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?

- **Up-to-Dateness:** Are all values in the dataset current and reflective of the most recent information, or is there a mix of up-to-date and outdated values? Is the update process periodic or synchronized?
- **Revisions:** Are there revisions due to error corrections, restatements, or other changes? Are these revisions done on a regular schedule? Is the researcher aware that this can cause data downloaded at different times to have different values, potentially leading to different results?

Consistencies

- **Format:** Is the format of the data variables (e.g. name, address, units of measurement) consistent? If not, is it costly to standardize and clean the data?
- **Classification:** Is a classification system (e.g., NAICS) used to categorize the data? Is the application of classification codes and their level of granularity consistent? When the classification system is updated (e.g., NAICS 2017 to 2022), has reclassification been consistently applied over time? Is the researcher aware of potential issues in comparing, aggregating, or merging data caused by classification inconsistencies?
- **Breaks:** Are there changes in data sources, data availability, collection methods, levels of measurement, geographic boundaries, regulatory policies, or other reasons that affect consistency over time?
- **Multiple Sources:** For data collected or aggregated from multiple sources (e.g., point-of-sales data from different stores) or geographic areas, are there differences in data availability, collection methods, or levels of measurement across these sources and areas?
- **International Data:** What data is available for international coverage? How is the data translated? How are measurement units (e.g., currency, weights, lengths) converted? How is standardization implemented? Is the translation, conversion, and standardization consistent across geographic areas and over time?
- **Duplicate Records:** Are there duplicate records in the dataset? If so, is the data value for these duplicates consistent? Do these duplicates indicate larger issues, such as problems with data entry, dataset merging, or system flaws?

If the librarian and researcher have identified data quality issues, but the database is still the best available option on the market, these quality concerns can be factored into the price negotiation.

References

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, inspired by 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].