Building a Taxonomy and Lexicon of Terms and Concepts at BLS

Daniel W. Gillman, Elizabeth Ashack, Daniel Chow, Ronald Johnson, Karen Kosanovich, Nicole Nestoriak, Ann Norris, Garrett Schmitt, Thomas Tedone, Clayton Waring

US Bureau of Labor Statistics; 2 Massachusetts Ave, NE; Washington, DC; 20212

Introduction

The US Bureau of Labor Statistics (BLS) strives to meet the information needs of its widely varying customers by measuring labor market activity, working conditions, and price changes in the economy. To meet this objective, in part, BLS is improving the presentation and delivery of its data and products. These improvement activities include the development of a taxonomy and lexicon of terms and definitions that are used to name and describe BLS data.

BLS has provided several ways for users to access its data. There is an API (Application Program Interface), single and multi-screen data extraction tools, downloadable ASCII text files, and many pages of links to tables, reports, and other data. The single and multi-screen data extraction tools were designed for accessing data by subject matter. Different tools were built for each program within the agency. This has created a conundrum, as users are required to know the subject matter each tool addresses beforehand, and it is difficult to assess the applicability of data or make comparisons from several sources at once.

Now, the BLS is building an agency-wide dissemination tool for all time-series data. This requires an interface that provides users the means to differentiate among all the data the BLS produces. BLS began building a taxonomy of terms and definitions to solve this problem.

The paper contains a description of the team formed to build the taxonomy and lexicon, a description of the work completed so far, current efforts, and prospects for future work.

Taxonomy and Lexicon Team

The Taxonomy and Lexicon team (hereafter, the team) was chartered in June 2013. The team was originally composed of at least one representative from all of the BLS program offices and from administration, field, publications, and research. The current membership is about the same, but the administration office dropped out, as their focus was more on records management than on dissemination and web site design. Participation by as many offices within BLS as possible was the only way to ensure the legitimacy and buy-in of the work. As such, the inclusion of each of the four main program offices was considered vital.

The team's goal is to design and develop a taxonomy and lexicon (Soergel, 1974; ISO, 2011) of technical terms in support of data dissemination and document tagging. Web site redesign is an important adjunct to this. The taxonomy portion of the work supports the design and functioning of the user interface to an agency-wide time-series data-dissemination tool called Data Finder. The concepts and terms (ISO, 2000; ISO 2009) proposed by the team for the taxonomy will be implemented in Data Finder. The lexicon portion will provide keywords for tagging documents that BLS publishes. Searches for data organized via the taxonomy and documents tagged via the lexicon should produce similar results. This means that documents and data addressing the same subject should be able to be found using the same search terms.

The technical language BLS uses is not always clearly understood by the public. For instance, programs use terms such as "class of worker" and "worker status" when distinguishing between self-employed persons and persons employed by a separate entity for a wage or salary. Without additional context, a user, even using a dictionary, would not necessarily associate these terms with the distinctions that they encompass. In this instance, the team chose the more verbose but clearer heading "Relationship between Workers and Employers." Here and in other places, the taxonomy is being designed to include plain English words to help guide novice users who may not be aware of fine technical distinctions.

Since its inception, the team has worked in a series of phases, each of about a six-month duration. In the first phase, the team evaluated previous work at BLS and work done by other agencies. The team laid out the basic structure of

the taxonomy and initiated work to identify important plain English words. In the second phase, the team constructed the initial taxonomy from the documentation associated with downloadable ASCII data files. These documentation files serve as database mapping files, linking a particular data series to the attributes associated with it, such as whether it is seasonally adjusted or addresses a specific industry. The team completed the plain English word evaluation and incorporated the results into the titles of high-level categories in the taxonomy. A very important design decision at this time was to separate those terms that describe the data series presented, such as whether it is a count of persons or a productivity index (i.e., measures), from the terms describing the population to which the data series refers, such as the age of the persons counted or the industry whose productivity is measured (i.e., characteristics). The durations of each phase so far were relatively short, so the product at the end of Phase Two needed substantial revision. This has been the focus of the current third phase. These three phases are described in more detail below.

Phase One

The team was organized into two subgroups: a Plain-English subgroup and a High-Level Terms subgroup.

The Plain-English subgroup was formed to collect data from BLS staff whose duties include answering inquiries via phone or email from data users. The subgroup found that data users often think about data based on what they have seen or heard about them in the news media. Another key finding was that the data user's definitions of words and terms often don't match the definitions used by BLS. Additionally, the subgroup found that BLS acronyms can be problematic for data users.

The High-Level Terms subgroup identified broad concepts for locating information within the BLS. The team recommended that the plain-English high-level terms be organized into two sets (See Figure 1 below). The first set, referred to as Measure categories**,** consists of plain English words that are used to describe the types of data observed by BLS. The second set, referred to as Characteristic categories**,** consists of plain English terms used to describe the population for which data exist. Expressed differently, given some data for a population, its characteristics consist of the aspects of that population that are fixed in advance, whereas its measure consists of an aspect of that population that is open to observation and has not been fixed in advance.

