Data Documentation Initiative through a BLS Example

Daniel Gillman

U.S. Bureau of Labor Statistics

Workshop on Transparency and Reproducibility in Federal Statistics; Washington, DC
22 June 2017

Data Documentation Initiative

- Standards and products for statistical metadata
- Developed
  - Under DDI Alliance, consortium managed by ICPSR
  - By statistical offices, libraries, archives, researchers
- Products
  - Codebook, DDI v2.x (current 2.5)
  - Lifecycle, DDI v3.x (current 3.2)

Data Documentation Initiative

- Products cont'd
  - Resource Description Framework vocabularies
    - Discovery (Disco)
      - Data set discovery
    - Physical Data Description (PHDD)
      - Rectangular file description
    - eXtended Knowledge Organization System (XKOS)
      - Describe statistical classifications

Data Documentation Initiative

- Current work
  - DDI-4 Moving forward
    - Generic (UML) model to support all current products
    - Expanded description of survey lifecycle
    - Multiple bindings
      - XML
      - RDF
      - SQL (planned)
- Many implementations / covers all DDI products

Consumer Expenditure Surveys

- Measures how US people and households spend money
- Conducted by BLS; data collected by Census
- Consists of 2 surveys
  - Interview (Quarterly)
    - Includes large or recurring expenses (e.g., rent)
  - Diary (2 Week)
    - Includes small, frequent expenses (e.g., groceries)
- Processing
  - 4 subsystems

Motivation

- Complete survey processing redesign
- Currently, variables managed through
  - Independent MS Access databases
    - One per subsystem
    - Tracking variables across DBs is very hard
- Single system for managing variables
  - Across surveys (Interview and Diary)
  - Throughout life-cycle
    - Including dissemination
  - Over years

Goals

- Want to show:
  - How similar variables change over time
    - Changes occur in odd years
  - How similar variables change over life-cycle
    - Including code list differences
    - Groupings – expenditure and UCC
  - Entire life-cycle
    - Questions to final variables and tables
    - Include all production subsystems

Goals

- Want to show:
  - Full processing steps
    - Links to variables as inputs/outputs
    - Show flow through and between each subsystem
  - Instrument design
    - Including wording and skip flow
    - Ability to input / output transferrable questionnaire

Path to Solution

- Need metadata system
  - Selected Data Documentation Initiative (DDI)
    - Version 3.2 - Lifecycle
  - Selected DDI commercial software
    - Colectica Designer and Repository / Portal
- Develop system iteratively
- Start small
  - Series of pilot systems

Results

- CE documentation needs
  - Pilot 1
    - Showed DDI is sufficient for needs
  - Pilot 2
    - Showed Portal is sufficient for needs
    - Details to follow

