Taxonomy and Lexicon Project at the US Bureau of labor Statistics

Daniel Gillman
Office of Survey Methods Research
US Bureau of Labor Statistics

APDU 2017
14 September 2017

Drivers

- Current data access
  - Subject matter specific tools
- Limitation
  - Can’t download 2 or more “unrelated” data sets at once
- Desire
  - One tool for all data
  - See DataFinder under “beta” tab at BLS.Gov

Drivers

- Desire cont'd
  - Need user interface for DataFinder
  - Taxonomy of terms
  - Needs to be "user friendly"
    - Reduced reliance on technical terms
    - Use plain English words at higher levels

Drivers

- Current search engine needs upgrading
  - Results inconsistent
    - Data series don't match data in
      - Reports
      - Documents
      - MLR articles
    - Retrieved items lack relevance
- Desire
  - Consistent tagging of data and documents
  - Need lexicon derived from taxonomy for tagging

Drivers

- Data retrieval issues
  - How to know all data on some subject is provided?
  - What data do we have about Boston?
    - Boston has 6 definitions!
  - What data we have on the nursing industry?
    - Note – nursing is not an industry!
    - Confusion between industry and occupation
      - "Field of Work" is commonly used phrase
  - Compare and contrast
    - Total employment (e.g., CES versus CPS in employment situation)

Drivers

- Data retrieval issues
  - Compare and contrast cont'd
    - Income, Wages, Earnings, Benefits, Compensation
      - Average weekly wages: CES based on hours
      - QCEW based on quarter of 13 weeks
    - "job creation"
      - CES – based on establishment counts
      - BED – job creation and destruction

Drivers

- Web site organized by program
  - Leads to specific data tools
- Subject matter details are missing
  - Geography – all levels
  - Industry
  - Occupation
  - Injury and Illness
- No direct link from subject matter
  - Can't get all data for a city, MSA, county, or state
  - Can't get all data for specific industry or occupation
- Need tagged items and search engine

Technical Details

- Focus on time series; BLS produces millions
- How do we break these up into terms?
- What is a good way to organize the terms?
- What can't we do?
  - Say what BLS does not have
  - Say how searching for data will proceed
- Tasks
  - Build taxonomy
  - Don't worry about applications

Technical Details

- Main task - disentangle data, for example
  - Urban Consumer Price Index
    - For apparel
    - In Washington, DC
- Measure
  - Def'n: Quantitative estimate for some set of units
  - Consumer price index
- Characteristics
  - Def'n: Dimensions in which measure is stratified
  - Product, Geography (MSAs)

Technical Details

- Identified Characteristics:
  - Geography
  - Industry
  - Occupation
  - Products and services
  - Benefits
  - Demographics
  - Workers
  - Establishments
  - Workplace illness and injury
  - Time

Technical Details

- Measures
  - Estimates on set of units
    - Which basic ones?
  - Statistical metadata community
    - Unit types – basic kinds of units
      - People
      - Establishments
      - Workers
      - Workplace
      - Injuries
      - Etc.
  - Allows user to find comparable data
    - E.g., (all) data on Workers

Technical Details

- Measures cont'd
  - Quantitative evaluation of some feature
  - For CPI
    - Unit type – consumers
    - Feature – prices
  - For labor Force
    - Unit type – people (civilian, non-institutional)
    - Features – next page

Technical Details

- Partial list of features – labor force
- Labor force
  - Total
  - Participation rate
  - Employed
    - Total
    - Employment-population ratio
  - Unemployed
    - Total
    - Unemployment rate

Technical Details

Datatypes for features – labor force

- Labor force
  - Total — count
  - Participation rate — ratio as percent
  - Employed
    - Total — count
    - Employment-population ratio — ratio as percent
  - Unemployed
    - Total — count
    - Unemployment rate — ratio as percent

Technical Details

- Measures cont'd
  - What does evaluation look like?
    - Quantity is based on some formula or procedure
    - Number has computational properties
  - Akin to datatype in computer science
    - For CPI, this is an index
    - For unemployment rate, a percent
  - Next page for Labor Force example

Taxonomy Structure

- Two main dimensions
  - Measures
  - Characteristics
- Characteristics
  - Independent facets, some multi-level hierarchies
- Measures
  - Unit types
  - Features – interleaved hierarchies
  - Datatypes

Taxonomy Work

- Currently in 5th 6-month phase
- Plan for cognitive testing at end of this phase
- Incorporation into DataFinder user interface
- Guide for web site reorganization
- Basis for document tagging
- Maintenance schedule

Contact Information

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