DDI-4 Cross-Domain Integration: Metadata for a New World

Dan Gillman
US Bureau of Labor Statistics
Office of Survey Methods Research
FedCASIC
April 14, 2021

DDI

- Data Documentation Initiative
- Program building statistical metadata standards
  - For data libraries, archives, producers, researchers
- Work managed under DDI Alliance
  - Secretariat at ICPSR, University of Michigan
- Currently, 2 standards
  - DDI Codebook (DDI-2, version 2.5)
  - DDI Lifecycle (DDI-3, version 3.3)

DDI-2 Codebook

- For describing a single study or data set
- No links between different codebooks
- Simple framework, easily adopted
  - Variables / Questions / Code lists / Data sets
  - Brief descriptions of methodology
- International Household Survey Network
  - Managed by World Bank
  - For documenting surveys in developing countries
  - Free software – building and querying Codebooks

DDI-3 Lifecycle

- For describing the survey lifecycle
  - Consistent with phases in UNECE GSBPM
    - Generic Statistical Business Process Model
- Built for data producers
  - E.g., federal statistical agencies
- Allows complex linking of metadata
  - Across collections from same survey
  - Across surveys and organizations over time

DDI-3 Lifecycle

- In use in many statistical offices, including BLS
  - Document Consumer Expenditure Surveys
    - Quarterly interview and Diary
  - Annual microdata release
    - Across surveys, over time, linking questions and variables

DDI-4 Moving Forward Project

- Begun in 2012
- Significant development in sprint meetings
  - Adjuncts to conferences
  - Special meetings (in a German castle)
- Plan: Manage Codebook and Lifecycle
  - Using model driven approach – UML
  - Manage further development of DDI standards
  - Automatically generate Codebook and Lifecycle
- But, effort stalled due to complexity

New World

- Several changes in recent years impact requirements
  - Larger research projects using data sometimes coming from external domains
  - More data, coming from a wider range of sources
  - Increased ability to compute with data (Machine Learning, etc.)
- Changes result in new requirements for data/metadata
  - More complete, machine-actionable metadata is needed
  - Improved “context” for data is needed (provenance, semantics)
  - New data formats/structures must be described and integrated
  - A broader range of technology platforms require support

New World

- Survey response rates diminishing
- New need new sources for data
  - Administrative data
  - Web scraping and other sources
- Evidence Act 2018

- Lack of results in DDI-4 effort
- Hence …

DDI-CDI

- Cross-Domain Integration
- New entry in DDI family
- Currently in draft form
  - Public review began April 2020
  - Intro webinar to UNECE statistical community
    - July 2020
- Planned release date
  - Summer 2021

DDI-CDI Design Goals

- Produce a useful, implementable product based on real use cases
- Produce a standard which would be useful across technology platforms (model-driven)
- Produce a standard which is more approachable and easier to understand
  - W3C specifications used as a model
  - Lots of examples at different levels

Model-Driven Standard

- Continue with UML based development
- Compatible with other standards
  - Especially W3C work
- Can generate multiple syntax representations
  - XML (currently exists), RDF, SQL, etc.
- UML itself is portable through use of XMI
  - XML Metadata Interchange
  - Language for sharing UML models
- From model, can generate profiles and packages
  - Targeted and useful subsets of entire model

DDI-CDI

- What's in CDI?

Foundational Metadata

- Building on years of work in DDI 4
- Sophisticated model for variables, conceptual underpinnings/application
- Works flexibly with different ontologies/concept systems/thesauri
- Well-aligned with DDI-Lifecycle
- Compatible with new DDI-SDTL
  - Structured Data Transformation Language
  - Documentation for individual processing steps

DDI-CDI Enhanced Functionality

- Many data structures:
  - (Not just) Rectangular/unit-record
  - Event history
  - No-SQL/"big data" – key-value
  - Multi-dimensional – data cubes and time series
- Describe data provenance/process
  - Procedural process
  - Declarative process
- Describe "foundational" metadata
  - Codes/categories/classifications
  - Concepts, variables, etc.

DDI-CDI Data Description

- "Datum-Centered Approach"
  - Specify different roles for data in data sets
    - measure, descriptor, identifier
- Describe 4 types of data structure
  - The model can easily extend to describe others
- Data transformation tools perform this kind of thing all the time
  - DDI-CDI can express the relevant metadata for tracking datums across different structures
  - No other standard has this capability

DDI-CDI Process and Provenance

- DDI standards don't describe the processes that are combined to actually produce data
  - Focus has always been on low-level data processing (stats packages/SDTL)
- DDI-CDI describes processes at a higher level, and connects them with low-level processing descriptions
- Directly implements common models for provenance and process (PROV, BPMN)
- Supports "black box" parallel processing as well as stepwise "flow" processing
  - New feature of DDI
  - Becoming common in the real world

DDI-CDI Alignment with Other Standards

- DDI-CDI directly implements other standards at the level of UML ("trace" relationships)
- DDI-CDI is *domain-neutral*
- Aligned with other flavors of DDI (Codebook. Lifecycle, etc.)
- Directly implements process/provenance standards (BPMN, PROV)
- Supports GSIM/GSBPM
- Designed to integrate with discovery standards (Schema.org, DCAT)
- Aligned with other data description models (CSV on the Web, SDMX, Data Cube, Observable Properties, SOSA/SSN, etc.)
  - Some work remains in testing these alignments

Questions

Contact Information

Dan Gillman
Gillman.Daniel@bls.gov