Measure Subject Areas	Characteristic Categories
Jobs Employment Hours Wages Benefits Earnings Additional measures Prices Consumer Producer Wages Cost of Living Consumer Expenditures Inflation Additional measures People List of measures Employer List of measures Additional subject areas as needed	Industries List of industries Sectors List of sectors Commodities List of commodities Occupation List of occupations Labor Force Categories List of labor force categories (education, union, etc.) Demographics List of demographic characteristics Geographic Areas List of regions, states, counties, cites, etc. SOII Categories Lists of events, parts, natures, and sources Additional characteristic categories as needed

Figure 1: Initial High Level Categories 1

Phase Two

The team achieved several goals in this phase:

- Divided terms by whether they primarily address measures[1] or characteristics[2]
- Built hierarchy of Measure terms
- Built hierarchy of Characteristic terms
- Identified and mapped plain English equivalents
- All terms were mapped to dissemination database (called LABSTAT) series ID's

The team delivered two major products at this time:

1. Hierarchy and mapping of Measures and Characteristics terms to series ID's
2. Mapping of plain English words to Measures and Characteristics categories

The combination of these products formed the first version of the BLS Taxonomy of Terms.

During Phase Two, the team was specifically charged to produce a taxonomy of terms that could be incorporated into the Data Finder user interface. This taxonomy included plain English words users often have in their minds when looking for data at BLS, as users of BLS data often do not know BLS-specific terminology used to refer to BLS data.

While many significant elements of BLS terminology were mapped to plain English equivalents and added to the taxonomy, the team postponed producing a full lexicon until the end of Phase Three. That phase included a substantial effort to improve the quality of the taxonomy. Since the lexicon will be generated from the taxonomy, improving the quality of the taxonomy before generating the lexicon will save work. The taxonomy also received priority because of its prominent incorporation in the BLS Data Finder tool, which at the time underwent accelerated development.

In this second phase, the team divided into two subgroups: 1) Measures and Characteristics; and 2) Plain English.

The Measures and Characteristics sub-team took the result of the first phase and expanded it. The main work of the sub-team was to use files describing data in LABSTAT to test and refine the proposed structure. The team refined the taxonomy consisting of two separate dimensions, Measures and Characteristics, to classify BLS data products. Each dimension had several large groups of concepts, with examples of specific categories within them. The goal was to use plain language to label each level of the taxonomy. As a result, the sub-team made changes and additions to the taxonomy to better tailor it to the diversity of BLS programs.

Multi-level hierarchies were developed for both Measures and Characteristics. The levels consisted, in increasing order of specificity, of "groups", "categories", and "category details". The lowest level was mapped to the attributes used to describe data series in each program's published database. This connected the taxonomy to every time series published by BLS in its LABSTAT database. Much work was devoted to developing the categories, which exist at an intermediate level of the hierarchy used in the taxonomy. These categories represent natural groupings, and choosing categories that could both encompass data from multiple BLS programs and divide intelligibly and meaningfully that data for the public proved challenging. Both the words used to label categories and the content of the categories themselves are open to revision. The other time-consuming aspect of the work was linking terms to the individual data series via the attributes provided in their associated mapping files. Because the purpose of this phase was to produce a first draft of a taxonomy that could function within the version of Data Finder under development, the linkages were an important part of the work.

As described above, each Dimension was broken down further into Groups (as shown in the table below), Categories, and Category Details. As an example, under the Dimension of "Measures" there is a Group called "Prices", under "Prices" there is a Category called "Consumer Prices and Inflation", and under "Consumer Prices and Inflation" there is a Category Detail called "Department Store Inventory".

[1] Measure – an estimate on some population
[2] Characteristic – a means of partitioning or stratifying a population

Under the Dimension of "Characteristics", one traces a path from "Geography" to "Region and Division" and finally to "Midwest". The structure uses plain English labels and seeks broad applicability so as not to split similar concepts that happen to exist separately among different programs because of relatively subtle technical and methodological distinctions. However, the sub-team did not dwell on the choice of labels, instead focusing on the usefulness of the groupings of concepts underneath them. See Figure 2 below.

Measures	Characteristics
Jobs	Occupation
People	Geography
Employers	Demographics
Prices	Worker Characteristic
Others	Unemployment and Not In Labor Force
	Worker Injury and Illness
	Industry
	Establishments/Businesses/Firms
	Products/Commodities/Services
	Other, Misc., and N.E.C.

Figure 2: Measures and Characteristics – Top Levels

The sub-team obtained the mapping files for all LABSTAT databases, both current and discontinued. Mapping files describe every data attribute for a given LABSTAT series. An attribute is one facet by which a data series is described. For example, a data series with an industry attribute will be associated with a mapping file showing each possible industry and the label for each.

