New Developments with the Data Documentation Initiative (DDI)

Dan Gillman
Information Scientist
Office of Survey Methods
UNECE ModernStats World Workshop
30 October 2020

Data Documentation Initiative (DDI)

- DDI is not one standard
- Instead, family of standards and other work products
  - Codebook (DDI 2.5)
  - Lifecycle (DDI 3.3)
  - XKOS – statistical extension of SKOS
    - eXtended Knowledge Organization System
  - Controlled Vocabularies
    - Datatype
    - Analysis unit

Data Documentation Initiative (DDI)

- New draft work products
  - DDI-CDI
    - Cross Domain Integration
    - Draft released in April
    - Final expect in 2021 Q1
  - SDTL
    - Structured Data Transformation Language
    - Draft released in July
    - Final expected in early 2021

Data Documentation Initiative (DDI)

- Discuss CDI and SDTL
- And include
  - Main features
  - Innovations in CDI
  - Potential effects for NSOs
  - Ways CDI and SDTL work together
  - Compare SDTL with VTL

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 Lifecycle

- Numbered DDI-3.x, currently 3.3
- Based on statistical lifecycle
  - Phases similar to GSBPM
- Supports reuse, for any class of objects
- Uses variable cascade, similar to GSIM
- Features for describing designs (new in v3.3)
  - Sampling, Questionnaire, Weighting, etc.
- Written in XML-Schema, immediately implementable

XKOS

- eXtended Knowledge Organization System
- Extensions to SKOS
  - W3C Simple Knowledge Organization System
  - Used to build concepts systems (hierarchies, taxonomies, ontologies)
  - Supports hierarchical relations (subtype, part of, instance of)
- Support for levels for statistical classifications
  - Consistent with Neuchâtel Classification model
- Allows for concepts associated with each level
- Written in RDF, to integrate with SKOS

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

New: 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 cascade
    - Added differentiation: datatypes (intended vs actual), value domains (substantive vs sentinel), units of measure
  - Expanded process model
    - Recording provenance
    - Implementation profile of BPM to make GSBPM computable
  - 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

DDI-CDI Logical Data Structures
Wide

Person ID	Sex	Reside	Born	Died
Marie	female	Maryland, USA	17 Feb 1930	6 Jul 2020
Henry	male	California, USA	12 Jun 1928	20 Jan 2016

DDI-CDI Logical Data Structures
Long

Case ID	Variable ID	Value
Marie	Sex	female
Marie	Reside	Maryland, USA
Marie	Born	17-Feb-1930
Marie	Died	6-Jul-2020
Henry	Sex	male
Henry	Reside	California, USA
Henry	Born	12-Jun-1928
Henry	Died	20-Jan-2016

DDI-CDI Logical Data Structures
Key-Value

Key	Value
MaSe	female
MaRe	Maryland, USA
MaBo	17-Feb-1930
MaDi	6-Jul-2020
HeSe	male
HeRe	California, USA
HeBo	12-Jun-1928
HeDi	20-Jan-2016

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

SDTL

- Versus VTL (Validation and Transformation Language)
  - SDTL – still in draft
    - Based on needs of statistical packages
    - Intended to work with DDI standards
    - Emphasis on microdata
  - VTL – in use in production
    - Formalization of GSIM, very mathematical
      - Lower level specificity
    - Based on Object Management Group / Meta-Object Facility
      - 4 tier architecture
    - Emphasis on aggregates

VTL	SDTL
Includes both data transformation and validation commands	Only data transformation commands
Designed to be executable	Designed for documentation, not execution
Requires parsing according to syntax rules	Machine readable without parsing; Structured with tags and nesting (e.g., JSON, XML, RDF);
VTL may be translated into other languages for execution	Intermediate language that can be used for translation between languages that requiring parsing (including VTL)
	Includes schema and software for translation into natural language

Review Packages

- DDI-CDI review
  - https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/860815393/DDI+Cross+Domain+Integration+DDI-CDI+Review
- SDTL review
  - https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/1120370729/SDTL+Review

Questions

Contact Information

Dan Gillman
Information Scientist
BLS/OSMR/MSRC
www.bls.gov/osmr
202-691-7523
Gillman.Daniel@bls.gov