Further Developments of the Taxonomy of Terms & Concepts at the BLS

Daniel W Gillman

Team Leader, Taxonomy and Lexicon Team

Beyond the Numbers,
St. Louis Federal Reserve Bank

October 7, 2016

Motivation

- Questions from users –
  - I want data about the nursing industry.
  - What information do you have about Boston?
- Technical improvements –
  - Data access is minimally subject matter related.
  - Impossible to pull more than one data series at once.

Observations

- Nursing industry?
  - Nursing is an occupation
  - How do we explain the difference?
- Boston?
  - We have 6 definitions of Boston
  - How do we steer the user to the right geographic definition?

Observations

- Data access
  - Subject area dependent
  - Not accessible by details
    - Nursing
    - Boston
  - One series at a time
  - Documents and Data not consistently searchable

Vision

- Single time series data access point
- Consistent retrieval of data and documents
- Access to data via broad and detailed subjects
- Web site able to accommodate this

Path

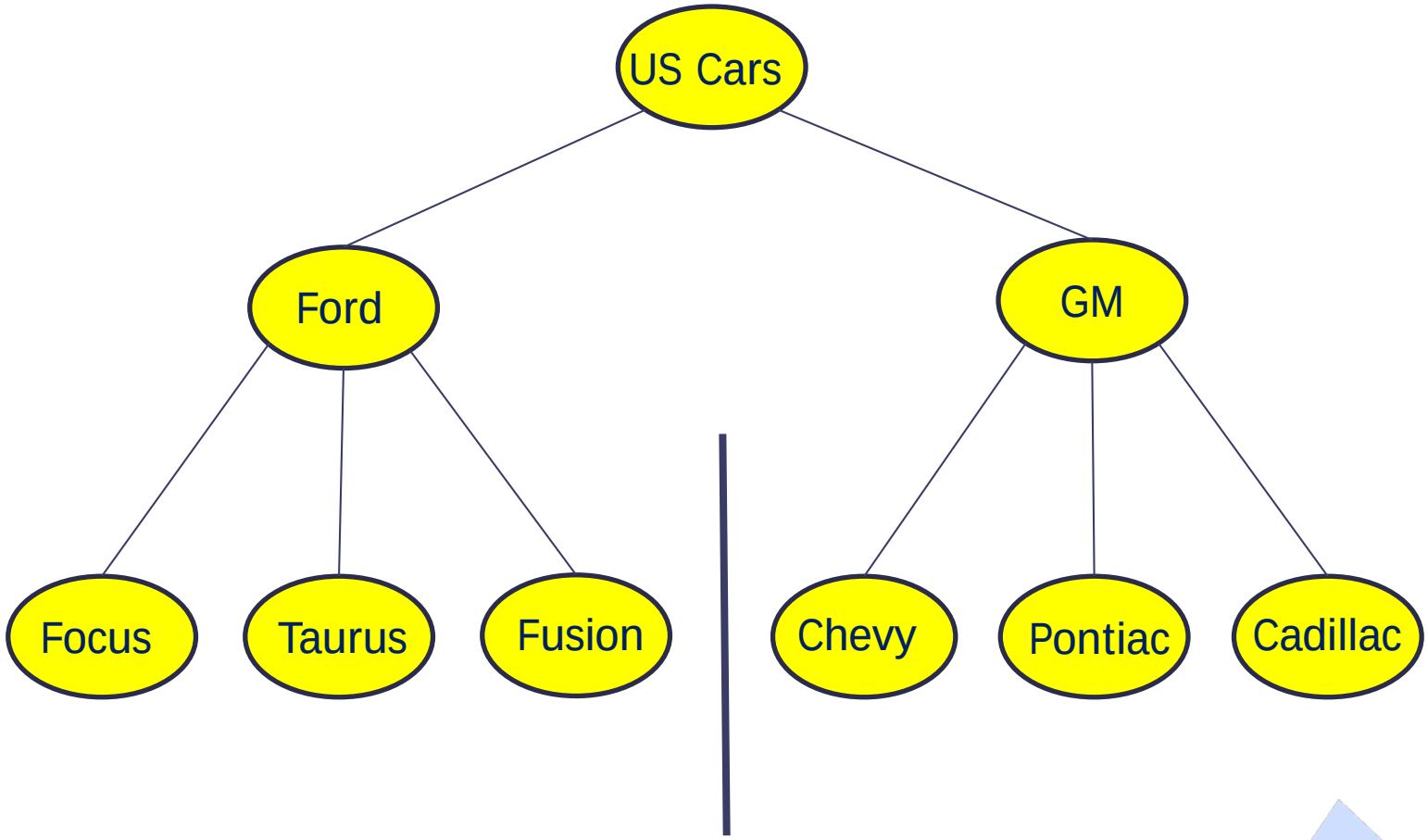
- Build taxonomy and lexicon of terms for BLS
- This begs lots of questions

What is a taxonomy?

