# Highlights of CNSTAT Report on Transparency in Statistical Information :

## The Use of Metadata Standards and Tools for Greater Transparency of Official Statistics

**Dan Gillman**

*US Bureau of Labor Statistics*

FCSM 2022

Session G2

October 27, 2022

# Transparency

- In the report, transparency is defined as
  - Transparency is the provision of sufficiently detailed documentation of all the processes of producing official estimates.
  - The 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

# Technical Specifications

- Schema is a kind of technical specification
- Formalized set of requirements
- Conform to specification
  - Satisfy all requirements
- Standards are examples
  - Technical specifications developed under Open, Fair, Balanced, Transparent, Consensus Process

# Standards

- Metadata Standards
  - Technical specifications
  - Define how metadata are organized, usually with a schema
  - Systems designed to implement standards
    - Achieve conformance by satisfying requirements
    - Guarantees enough metadata is available
  - Transparency, necessary condition
- Many metadata standards in statistics
  - DDI, SDMX, GSIM, GSBPM
  - Other statistical and generic standards

# Value of Metadata Standards

- Fit-for-purpose best practices from official statistics community
- Increase compatibility, interoperability of processes and systems
- Reduce development cost and maintenance burden
- Improve time-to-market with existing tools, methodology
- Improve quality with tried-and-tested methods, systems, processes
- Increase collaboration with international statistics community
- Use existing capacity building, staff with existing knowledge can be operational quicker

# Standards Explained 1/5

- GSBPM: Generic Statistical Business Process Model
  - UNECE – developed and maintained
  - Describes the activities and processes of official statistics offices
  - Some uses – classifying survey design or production systems, and system development activities
  - Adapted by the Census Bureau and BLS
  - Broad worldwide adoption

# Standards Explained 2/5

- GSIM: Generic Statistical Information Model
  - UNECE – developed and maintained
  - Conceptual, reference framework for statistical information
  - Describes inputs/outputs (e.g., data set, variable) for GSBPM processes
  - Used for designing and standardizing data architectures
  - Not directly implementable
  - Examples – National statistical offices, especially in Europe and Australia

# Standards Explained 3/5

- DDI: Data Documentation Initiative
  - DDI Alliance – developed and maintained
  - Suite of metadata standards for social and behavioral science data
  - All have an XML implementable representation
  - Codebook (2000), Lifecycle (2008), Cross-Domain Integration (late 2022)

# Standards Explained 4/5

- DDI: Data Documentation Initiative
  - Codebook – description of a data set or study, contains variables, questions, data structure
    - Example 1 – International Household Survey Network (IHSN)
    - Example 2 – Documentation of archived data sets at ICPSR (University of Michigan)
  - Lifecycle – Supports GSBPM, like GSIM, provides linkages across surveys and time
    - Example 1 – BLS Consumer Expenditure Survey public use microdata
    - Example 2 – MIDUS (Mid-life in the US) study at the University of Wisconsin
  - Cross-Domain Integration (still a draft, expected release early 2022)
    - Supports multiple data structures; linkages across variables, time, data sets; supports data integration
    - Independent of statistical domain; gaining usage in scientific and social data communities, including BLS and DOL

# Standards Explained 5/5

- SDMX: Statistical Data and Metadata eXchange
  - Mainly used to describe data and metadata sets and how they are exchanged/reported
  - Directly implementable in XML, CSV, JSON
  - Automate exchange/dissemination through standard web service interfaces
  - Has a metadata repository standard to allow distributed metadata storage
  - Numerous open-source tools available
  - Focus was on aggregated, international exchange. New version has more microdata features
    - Examples – many around the world, especially national and international banks, national statistical offices
    - Example in US – federal statistical agencies report national indicators to IMF DSBB via SDMX

# Metadata Systems

- Repository is the database for metadata
- Interface is the means to interact
- Combination is metadata system
- System can be combined with others
  - Makes metadata useful
  - Improves user experience

# Building Systems

- Obtain upper management support
  - Without this, long term success is unlikely
- Select technical specification
  - Existing standard is preferable
  - No reason to reinvent the wheel
  - Increase interoperability and consistency
- Don’t try to build a cathedral at the start
  - But use long-term plan as a guide

# Iterative Approach

- Build slowly, use iterative approach
  - Add useful new functionality at each stage
  - Easier to get funding for well-defined, small steps
- At each step
  - Build
  - Test
  - Deploy
  - Get feedback
  - Plan new functionality (based on feedback)
  - Repeat

# Contact Information

**Dan Gillman**

Information Scientist

Office of Survey Methods Research

www.bls.gov/osmr

(w) 202-691-7523

(c) 410-624-9582

Gillman.Daniel@BLS.Gov