Creation of a Web-Enabled Framework for a Taxonomy of Terms at the Bureau of Labor Statistics November 2017

Daniel W Gillman[1], Randall Powers[1]
[1]US Bureau of Labor Statistics, 2 Massachusetts Ave, NE, Washington, DC 20212

Abstract
This paper contains preliminary results and plans for a research project at the Bureau of Labor Statistics (BLS). BLS is building a taxonomy of terms for describing time series data to be used as the content of the search section of a new DataFinder tool for all BLS time series. The work reported here details the effort to use the precepts of Linked Open Data and its associated standards to build a semantic query engine and visualization suite for BLS time series. Goals include a determination whether using Linked Open Data will be an effective strategy for constructing tools for locating and downloading BLS time series. Progress made so far and plans for the future are discussed. The effort reported here is a research project only and is not in competition with the DataFinder development.

1. Introduction

The US Bureau of Labor Statistics[1] (BLS) is attempting to modernize and consolidate its services for disseminating data to the public. Currently, BLS provides and maintains separate time series data access tools for each subject matter area[2]. These tools are customized for the measures within each area. There is no way to download two or more series at once except within one subject area.

The BLS is developing the DataFinder[3] system, which will provide one tool for access to all time series data the agency offers. The DataFinder will provide the ability to download many series at once, however the much more complex problem is how to guide users to the appropriate data through a user interface. BLS data cover many subjects, and most measures are stratified through several dimensions. An example is the Consumer Price Index, which is reported for the US. However, it is available for many MSAs (Metropolitan Statistical Areas[4]) and products and services.

To accomplish the goal of organizing all the subjects around the measures and the various dimensions for stratifying them, the BLS is building a taxonomy[5] of terms. This taxonomy is being incorporated into the user interface for DataFinder. The taxonomy is still under development, but a draft of the final version is expected at the end of the 2017 calendar year.

As a research effort at BLS, and not in competition with the development of DataFinder, some staff (the authors) are looking to transform the taxonomy for use within the Semantic Web[6] using Linked Open Data (LOD). LOD is a technique popularized by Tim Berners-Lee, the inventor of the World Wide Web

[1] https://www.bls.gov
[2] https://www.bls.gov/data/
[3] https://beta.bls.gov/labs/
[4] http://www.w3.org/standards/semanticweb/ https://www.bls.gov/lau/lausmsa.htm
[5] https://www.cbd.int/gti/taxonomy.shtml
[6] http://linkeddata.org/

(hereinafter, the Web). Through the use of the Resource Description Framework[7] (RDF), a standard for adding meaning to the hyperlinks in the Web, LOD establishes meaningful relationships among a variety of data and other resources. For the purposes of the work reported in this paper, the idea is to link the terms in the taxonomy to specific measures the BLS produces.

The paper, then, contains an overview of LOD including RDF, a description of the taxonomy, a short synopsis of R-Shiny[8] (to be used as the basis of the user interface of the resulting system), and an outline of the work plan and summary of the effort so far. Again, the paper here reports on a research effort, the work is not yet complete, so this is an interim report.

2. Linked Open Data

Linked Open Data (LOD) is a method for structuring, linking, and querying data published on the Web. LOD methodology includes the use of some W3C (World Wide Web Consortium) and Internet standards. This provides a uniform way to represent and query data that are linked under LOD principles. The most important of these standards is RDF, a model based on a very simple logic using a sentence structure comprising a subject, predicate, and object. Subjects are Web resources that are linked through Predicates to Objects, which are either Web resources (therefore can serve as Subjects for new triples) or contain literal content, such as data. The simplicity of the model gives it power and makes it able to represent any kind of information, for example statistical time series data. Each subject-predicate-object sentence is commonly referred to as a "triple", and a database of linked triples is called a "triple store". The model used by RDF is also commonly referred to as a "graph model[9]" consisting of "nodes" (the subjects and objects) and "edges" or "arcs" (the predicates).

The RDF model is very basic and extremely powerful. It formalizes any information in a uniform way. This facilitates the linking both between and within data sets. Links across data provided by different publishers is possible. Each resource is uniquely identified on a global scale and referenced rather that copied as specific context. The subject – predicate – object model also provides a simple but strong logical underpinning, which allows for semantic query and inference of new facts.

See Figure 1 for an example of an RDF graph:

[7] https://www.w3.org/RDF/
[8] https://shiny.rstudio.com/
[9] http://infolab.stanford.edu/~ullman/focs/ch09.pdf

