Transparency and Data Quality

Dan Gillman
Information Scientist
US Bureau of Labor Statistics
FCSM Conference
25 October 2022

Transparency Report

- CNSTAT Panel on Transparency Reproducibility in Federal Statistics
  - Began April 2019
  - Report issued November 2021
- Report: Transparency in Statistical Information for NCSES and All Federal Statistical Agencies
- Report organized into 7 chapters and 2 appendices
  - Metadata, Standards
    - Chapter 5, Appendix A, and Appendix B

Transparency

- In the report, transparency is defined as
  - provision of sufficiently detailed documentation of all the processes of producing official estimates.
- Goal of transparency is to
  - enable consumers of federal statistics to accurately understand and evaluate how estimates are generated
- From this, there is need for documentation
- Documentation and metadata
  - 2 sides of the same coin

Metadata

- Data used to describe some resource(s)
  - Role for data, not a kind
- Same as documentation, only more formal
  - Documentation – typically in text form
    - Word, PDF, HTML documents
  - Metadata – typically in a database (repository)
    - RDBMS (relational), XML (hierarchical), RDF (graph)
- Not all documentation can be formalized
  - Rationales – reasoning supporting some decision

Metadata Schema

- Organized by a schema
  - Framework for structuring and organizing
  - Similar to a model
  - Contains bins (elements) for entering metadata
- Schema is a template for metadata
- Filled in schema is an instance

Metadata Schema and Instance Example

- Describe variables using metadata schema

- Name
- Meaning
- Universe
- Datatype (intended)
- Allowed values

- NAICS code
- Industry classification, 6 digits
- All mines in the US
- Nominal (categories, no order)
  - 21120 Crude oil extraction
  - 21130 Natural gas extraction
  - …

Technical Specifications

- Schema is a kind of technical specification
- Formalized set of requirements
- Conform to specification
  - Satisfy all requirements

- Necessary condition for transparency
  - Conformance to a metadata specification

Data Quality

- Transparency is a characteristic of quality
  - Transparency => sufficient documentation
  - Documentation (metadata) provides information to understand
    - Data and structures
    - Processes (data acquisition, editing, etc.)
    - Designs and methodologies
  - Provide a level of quality to data and their production
  - Provide the ability to assess quality

Provide a Level of Quality

- Data not understandable or interpretable => low quality
  - It is hard to use them
- Understanding entails many things
  - Meaning and allowable values for variables
  - Wording, order, and response choices for questions
  - Consequences of sample design
  - Editing and allocation procedures
- This is a role of documentation (metadata)
- Transparency => the necessary documentation is available

Provide Ability to Assess Quality

- Data quality considerations include
  - Are all reported values for each variable valid?
    - E.g., an age reported as 135 years
  - Are they accurate?
    - E.g., was the right NAICS code assigned to a business establishment?
  - Are they coherent?
    - E.g., do population estimates agree with other sources?

Provide Ability to Assess Quality

- Data quality considerations include
  - Are they consistent?
    - E.g., biological males reporting being pregnant
    - E.g., biological females reporting having prostate cancer
  - Are they timely?
    - Do the data represent the state of the current population or economy?
  - Are they useful?
    - Do they answer questions the public want to know?
- Metadata and documentation provide the answers

Metadata Quality

- All this works if the documentation (metadata) are good
  - Where good means “high quality”
- What does it mean to have quality metadata?
- Schema instance => declarative sentences
- From earlier example:
  - Name of the variable is “NAICS code”
  - Datatype of the variable is nominal
  - …
- The combination of declarative sentences is documentation

Metadata Quality

- Questions about these sentences:
  - 1) Do the instance values have the right format — Syntax
  - 2) Are the instance values true — Semantics
  - 3) Is there an important element left out? — Pragmatics
  - 4) Are there any irrelevant elements? — Pragmatics
- Gillman, D., *Achieving Transparency – A Metadata Perspective*
- Data-Intelligence, Special edition on metadata – To appear

Pathway to Metadata Quality

- Need to choose relevant schemas
- Metadata standards are a good source
- Look for standards development process that is
  - consensus, open, balanced, fair, and inspectable
- Several sources in statistical community
- Data Documentation Initiative

Statistical Metadata Standard:
Data Documentation Initiative (DDI)

- Managed under DDI Alliance at ICPSR
- Suite of metadata standards for social and behavioral science data
- Codebook (2000), Lifecycle (2008), Cross-Domain Integration (late 2022)
- All have an XML implementable representation
- Lifecycle was built with statistical agencies in mind

DDI Lifecycle Standard

- Supports survey lifecycle
  - Based on UNECE Generic Statistical Business Process Model
  - Like UNECE Generic Statistical Information Model
- Except
  - GSIM is a conceptual model
  - DDI Lifecycle based on XML – immediately implementable
- Supports reuse and linkages of metadata
  - Across surveys, revisions, and time
  - Many ways to group and organize metadata

DDI Lifecycle Standard

- Many statistical organizations around the world
  - Australian Bureau of Statistics
  - BLS
  - INSEE (France)
  - ISTAT (Italy)
  - Statistics Canada
  - Statistics Netherlands
  - Statistics New Zealand
  - Many others, including universities and data archives

Questions

Contact Information

Dan Gillman
Office of Survey Methods Research
www.bls.gov/osmr
202-691-7523
Gillman.Daniel@bls.gov