Pilot 2 Results

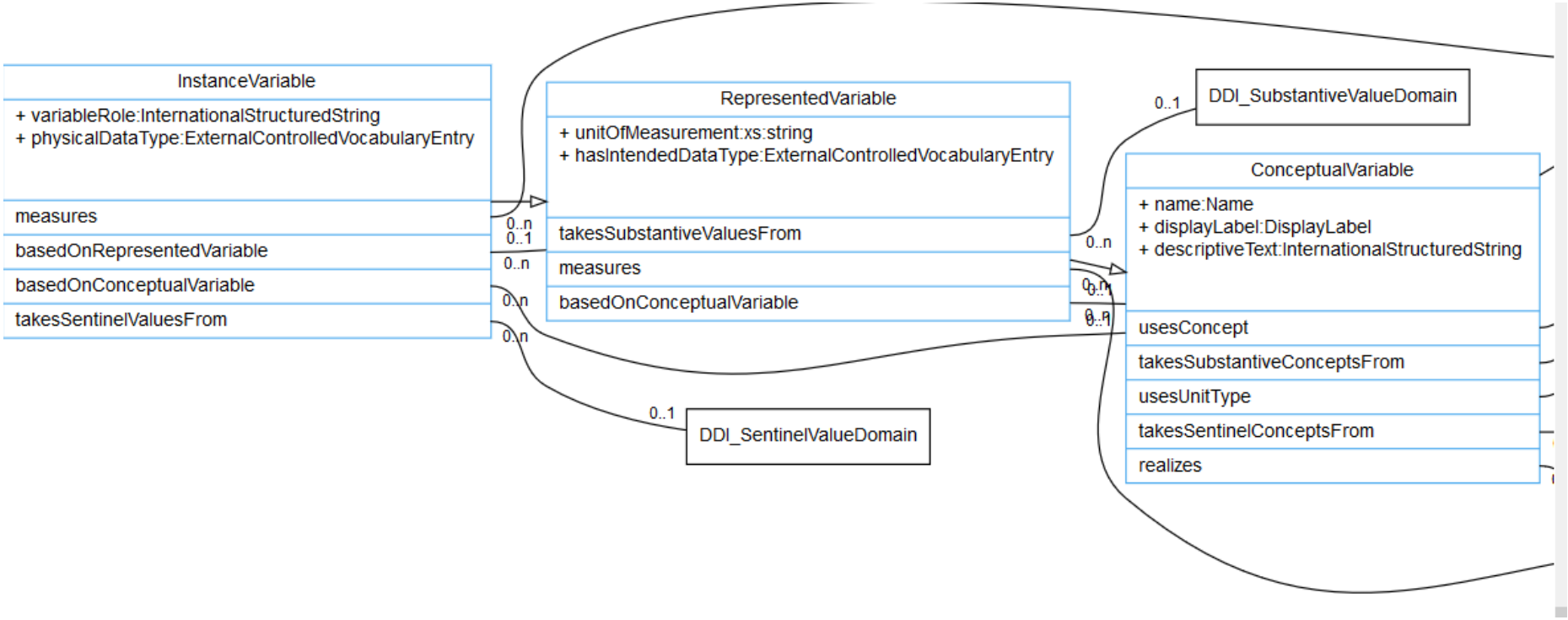
- Time / resource limited
- Show 2012 and 2013 metadata
- Illustrate with 2 variables
  - Education
  - EIHB expenditure set
    - Hospitalization and Health Insurance

Pilot 2 Results

- Compare variables
  - Using Correspondence Tree
    - Across surveys
    - Over time
    - Through lifecycle
  - Using Code Comparison
    - Observe changes to categorization
      - Substantive
      - Gratuitous

Variable Cascade in DDI

http://lion.ddialliance.org/package/conceptual



The InstanceVariable can reference RepresentedVariable and ConceptualVariable. It also inherits all properties and relationships from them.

R

STATISTICS

Basket 0

Label Highest Grade Completed?

Statistics | Code Comparison | Correspondence Tree

EDUCA — Highest Grade Completed?

EDUCA — Highest Grade Completed?

2012 EES Subsystem - MEMB.EDUCA — Highest Grade Completed?

2012 IES Subsystem - MEMB.EDUCA — Highest Grade Completed?

EDUCA — Highest Grade Completed?

2013 IES Subsystem - MEMB.EDUCA — What is the highest level of education that member has/you have completed?

Pilot 2 Results

- Educa code comparison
  - Highest grade completed
  - Substantive differences
    - 2012
      - 21 categories
      - One for each grade up to 12
      - Nursery/kindergarten not available choices
    - 2013
      - 8 categories
      - Nursery/kindergarten/elementary versus high school
      - One category for professional/masters/PhD

