Commercial Data Quality: Conversation with the Vendors

Commercial Data Quality: Questions to Ask

Commercial data vendors may be hesitant to share detailed documentation of their data. However, they are often willing to engage in discussions with data users to answer specific questions, provide clarity, or offer additional context regarding their proprietary methods. It is crucial for librarians or researchers to probe this "**black box**" as much as possible. Here are some questions that can help broadly assess the data quality and fitness for use:

- **Content:** What data fields, variables, identifiers, or classifications (e.g., North American Industry Classification System (NAICS)) are included? Check if the desired fields or variables are available. If the dataset needs to merge with other datasets, make sure there are common identifiers.
- **Geographic Coverage:** What is the geographic coverage of the dataset, its granularity, and how complete is the coverage? Be cautious of datasets claiming global coverage that might only include data for 10 countries, or those claiming county-level data that might only have detailed data for 20 states.
- **Temporal Coverage:** What is the temporal coverage of the dataset, its frequency, and how complete is the coverage? Be aware that while the dataset may have monthly data starting in 1995, the data might not be complete before 2000, or may only have annual data available for the earlier years.
- **Data Sources:** How does the vendor gather the data? The data may be obtained through one or a combination of the methods, including (1) collected data: obtained from public sources such as administrative records, surveys, public data, or web scraping; (2) created data: derived or modeled using proprietary algorithms or in-house analysts; (3) third-party data: purchased from other firms or vendors.
- **Methodology:** What is the data collection methodology, and is there any standardization or conversion of the original data? Is the data normalized or indexed to a specific baseline? How is this done? Is there a crosswalk to map data between different coding systems? Ask about the availability of detailed documentation, data dictionaries, file setup documents, or access to vendor experts who can answer questions. At a minimum, the vendor should provide a list of variables and their definitions.
- **Sample Size:** What is the number of records or observations included in the dataset, and how does this compare to the entire population? For example, in a public company dataset, what percentage of all public companies are included? Does the dataset also cover delisted firms, foreign firms listed on US stock exchanges, or US firms listed overseas?
- **Changes in Data:** Are there any changes to the datasets, such as changes in collection methods, classification, or shifts in coverage focus over time? How frequently is the data updated, and what is the time lag between data collection and its inclusion in the database? Is there a process to identify data errors and make corrections? Request documentation on gaps in coverage, change logs, imputations, cleaning, instruments, and any changes in collection methods, variable definitions, or geographic boundaries.
- **Data Product:** Is the data product a database, an interface, or a dataset? Web interfaces usually include the latest data and may not allow extensive data extraction. The cost of extracts or backfiles can be high.
- **Data Format:** What format is the data delivered? Is it structured or unstructured? Can it be delivered in common file formats (e.g., comma-separated values (CSV) file)? Is the format compatible with statistical software, or does it require additional cleanup?
- **Delivery Frequency:** How frequently will the data be delivered? Is it through daily, weekly, or monthly data feeds, or can it be provided as a one-time extract or annual subscription? Can updates be purchased on a scheduled basis or as one-time purchases?
- **Restrictions:** Does the data contain restricted variables? Researchers might need extra permissions to access restricted data such as health, labor, or social surveys that capture personally identifiable information. Does the vendor require

the submission of findings or full papers before publication? Do they retain the right to deny data usage in future publications? Can the data be shared for replication studies? How does the data need to be cited or attributed? Are there limitations on the number of records downloaded or types of analysis allowed? Some vendors prohibit machine learning or artificial intelligence applications on their data.

- **Access and Confidentiality:** Who can access the data, and what are the required processes to access the data? Besides the researcher, is it accessible to all campus affiliates, faculty, a specific department, or co-authors at other institutions? Does the project need an Institutional Review Board (IRB) review, or does the researcher need to sign a Data Use Agreement (DUA) or Non-Disclosure Agreement (NDA)? Consider the data's sensitivity, confidentiality, and required security measures, as well as any restrictions on data storage.
- **Resources Required:** How much time, money, expertise, and storage space are needed to acquire and make the data available to the researcher?
- **Trial or Samples:** Can the vendor provide a specific sample or set up a trial before purchase? Ensure the trial reflects actual database functionality, including download capabilities, exporting, searching, and printing.
- **Support:** What ongoing support is available for researchers regarding data errors, clarification, functionality, and methodology?

Issues Around Third-Party Data

Commercial data providers often provide third-party data on their platforms, but instead of giving direct access to third-party data, it may be integrated into their platform in different ways:

- **Input to their Models:** They may use third-party data as input into their proprietary models.
- **Proprietary Identifier/Rating:** They may append their data with proprietary identifiers or ratings from another company.
- **Convenience Data (Demographics Add-On):** They might enhance their data with third-party convenience data such as demographic data.
- **Aggregator (Tools for Automation):** The data provider may be an aggregator and provide tools for automation and all the data are from third parties.

When dealing with commercial data that includes third-party sources, here are some questions to ask:

- **Third-Party Source:** What is the third-party source, and how does the third party collect or create their data?
- **Licensing Rights:** How long are the licensing rights available to the source provider? Commercial data providers can gain or lose access to other third-party data or change to another data provider, which can alter the resource's methods and affect data quality.
- **Use Limitations:** Are there any limitations in using third-party data? Sometimes, the third-party data is view-only or cannot be easily exported or exported at all, so it would be hard to use for further analysis.
- **Enhancements:** What, if any, enhancements are made? Examples include adding demographic information from the Census, rounding, and modeling or forecasting data until observed. It is important to know what enhancements are made and ensure any alterations to the original data are transparent to the data users.
- **Vendor Consolidation:** Are there instances of data vendor consolidations, mergers, or acquisitions? If so, ask how data will be incorporated into the other product, if all data will come over, the timelines, and any changes to product specialists from the previous provider, methodology, or delivery and access method. Although we may not be able to do anything about a new and more tedious workflow, that is something to keep in mind as we evaluate those products.

Additional Questions on Competitive Products

To better understand the competitive landscape and the vendor's position in the market and open up new options to compare and further negotiate with the vendor, consider asking the following questions:

- What other sources do you compare your products to?
- Who are your competitors in the market?
- How does the data compare to another familiar or well-known source?

Vendors are usually quick to demonstrate how their data is different, highlight where they believe they excel, and are often upfront about areas where their data may not compare as favorably. This can also introduce you to other sources you had not considered or were unaware of.

After gathering this information and requesting a sample or trial, you are ready to investigate and discuss the data quality with the researcher. For further details, refer to **Data Quality Literacy Series 10: Commercial Data Quality: Conversation with the Researchers**.

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, inspired by the National Forum presentation from Stephanie Tully 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].