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
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
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

<table>
<thead>
<tr>
<th>Person ID</th>
<th>Sex</th>
<th>Reside</th>
<th>Born</th>
<th>Died</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marie</td>
<td>female</td>
<td>Maryland, USA</td>
<td>17 Feb 1930</td>
<td>6 Jul 2020</td>
</tr>
<tr>
<td>Henry</td>
<td>male</td>
<td>California, USA</td>
<td>12 Jun 1928</td>
<td>20 Jan 2016</td>
</tr>
</tbody>
</table>
## DDI-CDI Logical Data Structures
### Long

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Variable ID</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marie</td>
<td>Sex</td>
<td>female</td>
</tr>
<tr>
<td>Marie</td>
<td>Reside</td>
<td>Maryland, USA</td>
</tr>
<tr>
<td>Marie</td>
<td>Born</td>
<td>17-Feb-1930</td>
</tr>
<tr>
<td>Marie</td>
<td>Died</td>
<td>6-Jul-2020</td>
</tr>
<tr>
<td>Henry</td>
<td>Sex</td>
<td>male</td>
</tr>
<tr>
<td>Henry</td>
<td>Reside</td>
<td>California, USA</td>
</tr>
<tr>
<td>Henry</td>
<td>Born</td>
<td>12-Jun-1928</td>
</tr>
<tr>
<td>Henry</td>
<td>Died</td>
<td>20-Jan-2016</td>
</tr>
</tbody>
</table>
DDI-CDI Logical Data Structures
Key-Value

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaSe</td>
<td>female</td>
</tr>
<tr>
<td>MaRe</td>
<td>Maryland, USA</td>
</tr>
<tr>
<td>MaBo</td>
<td>17-Feb-1930</td>
</tr>
<tr>
<td>MaDi</td>
<td>6-Jul-2020</td>
</tr>
<tr>
<td>HeSe</td>
<td>male</td>
</tr>
<tr>
<td>HeRe</td>
<td>California, USA</td>
</tr>
<tr>
<td>HeBo</td>
<td>12-Jun-1928</td>
</tr>
<tr>
<td>HeDi</td>
<td>20-Jan-2016</td>
</tr>
</tbody>
</table>
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
<table>
<thead>
<tr>
<th>VTL</th>
<th>SDTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Includes both data transformation and validation commands</td>
<td>Only data transformation commands</td>
</tr>
<tr>
<td>Designed to be executable</td>
<td>Designed for documentation, not execution</td>
</tr>
<tr>
<td>Requires parsing according to syntax rules</td>
<td>Machine readable without parsing; Structured with tags and nesting (e.g., JSON, XML, RDF);</td>
</tr>
<tr>
<td>VTL may be translated into other languages for execution</td>
<td>Intermediate language that can be used for translation between languages that requiring parsing (including VTL)</td>
</tr>
<tr>
<td></td>
<td>Includes schema and software for translation into natural language</td>
</tr>
</tbody>
</table>
Review Packages

- DDI-CDI review

- SDTL review
Questions
Contact Information

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