The number of attributes varies greatly by program. For example, the Current Employment Statistics (CES) database has only a few attributes, such as data type, industry, super-sector, and whether the series is seasonally adjusted. Others, such as the Current Population Survey (CPS) database, have many more attributes. While some differences such as that described above originate in the nature of the data and what these data represent, other differences derive from the particular choices each program has taken in structuring its data. There is no standard way to structure the attributes.

The sub-team met regularly to discuss and agree upon changes required in the repertoire of Categories and Category Details to conform to available data. The sub-team then determined how to aggregate the lower-level Categories within higher-level Groups such that a customer could locate a relevant topic with a minimum of uncertainty. Some Categories fall under multiple Groups. For example, within Measures, the "Work Hours" Category appears under both the "Jobs" and the "People" Groups.

The plain English sub-team continued interviewing staff in the Regional Offices and the program offices to uncover English words commonly used by BLS data users. At the end of Phase One, the Plain English sub-team had talked to only a handful of BLS national office programs and regions and had recorded the findings from each one in its own document. By the end of Phase Two, all the Regional Offices and program offices gave feedback to the sub-team. Not all the information obtained was useful for the development of the taxonomy, and a substantial part of the work was to assess which findings were relevant. Most importantly, not only do many users not know BLS terminology, they are confused by the terms they hear. The sub-team mapped all plain English words mentioned in the interviews to the Measures and Characteristics categories built by the other sub-team.

In addition to recording plain-English synonyms used for BLS data series, data characteristics, or concepts, the sub-team sought to understand larger areas of confusion. When speaking with program staff, the sub-team asked about "problem" terms, language that caused confusion among users, as opposed to requests for known and understood data series. For example, a question such as "What was the national unemployment rate in August 2014?" would not be noted, whereas the question "What is the labor force? Does this include x, y, and z persons?" would be recorded. An example of this kind would let the team know that the definition of "labor force" causes confusion.

Another important cause of confusion is the distinction between industry and occupation. Often, users call to ask about data pertaining to their "field of work," rather than using the terms "my industry" or "my occupation". A user requesting occupational data for nurses might use the term "nursing industry" when he or she really means nurses as captured by the Standard Occupational Classification System (SOC), not the health services industry sector as captured by the North American Industrial Classification System (NAICS). With this information, the Plain Language sub-team recorded any terms discussed and added notes, if needed, about the terms.

The sub-team had lengthy discussions based on the interviews with the program staff to determine which Measures and Characteristics categories should be associated with each common term. In doing this, the sub-team kept the mindset of a non-expert BLS customer. While it is easy to review the common words list and immediately link familiar BLS concepts with their specific data series or program office, the sub-team had to work harder to ferret out associations for broader BLS concepts that cross multiple programs (e.g., wages, earnings, and benefits data).

By the end of Phase Two, the taxonomy contained a significant number of plain English words. It is not purely a taxonomy of technical terminology.

Phase Three

The third phase is devoted to improving the quality of the taxonomy produced in Phase Two. This phase consists of several parts: initial quality improvement, internal BLS review, cognitive testing, final review, and adjustments. There may need to be subsequent review phases in the future.

The initial quality improvement step has been finished. The team focused on the hierarchies under the Characteristics dimension of the taxonomy. At the end of the second phase, the taxonomy appeared as in Figure 3, and the next levels of Geography appear in Figure 4.

Taxonomy and Lexicon Outline

Measures

- **Jobs**
- **People**
- **Employers**
- **Prices**
- **Measure Survey and Attributes**

Characteristics

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

Figure 3: Taxonomy – Top Level

Each category under Characteristics needed significant work to improve the quality of that section. For the Geography section, the team focused on the need to create fewer choices at each level, reduce redundancy, and increase meaningfulness. Figure 4 shows the categories and detail under Characteristics - Geography - "Region and Division". The detail shows a lot of overlap between choices. For example, Census Regions and Divisions are listed together, even though each Census Division is entirely subsumed by a broader Census Region. The organization depicted in Figure 4 does not take advantage of this structure.

Geography
- Nation
- Region and Division
- States and territories
- County and Equivalents
- Cities and Metro Areas

==========

- Region and Division
  - East North Central
  - East South Central
  - Middle Atlantic
  - Midwest
  - Mountain
  - New England
  - North Central
  - Northeast
  - Pacific
  - South
  - South Atlantic
  - West
  - West North Central
  - West South Central
  - Other

Figure 4: Categories under Geography and Detail under Region and Division

Figure 5 shows how the Geography section has been reorganized. In particular, now the Regions and Divisions defined by the Census Bureau are made distinct and contained in a category called "Metro and Other Statistical Areas."