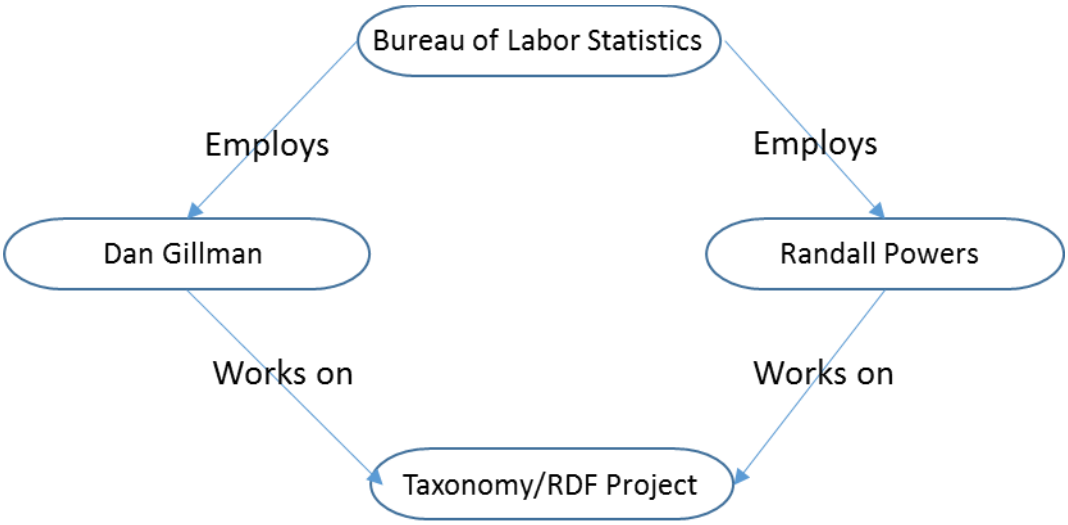

Figure 1: RDF Graph

Figure 1 depicts four triples, and as English sentences, they read as follows:

1. The Bureau of Labor Statistics employs Dan Gillman.
2. The Bureau of Labor Statistics employs Randall Powers.
3. Dan Gillman works on the Taxonomy/RDF project.
4. Randall Powers works on the Taxonomy/RDF project.

Note that sentences 1 and 3 can be meaningfully combined, as can sentences 2 and 4. This affordance is shown graphically through the fact that the objects in sentences 1 and 2 are the subjects in sentences 3 and 4, respectively.

Note, also, in Figure 1, two triples share the same Subject, and two triples share the same Object. This illustrates the inter-connectedness power of RDF.

Now, it needs to be noted, Linked Open Data is not a technique for just linking one datum to another. Instead, linking metadata and including links to the data is necessary. See Cotton and Gillman (2017) for a detailed description. In general, though, RDF graph conveys meaning through the content of the subjects and objects and through the predicate, or relationship, between them. But, if the subject and object are not data of interest, then they are descriptive of those data, because they are related to the data. This means the RDF graph contains and links metadata.

3. Taxonomy

The project to build a taxonomy of terms describing time series data at BLS began in 2013. The initial stated purposes of the project are for the taxonomy to serve as the user interface for a new agency-wide series dissemination tool called DataFinder and, as a flattened lexicon derived from the taxonomy, be used as the means to tag documents, reports, and Monthly Labor Review[10] articles consistent with the data to which those terms apply. As the user interface for DataFinder, it will be necessary to include plain English words at the higher levels to guide non-expert users to the technical terms describing the data they want.

[10] https://www.bls.gov/mlr/

As the work progressed, two other major applications arose. One is for the taxonomy to serve as a resource for understanding BLS technical terminology. A glossary is being built for providing definitions, but how terms relate to each other is a purpose of any taxonomy. Providing these relationships allows users to make distinctions between similar but different ideas.

The other new purpose is to guide a reorganization of the BLS web site. Currently, the web site is mostly organized by the internal structure of BLS. It is possible to find data based on some broad ideas, such as employment, but finding data on specific notions such as “Boston” or “nurses” is not so easy. Worse, even if some data are found, say, for Boston, there is no way to know if they are all the data BLS has on the subject.

The project is also building on some previous work done at BLS. In 2010, a team looking at BLS data in general, found that time series data are a combination of a measure, i.e., a quantitative estimate, some combination of characteristics, where each characteristic is a way to specialize data, and time.

More specifically, a measure is a quantitative estimate of some population. BLS produces several thousand of these. Many of these are point-in-time estimates, produced on a regular schedule, and can be compared from one iteration to the previous. A series of estimates for the same measure separated by a given time interval is a time series.

Measures can often be stratified through dimensions called characteristics. Examples of these are geography, industry, and occupation. Through the selection of categories, one from each of several characteristics (often classification schemes), the population on which the estimate is made is specialized. For example, the Consumer Price Index (CPI) can be found for each of the larger MSAs and for each of the many products and services within the US. For example, there is a CPI for clothing in the Washington, DC MSA.

In the taxonomy project, the development team identified all the characteristics relevant to describing time series data and organized them into a set of independent hierarchies. For the most part, any category within one characteristic will not be found in another. There are two major exceptions. One is the close relationship between industry and products and services. The other is the confusion among various levels of geography. Cities and MSAs are often referred to by the same name. For instance, a user could encounter six different uses of the term Boston describing BLS data.

There are 10 separate characteristics identified through the work so far. All are structured as hierarchies, and some are multiple ones. The set of characteristics consist of the following:

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

The Demographics, Workers, Establishments, and Workplace Illness and Injury characteristics each contain several hierarchies within them.

