Evaluating Dataset for Research Needs

Data quality depends on the fit between the dataset and the research question. A high-quality dataset allows researchers to answer the question of interest.

General Dataset Features that Indicate Data Quality

Dataset features that contribute to quality include:

- data elements that are relevant to the research question
- sample that allows inferences about the population of interest
- data provenance or documented trace of origin is understood
- data can be cross-checked against other sources
- the sample size is large enough to allow for meaningful statistical testing

Think Critically about Existing Datasets

Thinking critically about existing datasets involves asking:

- Is the dataset accessible to the researcher?
- Is the dataset available within a reasonable timeframe?
- Does the data include measures relevant to the research 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, and maximums sensible? Why are some elements missing data?
- How large does the effect need to be to detect it in this sample size?

Questions to Ask When Combining Datasets

When combining datasets, ask a much richer array of questions, including:

- Are there common identifiers between both datasets? The most ideal identifiers are those less likely to change over time such as Employee Identification Numbers (EINs), Central Index Keys (CIKs), or geographic codes such as Federal Information Processing Standards (FIPS) county codes or International Organization for Standardization (ISO) 3166 Codes for the representation of names of countries and their subdivisions. In contrast, a company name or ticker symbol may not be ideal since they may change due to restructurings or name changes. A geographic name may be formatted differently depending on the source.
- What is the smallest unit for which data can be combined and what are the implications for aggregation?
- Who does the combined dataset represent? What are the implications for sample size and representativeness?
- Are there differences in the definition of the same variables (e.g. "Year" can be defined differently in different sources)?

References

Kalinowski, A., & Hines, T. (2020). Eight Things to Know about Business Research Data. *Journal of Business & Finance Librarianship, 25*(3/4), 105–122. https://doi.org/10.1080/08963568.2020.1847548

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

The Knowledge Brief is compiled by Grace Liu, based on the National Forum presentation from Janet Currie and Barbara Esty, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Visit the project website to learn more!
https://www.dataqualityliteracy.org

INSTITUTE of Museum and Library SERVICES