Pilot 2 Results

- Educa code comparison
  - Highest grade completed
  - Gratuitous differences
    - Category labels change
      - Added blanks
      - "12th" versus "twelfth"
      - "professional school degree" versus "professional degree"
      - "high school (grades 9-12), no degree" versus "high school (grades 9-12, no degree)"
      - "high school graduate – high school diploma or the equivalent (GED)" versus "high school graduate - high school diploma or equivalent (GED)"

cs

ode Comparison | Correspondence Tree

	2012 IES Subsystem	2012 EES Subsystem	memi121x	memi122	memi123	memi124	memi131	2013 IES Subsystem	2013 EES Subsystem
	00	00							
h 11th grade	01-11	01-11	11	11	11	11	11		
iploma	38	38							
uate	39	39	39	39	39	39	39		
degree	40	40							
, occupational/vocational	41	41							
, academic	42	42							
e	43	43							
	44	44							
ol degree	45	45							
e	46	46	46	46	46	46	46		
npleted, or less than 1 year								1	1
arten, and elementary (grades 1-8)								2	2
des 9-12), no degree								3	3
uate - high school diploma or equivalent (GED)								4	4
t no degree								5	5
e in college								6	6
e (BA, AB, BS, etc.)								7	7
ional, or doctorate degree (MA, MS, MBA, MD, JD,								8	8
des 9-12, no degree)									
uate – high school diploma or the equivalent (GED)									
no degree									

Pilot 2 Results

- Path of variables through life-cycle
  - Use Lineage
    - For given year
    - Question
    - Each processing subsystem
    - Final output

R

STATISTICS

Basket 0

est Grade Completed?

Interview > 2012 > memi131 6

Variable | Details | Lineage | Concordance Variables

memi123 - EDUCA — What is the highest level of school the member has completed or the highest degree the member has received?

2012 EES Subsystem - MEMB.EDUCA — Highest Grade Completed?

2012 IES Subsystem - MEMB.EDUCA — Highest Grade Completed?

educa2012 — WHAT IS THE HIGHEST LEVEL OF SCHOOL (NAME) HAS/YOU HAVE COMPLETED OR THE HIGHEST DEG

memi131 - EDUCA — What is the highest level of school the member has completed or the highest degree the member has received?

2012 EES Subsystem - MEMB.EDUCA — Highest Grade Completed?

Pilot 2 Results

- Processing diagram
  - Four subsystems
  - Each broken into smaller systems
  - Initial Edit Subsystem (IES) uses 3rd level

- Limitation
  - No links to input/output variables

R

STATISTICS

Basket 0

CE Process Flow

Common Metadata | BLS CE Instruments

Flowchart | Details

Start

Census

IES

EES

Tables and Microdata

End

R

STATISTICS

Basket 0

CE Process Flow

Common Metadata | BLS CE Instruments

Flowchart | Details

Start

- Census
- IES
  - Bundle 1
  - Bundle 2
  - Bundle 3
    - **Dynamic EoP** — Dynamic end of processing review
- EES
- Tables and Microdata

R

STATISTICS

Basket 0

Bundle 2

Month Verification	This edit verifies that the month of expenditure is a valid month for that processing month and questionnaire
RPA Part 3	This edit assigns current month expenditure information to the subsequent interview. It is set up to handle the only sec instrument, tables EOPI and ECRB. It is referred to as "Part 3" since parts 1 and 2 are now part of the CAPI instrument
Tenure	This edit derives CU level QTENURE. After QTENURE is derived, checks are done to determine whether related housing QTENURE values. The results of these checks may include the creation or deletion of records. Inconsistencies are corre
Inventory Consistency	This edit does something to inventory to make it consistent :-)
Change Property Loan	This edit handles the case where a mortgage or lump sum home equity loan is changed during the reference period. Th Section 3F, OPF. The information about the change to the loan, e.g., refinance, paid-off, etc. is in section 3J, OPJ.
NOTE Inventory	The Phase 3 Screens and other data reviews require access to the notes to help determine whether an expenditure is c store any first Interviews or Type A CUs, the inventory of the notes needs to be done in Phase 2 processing in order to
Demographics	This edit ensures that for selected data items, a valid data value is contained on the data file and ensures certain consis questionnaires which are interviews and Type A non-interviews. The field MEMBRACE is created in the first step followe the AGE field

Contact Information

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

Pilot System

- http://bls-eval.colectica.org