Introducing the DDI Metadata Standards in Support of Transparency in Federal Statistics

Dan Gillman
Office of Survey Methods Research
US Bureau of Labor Statistics

COPAFS Seminar
16 March 2022

Outline

- Transparency
- Metadata
- Standards –
  - DDI (Data Documentation Initiative)

Transparency

- Many people / reports use the term
  - Presented as a worthy goal to achieve
  - Often the term is not defined
- Often stated in conjunction with privacy
  - There is a tension between the two
- But what does it mean?
  - Dictionary definition
    - Condition of being easy to perceive or detect

Transparency

- The Promise of Evidence-Based Policymaking
  - Report of the Commission on Evidence-Based Policymaking
- Transparency means
  - giving the public information
  - how the government is using their data
  - improve its effectiveness and efficiency
- Report mentions transparency 48 times prior

Transparency

- NAS/CNSTAT Report – November 2021
  - Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies
- Transparency means
  - provision of sufficiently detailed documentation
  - all the processes of producing official estimates

Transparency

- From the definitions
  - Transparency requires documentation
  - Need to find, understand, and use objects
- But documentation is situation specific
  - Documentation differs depending needs
  - Different objects => different documentation
    - Example:
    - Variables vs Questions

Transparency

- What documentation is necessary – example 1
  - Variable
    - Name
    - Definition
    - Datatype
    - Universe
    - Set of allowed values

Transparency

- What documentation is necessary – example 2
  - Question
    - Name
    - Links to subsequent questions
    - Previous question link
    - Universe
    - Response choices
    - Question text

Transparency

- Variables & questions described differently
- So, needs for transparency differ

Metadata

- Documentation
  - Usually, informal textual descriptions
- Metadata
  - Usually, formalized descriptions
- The terms sometimes used interchangeably
- Metadata
  - Data being used to describe some objects

Metadata

- When formalized, metadata are
  - Based on a set of elements
    - Elements used in description of every object in some class
  - Elements organized through a schema
    - Set of formal rules
    - Value formats and datatypes
    - Relationships between elements
  - Set of values for schema elements called an instance
- Advantage to managing metadata: reuse
  - Write once, use many times

Metadata

- Metadata schema and instance
  - Variable example:
  - Schema
    - Name
    - Definition
    - Datatype
    - Universe
    - Set of allowed values

Instance
marital status of person
legally defined marital state
nominal (unordered categories)
adults
<s, single>
<m, married>
<d, divorced>
<w, widowed>

Metadata

- Metadata as traditional documentation
- Combination of schema and instance tells a story
  - same as textual documentation
- Take the variable example, again:

The variable named *marital status of person*, which means the "legally defined marital state", is applied to the universe of adults. The set of allowed values, categories and their codes, for representing data are "s" for single, "m" for married, "d" for divorced, and "w" for widowed. This unordered set of categories is given the *nominal* datatype.

Metadata

- Achieving transparency
- Two criteria for metadata
  - Instance must adhere to rules of schema
  - Metadata must tell the right story
- More technically
  - Adherence to rules of schema --- conformance
    - Satisfying all requirements
  - Telling the right story
    - Metadata quality

Metadata

- Metadata schema
  - Aka technical specification
- Metadata standards contain technical specifications for metadata
- Scope of the standard directs what kinds of objects can be described by the spec
- Adherence to the spec is conformance to the standard
- Therefore, metadata standards support transparency

Standards in General

- Why use standards?
  - Consensus open standards have broad applicability
  - Reduce development time
    - Metadata model doesn't have to be designed
  - Conformance to a standard
    - Promote interoperability
      - Sharing between systems simplified
    - Promote transparency across agencies
      - Similar things (e.g., variables) described the same way

DDI

- Data Documentation Initiative
- Suite of metadata standards and other products
- Developed for Social, Behavioral, Economic data
- Standards process is equitable & consensus driven
  - Open – any interested agency can join the effort
  - Balanced – participants span user community
  - Visible – documents and process are inspectable
  - Fair – every participant has the same rights

DDI

- DDI managed by DDI-Alliance
  - Consortium of member organizations
    - National statistical offices
    - National social science data archives
    - University data libraries
    - University research centers
    - Consortia
    - Others
  - Over 40 members, including BLS
- Alliance officially under University of Michigan

