Active Metadata at the US Department of Labor

Dan Gillman
Information Scientist
US Bureau of Labor Statistics
COSMOS
12 April 2024

US Department of Labor

- Over 20 agencies
- Devoted to different aspects of the labor force
- 5 enforcement agencies
  - Seek compliance with labor laws
  - Produce data on inspections for public use
  - Existing metadata is scant
- Several policy agencies as well
  - Executing programs for improving and measuring working conditions

US Department of Labor

- Office of Data Governance
  - Under Chief Data Officer
  - Independent of CIO
- Devoted to improving FAIRness and quality of DOL data
- Developed multi-stage approach
  - Without any mandate and very small budget
  - Using examples, building in small steps, and generating goodwill

Data Governance

- Strategy consists of layers
  - EDI – inventory for all DOL data sets (public, restricted, private)
  - API – access to publicly available DOL data
  - Portal – web site to guide users to DOL data
- Metadata – in 3 layers
  - Top: EDI catalog
  - Middle: Data content at data set level
  - Bottom: Description of variables, based on Instance Variable
    - IV as described in DDI Cross-Domain Integration

Metadata

- EDI level
  - Descriptors for each data set
    - Agency, Contact, Name, etc.
- Mid level
  - Descriptors for general content
    - Geographic coverage and detail
    - Classification schemes used
    - Unit types associated with the data records

Metadata

- Mid level, cont'd
  - Descriptors for general content
    - Applicable laws and regulations
    - Data set structure (how data are logically organized)
- Low level
  - Descriptors for variables
    - Universe
    - Datatype (application and intended)
    - Definition
    - Range limits, Units of Measure, Precision

Metadata

- Low level
  - Descriptors for value domains
    - Structure
      - Range, Rule, List, Reference
    - Allowed value description
      - Generic numeric range, Regular expression, Code/value pairs, URL
    - Type
      - Substantive, Sentinel

Metadata

- Low level, cont'd
  - Summary Statistics
    - Numeric
      - Min/max, 1st, 3rd quartiles, mean, median
    - Categorical
      - Top 10 categories, by percentage
  - Valueless entries
    - Entries under a variable not among allowable rules

Active Metadata

- All metadata in machine interpretable form
- Controlled vocabularies used
- Regular expressions for text and identifiers
- Numeric ranges interpretable as integer, real, or currency
- Categories linked to concepts
- Codes clearly differentiated from their meanings

Metadata Driven Data Quality

- Data Quality
  - DOL agencies don't take the time to ensure quality
  - Data are not first order of business
  - Quality checks are for finding errors
- Valueless entries
  - Uninterpretable entries – determined through Value Domain
  - List generated for each variable in each data set

Metadata Driven Data Quality

- Range checks – find range values outside predefined limits
- Consistency checks – consistency across multiple variables
  - Example: sex – male; age – 90; pregnant – yes ????
  - Consistency of geographic variables; ZIP and state/county + city
  - Consistency of classification codes and descriptions
- Geographic location
  - Geo-coding and address standardization

Need for Consistent Identifiers

- DOL cannot check if establishments are serial violators
- No standard ID
  - Cannot use BLS business register ID – confidentiality pledge
- Address standardization for businesses helps, but no guarantee
  - Close then reopen under new ownership / type
  - Move
  - Change name, but otherwise remain the same

Questions

Contact Information

Dan Gillman
Office of Survey Methods Research
www.bls.gov/osmr
+1.202.691.7523
+1.410.624.9582
Gillman.Daniel@bls.gov