Perceived Documentation Quality for Social Science Data

Jinfang Niu & Margaret Hedstrom

School of Information, University of Michigan
Outline

- Secondary data analysis & Documentation
- A larger research project
- Previous work
- Findings
- Future work
Secondary data analysis

The analysis of data for a different purpose than what the data were originally collected for, possibly by the original data producers themselves, or in collaboration with other people, or by entirely different people.
Documentation

Definition

Knowledge about data that is recorded and transferred to secondary users.

Examples

Codebooks, project reports, data collection instruments, previous publications, user guides or handbooks, statistical manual, data extraction software, IRB materials, workflows
Based on cataloging information

Consistency between data and documentation

Based on understanding of data (obtained from documentation)
A larger research project

- Identify impacting factors of user Perceived Documentation Quality (PDQ)

- Study the effect of PDQ on secondary data use
  - Impact on users’ incentive to use secondary data?
  - How do users overcome inadequate documentation?
Previous work

- A Documentation Evaluation Model was constructed
- Possible impacting factors identified & Hypotheses formulated
- Data collected
Documentation Evaluation Model: How it was constructed?

- Document Quality Indicators (DQI)
- Technology Acceptance Model (TAM)
- Data Documentation Initiative (DDI)
Documentation Evaluation Model: what it looks like?

Perceived Quality

Ease-of-use
- Learn
- Find
- Content

Sufficiency
- Complete
- General perception
Impacting Factors

1. Producers’ Incentive & Capacity
2. Existence of intermediaries
3. Vulnerability to tacit knowledge
4. Users’ absorptive capacity

Perceived Documentation Quality
Exploratory interviews

- Become familiar with secondary data user & users, help decide survey sample and units of analysis, and create metrics for absorptive capacity.

Survey

- Population: people who use secondary data to conduct social science research
- Units of analysis: the most recent use case of single datasets.
Findings

- Checking the validity and reliability of DEM
- Testing the effects of the impacting factors identified
Reliability and Validity of the Model

Alpha for ease-of-use : 0.95

Hard to find
  - hard copy only; information dispersed; no cross-references between various parts; unorganized and overwhelming

Hard to understand
  - too tersely written; terminology not clear; scanned codebooks blurry and difficult to read

The item “difficulties in learning to use documentation” should be dropped from the model.
  - Alpha for the remaining 3 items: 0.94
Reliability and Validity of the Model

Alpha for sufficiency: 0.83.

Completeness: users complained about the absence of certain elements or incomplete descriptions.

The item “with documentation, I did not need to seek additional information to use the data.” was not a good indicator of sufficiency.
Reliability and Validity of the Model

Accuracy

Errors detected based on the inconsistencies between data and documentation.

Not included in DEM because:
- Consistency is very closely related to data.
- Hard for secondary users to detect errors in documentation besides inconsistency.

Accuracy needs to be included to evaluate the quality of both data and documentation.
Effects of producers’ incentive

Documentation of data produced for sharing is more sufficient* and easier to use* than data produced for self-use.

*: p<0.01, **: p<0.05 ***: p<0.10
Effects of Intermediaries

- Documentation of data produced for sharing and distributed by intermediaries are more sufficient** and easier to use** than data produced for sharing and distributed by data producers
Effects of Vulnerability to tacit knowledge

- Documentation for quantitative data is more sufficient*** and easier to use*** than documentation for qualitative data.

- Documentation for survey and census data is more sufficient* and easier to use** than administrative records and interview data.
Effects of absorptive capacity

Professors perceive the documentation they use as more sufficient* and easier to use*** than students.

Users familiar with the topics of the data perceive the documentation they use as more sufficient* and easier to use* than users not familiar with the topics of the data.

Users experienced in using the same data perceive the documentation they use as more sufficient* and easier to use* than users not experienced in using the same data.
Effects of absorptive capacity

- Users experienced in secondary data analysis perceive the documentation they use as more sufficient* and easier to use* than users not experienced in secondary data analysis.

- Users more experienced in collecting and analyzing self-collected data perceive the documentation they use as more sufficient* than users not experienced in collecting and analyzing self-collected data.
Conclusions

- Perceived documentation quality includes three aspects
  - Ease-of-use & Sufficiency & Accuracy

- DEM is reliable and valid in general with several exceptions

- Perceived documentation quality is affected by four factors:
  - Producers’ incentives
  - Existence of intermediaries
  - Vulnerability to the tacit knowledge problem.
  - Users’ absorptive capacity
Future Work

• Identify impacting factors of user Perceived Documentation Quality (PDQ)

• Effect of PDQ on secondary data use
  • Impact on users’ incentive to use secondary data
  • How do users overcome inadequate documentation?
Thanks!

NSF Award # IIS 0456022
Rackham Research grant for dissertation
Rackham one-term dissertation writing award