Evaluating Data Documentation

Data Documentation

Data documentation provides the contextual information needed to discover, understand, access, and reuse data. Examples include README files, metadata, data dictionaries, codebooks, methodologies, survey instruments, and lab records.

Indicators of Documentation Quality

Generally, data documentation is of higher quality if:

- it is produced for use by researchers;
- it is transparent and well-documented by its creator;
- it is validated via a peer review process; and
- time and expertise are invested to curate it by experts.

Characteristics of Good Data Documentation

Good data documentation tells the prospective user:

- **Why** the data is collected (e.g., project context or descriptions);
- **How** the data is collected, structured, and managed (e.g., methodology, study design, sample, universe/population, questionnaires, restrictions, revision history);
- **Who (or what)** the unit of analysis/observation is (e.g., individual, household, business establishment);
- **Where** is being covered and its granularity (e.g., geographic coverage and smallest geographic unit);
- **When** is covered (e.g., time periods covered, date of collection, or dates for different waves of collection); and
- **What** concepts are being measured (e.g., variable data dictionary)?

Fill Data Documentation Gaps

If you obtain a dataset without documentation or if the documentation is confusing or missing information, consider the following:

- **Who collects the data?** Is it a commercial vendor, government/organization, or individual researcher? You can research and connect with the potential data producers to request the details.
- **Who distributes the data?** Are they libraries or data archives/repositories, commercial/independent organizations, or other researchers? You may check your license, or tap into your network of other libraries or repositories that may purchase or distribute the data.
- **Who uses the data?** Search for published articles, working papers, or unpublished works to find who uses the data. Connect with researchers to work out gaps in documentation. If many people use it, it is often another indicator of good data quality.

Comparing Documentation with Datasets

After evaluating the critical information included in the data documentation, a further step is to compare the documentation with the datasets. This helps identify any undocumented variables, out-of-range codes, unexplained missing data, illogical skip patterns, or other discrepancies between the data and its documentation.

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, based on the National Forum presentation from Ron Nakao 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].