularity) sufficient for the analysis?

Clarity

- **Data Sources:** Is the data collected, created, or purchased? Is the methodology for data collection, creation, or aggregation valid and reliable? If the data originates from a survey, questionnaire, or form, are the questions used and detailed methodology available for review?
- **Variable Definitions:** Are all data variables and their values clearly defined and documented? When date or year is a variable, is it clear whether it refers to the date/year the data was collected or reported?
- **Actual vs. Modeled Data:** For variables with assigned values, how are these values calculated? Do the data represent actual values, or are they averaged, estimated, or projected (which is common for variables such as private companies' revenues or industry sizes)? Is the researcher aware that different sources can present widely divergent estimates?
- **Changes to Original Data:** Are there any standardization, conversion, normalization, or indexing made to the original data? Will these changes affect the data's fitness for use?

Completeness

- **Sample Size:** It is rare for a dataset to be 100% complete; it often consists of a sample rather than the entire population. Does the sample size allow for valid inferences about the population of interest? Is the sample large enough to support meaningful statistical testing?
- **Sample Representativeness:** Does the dataset include all necessary subgroups for the intended study? Can the database produce a sample that represents the broader population? Is the data a proper proxy or measurement for the phenomenon under study (e.g., data about public firms can be a poor proxy for calculating actual industry concentration)?
- **Missing Values:** Are there missing values for any variables? Is there a discernible pattern to the missing values, or are they sporadic? Can these missing values be filled through imputation or other methods? Is the proportion of missing values acceptable?
- **Header Data:** Are there data variables, such as company name, ticker symbol, stock exchange, industry code, headquarters location, or other demographic variables that only contain the latest available values (which may not be the most current)? Is the researcher aware that the header data can lead to misclassifications or mismatches in time-series and cross-sectional analyses, potentially distorting trends and comparisons by organizing data into incorrect categories?

Accuracy

- **Errors:** Are there observable data errors, such as typos, arithmetic errors, coding errors, date errors, classification errors, or outliers? Can the data be cross-checked or spot-checked against other sources? Is the researcher aware of the “database effect,” where using different datasets can yield different research results?
- **Biases:** Are certain groups excluded, underrepresented, or not timely included in the datasets? Is the researcher aware of potential biases, such as sampling bias, under-coverage bias, survivorship bias, or look-ahead bias? Can procedures be applied to mitigate these biases?

Evaluating Commercial Data Quality

Clarity

○ **Actual vs. Estimated Data** (be aware that sources often present widely divergent numbers in their estimates.)

○ **Standardization** (may lead to inaccuracies in certain prediction models.)

○ **Superseded Data** (may lead to different data when downloading at different time.)

○ **Reporting Time** (Improperly recorded reporting time can lead to look-ahead bias or selection bias.)

● **Misuse of Data**

○ **Improper Proxy** (may lead to unreliable research results.)

Data Quality Literacy: :
A Knowledge Brief

01 02 03 04 05 06 07 08 09 **10** 11 12 13

Commercial Data Quality: Conversation with Researchers

After one or more conversations with the vendors (Refer to Data Quality Literacy Series 09: Commercial Data Quality: Conversation with the Vendors), librarians can work with the researcher to examine the data quality more closely through the sample or trial. Here are some further questions to consider to assess its fitness for use:

Coverage

- **Data Fields and Variables:** Does the dataset include all relevant data fields, variables, identifiers, and classifications needed for the research?
- **Geographic and Temporal Scope:** Does the geographic and temporal coverage align with the research needs? Is the level of detail (granularity) sufficient for the analysis?

Clarity

- **Data Sources:** Is the data collected, created, or purchased? Is the methodology for data collection, creation, or aggregation valid and reliable? If the data originates from a survey, questionnaire, or form, are the questions used and detailed methodology available for review?
- **Variable Definitions:** Are all data variables and their values clearly defined and documented? When date or year is a variable, is it clear whether it refers to the date/year the data was collected or reported?
- **Actual vs. Modeled Data:** For variables with assigned values, how are these values calculated? Do the data represent actual values, or are they averaged, estimated, or projected (which is common for variables such as private companies' revenues or industry sizes)? Is the researcher aware that different sources can present widely divergent estimates?
- **Changes to Original Data:** Are there any standardization, conversion, normalization, or indexing made to the original data? Will these changes affect the data's fitness for use?