- Need to understand a hierarchy

- *Hierarchy –*
  - *A system or organization in which things are ranked one above the other according to some criteria –* Adapted from on-line definition

Hierarchy - example



Taxonomy - definition

- *Taxonomy –*
  - *A collection of controlled vocabulary terms and words organized into a hierarchical structure. Each term or word is in one or more parent/child relationship(s) to others in the taxonomy.* – M. Hlava

Taxonomy - examples

- Biological Classification of Living Things
  - http://anthro.palomar.edu/animal/animal_1.htm
- North American Industrial Classification (NAICS)
  - https://www.census.gov/eos/www/naics/
- Standard Occupational Classification (SOC)
  - http://www.bls.gov/soc/

What is a Lexicon?

- *A vocabulary (i.e., set of terms) of an organization or branch of knowledge* – Modified from on-line definition
  - More generally, includes dictionaries for Natural Language

- BLS uses
  - Specifically defined terms
  - Covering Labor Economics and Statistics

Lexicon Examples

- Oxford English Dictionary
- Any glossary of technical terms
  - E.g., BLS Glossary
    - http://www.bls.gov/bls/glossary.htm
- Urban Dictionary
    - http://www.urbandictionary.com/
- Legal Dictionary
    - http://dictionary.law.com

Project Objectives

- Build taxonomy of terms and concepts
  - Support finding BLS data – time series, tables
- Build a lexicon
  - Based on the taxonomy
    - "Flattened", no structure
  - Supports
    - Tagging
    - Searching
    - Retrieving documents – e.g., articles, news releases

Project Objectives

- Incorporate plain English words
  - Help guide non-technical users
- Goals
  - Consistent retrieval of
    - Data
    - Documents
  - Provide user interface for data retrieval tools
  - Guide to reorganization of web site
  - Resource for meaning of terms

Identifying Plain English Words

- Interview each
  - Program area
  - Regional office
- Met with staff that interact with public
- Identify areas of confusion public often have
- Map plain English to BLS terminology

Plain English - examples

- Inflation – maps to CPI
  - A very common idea
- Field of work – maps to Industry & Occupation
  - Commonly found
  - Public don't differentiate industry & occupation
  - "I want data about the nursing industry"
- Pay, Salary, Wages, Income, Compensation
  - Many possible confusions

Taxonomy Development

- Work done in phases
- Limit initial scope to
  - Taxonomy only; do Lexicon later
  - Describe data associated with Time Series
- Divide Taxonomy into 2 facets:
  - Measures – estimates on some population
    - Unemployment Rate
    - Consumer Price Index
  - Characteristics – categories for stratifying measures
    - Geography
    - Product/Commodity

Measures Facet

- Develop high level taxonomy to drill down to measures
- Use plain English words
  - Name categories in top levels
  - Develop meaningful paths to technical terms
  - But, violate strict hierarchy by
    - Build multiple paths to some measures

Measures Facet

- Jobs
- People
- Employers
- Prices

People Category - expanded

- People ->
  - Consumer Assets and Liabilities
  - Consumer Prices and Inflation
  - Consumer Spending
  - Employed *
  - Labor Force
  - Pay and Benefits *
- People ->
  - People and Families
  - Strikes and Union Membership *
  - Unemployed
  - Work Hours *
  - Workplace Injuries *
  - Not in Labor Force

Characteristics Facet

- Classifications
  - Used to stratify some measures
- Develop hierarchy for each subject set
  - Make single hierarchy to incorporate all versions
    - E.g., Industry contains
      - NAICS 2012, NAICS 2007, Census 2000
- Use plain English to name
  - High level categories
  - Categories that bridge versions

Industry Example

- Example – Monthly Employment Situation
  - Table B – Establishment Data
    - Employment by Selected Industry
      - Government
- "Government" is not a NAICS category
- Taxonomy has to account for it
- Therefore, adopt NAICS-like structure
  - But, expanded from NAICS, Census, and even SIC

Injury and Illness Example

- OIICS
  - Occupational Injury and Illness Classification
- 4 main facets

Facets	Example
Nature	sprain
Body Part	ankle
Source	concrete floor
Event	fall from step

Injury and Illness Example

- What do these facets mean?
  - Body part? – pretty easy
  - Nature? – maybe not so clear
  - Narrative – I sprained my ankle on the hard concrete floor by falling off a step.
- Will this seem natural to new users?
- Are there better synonyms or phrases?
- Also, can't get too complex

List of Characteristics

- Geography
- Occupation
- Industry
- Establishments/Businesses/Firms
- Products/Commodities/Services
- Demographics - Characteristics of People
- Worker Injury and Illness
- Time
- Unemployment and Labor Force Status
- Worker Characteristics

Current Work

- Characteristics facet revised
- Undergoing program area review
  - Comment resolution
- Measures facet being revised
- Public review planned
  - Include cognitive testing
- Further program review planned

Future Work

- Develop Lexicon
  - Or, replace with text classification
- Expand beyond time series data
  - Include all data
- Expand beyond data terms
  - Include statistical/economic terms
- Develop maintenance plan

Contact Information

Daniel W Gillman
Office of Survey Methods Research
www.bls.gov/osmr
Gillman.Daniel@bls.gov
202-691-7523
Team Leader, Taxonomy and Lexicon Team
TaxonomyLexiconTeam@bls.gov