Geography
- Nations
- States and Territories
- Counties and Equivalents
- Cities and Towns
- Metro and Other Statistical Areas
  - Census Regions
  - Census Divisions
  - Combined Statistical Areas

Metropolitan Statistical Areas
Micropolitan Statistical Areas
Metropolitan Divisions
Special Areas

Figure 5: Geography, reorganized

Now geography is subdivided consistently by type, separating jurisdictions from statistical areas and separating in turn the different types of jurisdictions and statistical areas from one another. There is an organizing principle that is in use now here. A taxonomy built around organizing principles is easier to build, use, and maintain (Hlava, 2015). An overriding concern among the team is the needs of the user looking for BLS data.

Another addition is the "Special Areas" category. The BLS defines many areas that do not fall under one of the standard geographic subdivisions maintained by the Office of Management and Budget (OMB). Without a very complex hierarchical structure to define all these areas, they could not be found under the original design. One example is "Northeast Alabama non-metropolitan area." This was put under State under Special Areas. Another example is "Garrett County MD non-metropolitan area". This was put under County and Equivalents under Special Areas. Again, there is a principle in use. Both areas are classified as special areas, because they do not conform to either a jurisdiction or an OMB-standardized statistical area. The essential geographic region, state for the first example and county for the second, is how these areas are classified under Special Areas.

The team found all other categories under Characteristics, e.g., Industry, Occupation, Products/Commodities/Services, also in need of significant reorganization. This was the main objective of the first section of the Phase Three work. For example, even though the BLS follows and uses NAICS and SOC, each program may have special needs for reporting data that deviate from those classifications. For instance, the CES program publishes data to a Government category. However, Government is not a category within NAICS. Factors such as this have added significant complexity to the work.

The team found many other areas of difficulty that further complicated its work. Important examples of these are listed here:

- Characteristics such as Industry are represented using several classifications. For Industry this includes NAICS 2007, NAICS 2012, and Census 2000.
- Different programs do not always incorporate new versions of classifications at the same time. Thus, NAICS 2007 and NAICS 2012 are both currently in use.
- Different programs do not necessarily report data at the same level of a particular classification. For instance, the Occupational Employment Statistics (OES) program reports data to the lowest level of SOC, but the Census of Fatal Occupational Injuries (CFOI) does not.
- Many programs report data in categories that are not part of a standard. As noted above, CES reports Government data, but NAICS has no category titled Government. OES reports data for "Central Colorado Non-metropolitan Area," but this is not a standard geographic area managed by OMB.
- Classifications are revised on different dates and for different intervals.
- Incorporating several versions of some classification requires making changes to the taxonomy whenever a new version is put into use, and this includes adding cross-walks between the newest version and older ones.

Each of the issues described above points to the need to make compromises in the development of the taxonomy. Rather than being the authoritative source of all BLS technical language, the taxonomy will represent a common understanding of programs, data, and classifications used throughout the agency from a non-technical point of view.

Future

As indicated previously, the team is primarily tasked with building a taxonomy and lexicon for use in finding data and related documents on the BLS web site. However, the organization of the taxonomy into two main facets – Measures and Characteristics – leads us to conclude that the same design can be applied to the web site.

Currently, the BLS web site is organized mostly around statistical programs (e.g., surveys) and measures. For instance, one could theoretically find all data about Boston; however, it would be extremely time consuming. The same problem arises when one wants all the data about some industry, occupation, product, or any other item in a classification.

Reorganizing the web site to account for all the characteristics will be a big job; however, the taxonomy will provide a guide as to how to achieve this.

Another consideration is the long-term maintenance of the taxonomy. One design path the team could have gone down was to include every version of every classification in use to describe data, past or present, as discussed in the section on Phase Three. This would have meant including crosswalks between versions to tie all the classifications and terms together, and many published statistics, particularly aggregates, under one classification system have no true analogue in another related classification. Then, the staff would need to devote additional work to incorporate any changes in the future. This solution is not practical, and in all likelihood would take many years.

Instead, the team chose to concentrate on data that are currently being produced and map categories from past collections into the taxonomy as best as possible. As a consequence, Industry and Occupation sections under Characteristics will not be organized to look like NAICS and SOC. This is consistent with the discussion about Industry and Occupation in the Phase Three section above.

References

1. ISO 1087-1 (2000) *Terminology work – Vocabulary - Part 1: Theory and application*. ISO, Geneva
2. ISO 704 (2009) *Terminology work - Principles and Methods*. ISO, Geneva
3. ISO 25964-1 (2011) *Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval*. ISO, Geneva
4. Hlava, M.M.K. (2014) *The Taxobook – Principles and Practices of Taxonomy Construction*. Morgan and Claypool
5. Soergel, D. (1974) *Indexing languages and thesauri: Construction and maintenance*. Wiley Information Science Series, Los Angeles