Completeness

- **Sample Size:** It is rare for a dataset to be 100% complete; it often consists of a sample rather than the entire population. Does the sample size allow for valid inferences about the population of interest? Is the sample large enough to support meaningful statistical testing?
- **Sample Representativeness:** Does the dataset include all necessary subgroups for the intended study? Can the database produce a sample that represents the broader population? Is the data a proper proxy or measurement for the phenomenon under study (e.g., data about public firms can be a poor proxy for calculating actual industry concentration)?
- **Missing Values:** Are there missing values for any variables? Is there a discernible pattern to the missing values, or are they sporadic? Can these missing values be filled through imputation or other methods? Is the proportion of missing values acceptable?
- **Header Data:** Are there data variables, such as company name, ticker symbol, stock exchange, industry code, headquarters location, or other demographic variables that only contain the latest available values (which may not be the most current)? Is the researcher aware that the header data can lead to misclassifications or mismatches in time-series and cross-sectional analyses, potentially distorting trends and comparisons by organizing data into incorrect categories?

Accuracy

- **Errors:** Are there observable data errors, such as typos, arithmetic errors, coding errors, date errors, classification errors, or outliers? Can the data be cross-checked or spot-checked against other sources? Is the researcher aware of the "database effect," where using different datasets can yield different research results?
- **Biases:** Are certain groups excluded, underrepresented, or not timely included in the datasets? Is the researcher aware of potential biases, such as sampling bias, under-coverage bias, survivorship bias, or look-ahead bias? Can procedures be applied to mitigate these biases?

Major Sources

Major Sources: Examples of Free International

- https://www.ipums.org/ IPUMS provides census & survey data from around the world integrated across time & space. IPUMS integration & documentation makes it easy to study change, conduct comparative research, merge information across data types, and analyze individuals within family & community contexts.
- https://ghdx.healthdata.org/ Catalog of surveys, censuses, vital statistics, & other health-related data.
- https://microdata.worldbank.org/ Facilitates access to microdata that provide information about people living in developing countries, their institutions, their environment, their communities & the operation of their economies.

Major Sources: Examples – Free!

- https://fred.stlouisfed.org/ . Federal Reserve Economic Data, FRED. Over 841,000 economic data time series from national, international, public, & private sources.
- https://www.usa.gov/ Open data from the federal government.
- https://www.data-archive.ac.uk/ (United Kingdom)
- https://www.gesis.org/en/home (Germany)
- https://www150.statcan.gc.ca/n1/en/type/data (Canada)
- https://ada.edu.au/ (Australia)
- https://proyectos.inei.gob.pe/iinei/srienaho/index.htm (Peru)

Major Sources: Membership

https://www.icpsr International consortium of 800+ academic & research bodies. Provides leadership & training in data access, curation, & analytical methods for social science. Collaborating with funders, including federal US agencies, ICPSR develops themed data collections and projects. Established the Summer Program in Quantitative Methods in 1963, educating tens of thousands of researchers in introductory to advanced training in statistics, data analysis, and quantitative research methods.

- **American Economic Association Data and Code Repository**
- **American Educational Research Association Data Repository**
- **Health and Medical Care Archive (HMCA) (Robert Wood Johnson)**
- **National Archive of Data on Arts & Culture**
- **National Archive of Computerized Data on Aging**
- **National Archive of Criminal Justice Data**
- **Resource Center for Minority Data**
- **Social Media Archive**
- **ResearchDataGov (**web portal for discovering and requesting access to restricted microdata from federal statistical agencies.)

Major Sources: Additional

- https://ropercenter.cornell.edu/ipoll/ World's largest collection of public opinion. Subscription.
- https://www.datarescueproject.org/ Started in February 2025 as a coordinated effort of 3 data organizations, including members of IASSIST, RDAP, & the Data Curation Network to serve as a clearinghouse for data rescue-related efforts & data access points for public US governmental data that are currently at risk

Data Rescue Project

- Data and Statistical Services (Princeton catalog) Use to discover sources.

Princeton University LIBRARY

Data Ethics

- Personally Identifiable Information: https://csrc.nist.gov/glossary/term/personally_identifiable_information
- Human Subjects Research & Institutional Review Board (IRB) (National Academies)
- Independent Institutional Review Boards

Q&A

Bobray Bordelon
bordelon@Princeton.edu