Transparency and UNECE Metadata Standards

Dan Gillman
US Bureau of Labor Statistics
Office of Survey Methods Research
Modern Stats World Workshop
June 27, 2022

Outline

- Transparency panel report
- Conditions for transparency
  - Metadata schemas and instances
  - Conformance
  - Metadata quality
  - Usability

CNSTAT Panel on Transparency and Reproducibility in Federal Statistics

- Panel approved April 2019
- Sponsor agency: NSF/NCSES
- 15 panel members
  - US statistical agencies
    - Including Dan Gillman (US BLS)
  - International agencies
    - Including David Barraclough (OECD)
  - Academia, Archives, Consultants

Work of the Panel

- Periodic 2-day meetings
  - Day 1 – fact-finding with invited speakers
  - Day 2 – internal deliberations
- Covid interfered with schedule
  - No face-to-face meetings after February 2020
  - Drafting complete document was slowed
- 10-member review panel
  - Produced many comments
  - Comment resolution was time-consuming

Report

- Official Title of Report

  Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies

- Report issued November 2021
- Focus on transparency
- Divided into
  - Summary
  - 7 Chapters + 2 (substantive) Appendices
- Link to Report - https://www.nationalacademies.org/our-work/transparency-and-reproducibility-of-federal-statistics-for-the-national-center-for-science-and-engineering-statistics

Relevance to MSW Standards

- Chapter 5 – Metadata and Standards
  - Detailed description of metadata
  - Return on Investment
    - for metadata management and systems
  - Rationale for adopting standards
    - Includes argument for joining UNECE efforts
- Co-authored by:
    - David Barraclough, Dan Gillman

Relevance to MSW Standards

- Appendix A –
  - Statistical Metadata Standards – in detail
  - Description of UNECE, DDI, and SDMX standards
    - UNECE: GSBPM, GSIM, CSPA, CSDA
    - DDI: Codebook, Lifecycle, CDI, SDTL, XKOS, others
    - SDMX: SDMX, VTL
    - Other standards: DCMI, DCAT, PROV, ISO 19115, others
- Co-authored by:
  - David Barraclough, Dan Gillman, Michael Lenard, Andrea Thomer

How to Read the Report

- The report is long - 178 pages
  - From TOC to end of Appendix B
- For quicker and less technical read
  - Summary, Chapters 1 and 7
- All recommendations & conclusions in Summary
- Each chapter has its own recommendations
  - More contextualized, better understanding

Decisions

- Provide recommendations in each chapter
  - Devote chapter 6 to specifics for NCSES
- Definition of transparency:

  provision of sufficiently detailed documentation of all the processes of producing official estimates

- Focus on documentation -> need for metadata
- Reduce emphasis on reproducibility
  - Transparency is a pre-condition

Documentation

- Needed to find, understand, and use
  - Data
  - Methodologies
  - Processes (designs, algorithms, code)
- Documentation and Metadata
  - Generally, synonymous
  - Often
    - Documentation refers to textual explanations
    - Metadata is a more formalized way of explaining

Documentation

- Formal metadata conundrum
  - Textual descriptions “tell a story”
  - Formal metadata attempts the same thing
  - The information obtained from metadata
    - Must be at least as informative as text
    - Organized metadata can do more
      - E.g., comparability over time and studies
  - Hard to subdivide each kind of description
    - Consider descriptions for variables versus for rationales

Metadata

- In the formal case, metadata
  - Set of descriptors for a kind of objects
    - E.g., variables, questions
- What descriptors needed for variables?
  - Example
    - Name
    - Universe
    - Allowed values
    - Datatype
    - Related data sets
    - Related concept

Metadata

- What descriptors needed for questions?
  - Example
    - Name
    - Universe
    - Response choices

Question text

Previous question(s)

Following question(s)

Metadata Schema

- Set of descriptors = Schema
- Each descriptor = schema element
- Schema formalized by
  - Specific rules for
    - Element values (formats, etc.)
    - Relationships among elements
    - Optionality / Cardinality for elements or relationships
- Schema = kind of technical specification

Schema Instance

- Set of values corresponding to schema elements
  - Called a schema instance
  - Example of variable schema instance
    - Name — marital_status
    - Universe — adults
    - Allowed values — <S, single>, <M, married>
    - Datatype — nominal
    - Related datasets — CPS, NLS, CE, ACS, SIPP, others
    - Related concept — “legal marital state”

Schema Instance

- Example of question schema instance
  - Name — marital_status
  - Universe — adults
  - Response choices — Single, Married
  - Question text — What is your current marital status?
  - Previous question(s) — ?
  - Subsequent question(s) — Were you married previously?