For the measures, the work is not as straightforward. The team has decided to try to treat each measure similarly to a variable. This means it represents a set of units, specified by a unit type. For example, person, establishment, and worker are unit types. A measure is also a characteristic (in a different sense!) of the unit type. We will call this a measure characteristic. So, for instance, *average hourly wage* is a measure characteristic of the unit type *workers*.

Some of the measures BLS produces includes separable information in the description of the measure characteristic. In particular, measures of benefits often include the type of benefit in their descriptions. However, the same thing is being measured each time, for instance *percent of establishments offering benefits*. Obviously, here, the measure characteristic applies to the unit type *establishments*.

4. R-Shiny

The R-Shiny package combines the R[11] environment with the modern Web. It can be used to build Web applications using R and turn analyses into interactive applications. Knowledge of underlying Web languages, such as HTML or JavaScript, is not required. R also provides a package for SPARQL[12] (SPARQL Protocol and RDF Query Language), the Web query language for RDF triple stores.

The idea is to use R-Shiny to build the infrastructure around a system that allows a user to query the taxonomy (through SPARQL package for R). The result will be a well-defined series identifier that will be used for downloading a series from the BLS LABSTAT[13] database. Then, this data will be available for input into any of a set of visualization tools.

5. Work Plan

The work for the project will be done in several steps. Currently, as it is being developed and modified, the taxonomy is being managed in a MS Excel spreadsheet. This choice, though not optimal for taxonomy management, was an expedient resulting from availability, cost, time constraints, and team expertise.

The main steps are briefly described here:

1. The columns in each row of the spreadsheet contain terms in the path in the taxonomy to series terms. The series terms are technical language the BLS program offices use to name a series. A series may have many paths to it. By sorting the rows based on the values in the columns, we are able to develop all the triples implied by the taxonomy. Each triple is generated from the terms consecutive columns, but the resulting predicate is not meaningful. This will be produced later.

2. The resulting set of triples is first used to create single Web pages for each term. The parent term and any children terms are provided with hyperlinks to their pages. The hyperlinks correspond to the predicates. This set of Web pages is a simple navigational tool for the taxonomy. Creating pages this way allows us to uniquely identify each term in the taxonomy. This could be done more

[11] https://www.r-project.org/
[12] http://www.w3.org/TR/rdf-sparql-query/
[13] https://www.dol.gov/oasam/ocio/programs/PIA/BLS/BLS-LABSTAT.htm

directly, but we want render the taxonomy as a set of Web pages as its own resource. This step represents the work completed so far, and plans were made to continue with the project.

3. Next, we translate each of the web pages into TURTL[14] (Terse RDF triple Language) file using the SKOS[15] (Simple Knowledge Organization System) vocabulary. TURTL is a syntax for representing RDF triples. SKOS is a set of elements laid out in TURTL for representing a taxonomy (a knowledge organization system). Here, the content of each Web page will be added to the TURTL file.

4. Following this, we build an engine for generating BLS series identifiers from the series terms found through queries of the taxonomy. The combination of a measure term and any appropriate characteristics terms is necessary to generate the series identifier. For instance, the terms "Consumer Price Index", "Washington, DC MSA", and "Apparel" is enough to generate an identifier to obtain the series for the CPI for apparel in Washington, DC. To illustrate, this series identifier is CUURA311SAA, were CU refers to Consumer Price Index, U refers to not seasonally adjusted, SA311 refers to Washington, DC, and SAA refers to Apparel.

5. After this, we set up the SPARQL system (building a SPARQL endpoint, as it is called) for generating the queries to extract the terms naming each series. It should be noted that these series names could be series identifiers, but many of the terms have associated codes, so the series identifiers have codes in them instead of terms. Then, we tie the results of the SPARQL queries to the series identifier generator.

6. Using the series identifier and the BLS API[16] (Application Programmer Interface), we set up a system to download any data series. We save the files in a cart managed by R.

7. Finally, we create a switchboard in R-Shiny allowing the user to select a visualization tool to display the selected data sets.

There are many steps in the process, and the resources we have are very limited. As shown above, much of the work remains to be completed.

6. Conclusion

This paper contains the report of an effort to turn the taxonomy of terms under construction at BLS into an RDF database capable of being searched and able to produce the information needed to access particular BLS time series data. This is purely an experimental effort, and it is not to be confused with the effort to build the DataFinder data query tool.

There are several goals the work is expected to achieve:

- Provide the authors with experience in semantic web and LOD technology, including TURTL, SKOS, and SPARQL

[14] http://www.w3.org/TeamSubmission/turtle/
[15] http://www.w3.org/2004/02/skos/
[16] https://www.bls.gov/data/#api

- Determine if the RDF approach to finding time series data at BLS can be made to work and is effective
- Gain experience building applications using R-Shiny
- Build more user-friendly applications for displaying the taxonomy

References

Cotton, F., and D. Gillman (2017) Linked Open Statistical Metadata. In T. Prodromou (ed.) *Data Visualization and Statistical Literacy for Open and Big Data* (pp. 297-320). Hershey, PA: IGI Global.