DDI

- Three metadata standards
  - Codebook
  - Lifecycle
  - Cross-Domain Integration (CDI) still in draft, released soon
- Several products
  - XKOS (eXtended Knowledge Organization System)
  - SDTL (Statistical Data Transformation Language)
  - Controlled Vocabularies (for metadata interoperability)

DDI Codebook

- Numbered DDI-2.x, currently 2.5
- Designed to describe
  - Single use survey
  - Social science experiment
- Principal class: Study
- Reuse not part of the design
- Everything redefined or described in each instance
- Written in XML-Schema, immediately implementable

DDI Codebook

- Initial DDI standard
- Started in 1995
  - Based on electronic codebook work in 1980s
  - Implementation language
    - SGML
    - XML DTD
    - XML-Schema

DDI Codebook

- Major implementation
  - International Household Survey Network (IHSN)
  - Managed by World Bank
  - Help developing countries
  - Many free tools

DDI Lifecycle

- Numbered DDI-3.x, currently 3.3
- Based on statistical lifecycle
  - Phases taken from UNECE GSBPM
    - Generic Statistical Business Process Model
- Supports reuse, for any class of objects
- Uses variable cascade, same as UNECE GSIM
    - Generic Statistical Information Model
- Features for describing designs (new in v3.3)
  - Sampling, Questionnaire, Weighting, etc.
- Written in XML-Schema, immediately implementable

DDI Lifecycle

- Many tools
  - Open source, university research libraries
  - Commercial
- BLS application
  - Consumer Expenditure Surveys
  - Document yearly public use microdata release
  - Variables, questions, data sets, relationships (time, surveys, concepts)

DDI Lifecycle

- Many national statistical offices
  - Australia
  - Canada
  - France
  - New Zealand
  - Norway
  - Others

DDI-CDI

- Cross Domain Integration
  - Will be numbered DDI-4.x
- Intended to describe data from any source
- Supports description and integration of disparate data sets, such as
  - Traditional survey data
  - Administrative data
  - Sensor and web-scraped data
- Developed and maintained as UML model
  - XML-Schema syntax representation
  - RDF and OWL syntax representations planned
  - Others (e.g., SQL) possible

DDI-CDI

- New features in CDI:
  - Expanded variable description
  - Expanded process model
    - Recording provenance
  - Datum-centered approach: Ability to track each datum through
    - data sets, processing steps, etc.
    - shared concepts, but different representations

DDI-CDI

- New features in CDI:
  - Expanded logical data structures
    - Wide or Rectangular – typical statistical data sets, e.g., Excel file structure
    - Long – for event history data, each record has unit ID, var ID, and datum
    - Key-Value – for sensor data, each record is an ID and datum
    - Multi-dimensional – N-Cubes, Time Series
      - Ties back to microdata
      - Semantics-based structure
  - DDI provides means to transform from one to another

XKOS

- eXtended Knowledge Organization System
- Extensions to SKOS
  - W3C Simple Knowledge Organization System
  - Used to build concepts systems (hierarchies, taxonomies, ontologies)
  - Supports hierarchical relations (generic, part of, instance of)
- Support for levels for statistical classifications
  - Typically seen in NAICS and SOC
- Allows for concepts associated with each level

XKOS

- Written in W3C RDF, to integrate with SKOS
  - Resource Description Framework
- Collision with SKOS
  - Some additions in XKOS now added to SKOS
  - SKOS has semantics for many relations
- XKOS used by
  - France
  - ALPHA project (eastern Africa)

SDTL

- Structured Data Transformation Language
  - Mid-level specification
    - Intermediate language for representing data transformation commands
  - Used for
    - Documentation and description – e.g., provenance, processing
    - Translation between statistical languages
      - E.g., SAS, SPSS, Stata, R, and Python

SDTL

- Structured Data Transformation Language
  - Can be translated into natural language
    - Users need not know specifics of specific processing language
    - Provides means to classify text describing the steps in a process

Controlled Vocabularies

- Supports interoperability
- Provides language to consistently record commonly used values
- Examples
  - Units of analysis: individual, family, household, etc.
  - Telephone type: fixed, mobile, fax, etc.
  - Note type: comment, observation, system, processing, etc.
  - Many more

Contact Information

Dan Gillman

Office of Survey Methods Research

US Bureau of Labor Statistics

www.bls.gov/osmr

(w) 202-691-7523

(c) 410-624-9582

Gillman.Daniel@BLS.gov