Transparency

- Transparency depends on documentation
  - Could be provided as formal metadata
- What makes a variable or question transparent?
  - Have necessary metadata to support required needs
  - Necessary metadata expressed through
    - Kind of object: Schema
    - Specific object: Instance of the schema
- Schema instance = metadata for an object

Conformance

- Question –
  - How do we know an instance follows the rules?
- Schema is a technical (formal) specification
  - Contains requirements and other conditions
- Conformance to a technical specification
  - Satisfy all requirements
- An instance conforms to a schema
  - If the instance satisfies all requirements in schema

Conformance

- This does not say the values are correct
  - Only that they follow formatting rules
- This does not say the elements are effective
  - Schema might have missing elements
  - Schema might have irrelevant elements
- Conformance is only about requirements
  - Found in the technical specification

Transparency

- Necessary condition for transparency
  - Conformance to a schema
- Is this enough? Is this sufficient?
- No. Why?
- How good are the metadata?
  - They can follow all the requirements
  - But do they describe an object of interest well?

Metadata Quality

- Do instance values follow formatting rules?
  - Syntax
  - Formats, obligations, cardinality, relationships
- Are all instance values true?
  - Semantic
  - Formal truth theory
  - Follow Tarski's notion of truth in a formal theory

Metadata Quality

- Semantics continued
  - Formal statement “variable name is marital status”
    - Is true, if and only if
  - The name of the variable is “marital status”

- Now, consider all schema element / instance values
- Does combination tell the right story?
  - Pragmatics
  - Schema elements might be missing / irrelevant

Metadata Quality

- Operationalizing this – in 4 steps
- #1 Conformance - syntax
  - Instances must conform to a schema
- # 2 Truth - semantics
  - Is each schema / instance value combination true?
  - For example, for variables
    - Is the name of a variable the right one?
    - Is the assigned datatype appropriate?

Metadata Quality

- #3 The whole truth - pragmatics
  - Is the story incomplete?
  - Does the schema need more elements?
  - Is there some necessary information left out of the schema?
- #4 Nothing but the truth - pragmatics
  - Is the story confusing?
  - Does the schema include misleading elements?
    - For variables, don't include
      - Unnecessary: Number of letters in name of variable
      - Irrelevant: Current population of United States

Transparency

- Another necessary condition for transparency
  - Metadata quality
- Are there more conditions?
  - Yes.
- How good is the user/system interface?
  - Can the user get the system to return desired information?
  - Usability

Usability

- Usability:
  - the quality of users' experiences when interacting with systems
- 2 main usability concerns for transparency:
  - Interface design
    - Can the user make sense of the what's on the screen?
  - Available functions
    - Are required functions available through the interface

Usability

- User interface
  - Usual usability concerns:
    - Colors — Button placement and function
    - Clearly stated instructions
- Functions: Discovery and Understanding
  - Require metadata and schemas
  - Both input and output
  - Metadata must conform to schemas
    - Discovery — input metadata
    - Understanding — output metadata

Conclusion

- Transparency requires
  - Metadata
  - Schemas
  - Conformance to the schema
  - Metadata quality
  - Usable system interface
- Claim
  - These requirements are sufficient

Any questions?

Contact Information

Dan Gillman
Office of Survey Methods Research
www.bls.gov/osmr
(w) 202-691-7523
(c) 410-624-9582
Gillman.Daniel@bls.gov

Usability

- Discovery
  - Open world assumption
    - Can't find an object just means you couldn’t find it
    - Possible search criteria are not known in advance
  - Closed world assumption
    - Every object can be found through search
    - All search criteria are known in advance
  - Use of controlled vocabularies,
    - Not user defined keywords
    - Provides exact set of values in metadata instances