LEHD Public Use Data Schema Versioning

Lars Vilhuber, Heath Hayward, Matthew Graham ces.qwi.feedback@census.gov 2016-May-27

Introduction

The Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program will release a new schema for much of the data published by the Program. In this paper, we describe scope, history, and details of the schema, provide examples on how to use the machine-readable features of the schema as well as the user-accessible documentation of the schema, and provide an outlook onto future changes. All examples and files referenced can be found at http://lehd.ces.census.gov/data/schema/, unless otherwise noted.

Scope

The public-use data from the Longitudinal Employer-Household Dynamics Program, including the Quarterly Workforce Indicators (QWI) and Job-to-Job Flows (J2J), are available for download according to structural and file naming schemata. The data themselves are available as Comma-Separated Value (CSV) files through the LEHD website's Data page at http://lehd.ces.census.gov/data/ as well as through the LED Extraction Tool.

History

The first published schema for the Quarterly Workforce Indicators (QWI) was v3.5, used for QWI files releases through R2013Q1. No formal document describing the schema was released, but a user-contributed "Cheatsheet" was available. A restructuring of the data and file naming conventions lead to V4.0 for releases starting with R2013Q2. The newer schema was described in the form of a PDF document that was occassionally updated to reflect corrections and enhancements. All data releases were accompanied by a set of CSV files for allowable values of variables and flags, accompanying each collection of tabulations for each state.

Starting with release R2015Q2, a more formal, flexible, and machine-readable structure was implemented, and published as V4.0.1. As changes occur to elements of the schema, version numbers are incremented (see Versioning). Broader changes are first published as draft schemas (typically used by draft or "beta" releases of data), before becoming finalized. All versions are retained on the LEHD schema archive at http://lehd.ces.census.gov/data/schema/.

- v3.5 First documented schema
- V4.0 Second documented schema, change in file naming conventions; added and dropped variables.
- V4.0.1 First formally structured schema documentation of V4 schema.
- V4.1 Additional files and variables (not finalized as of May 2016)

Usage

Each *data* release is accompanied by a file specifying a compact notation for metadata. For instance, the R2015Q2 release of Missouri QWI by race and ethnicity for all firm types (archived here or here) would have a file called *version_rh_f.txt* with the following content:

```
QWIRH_F MO 29 1995:1-2014:3 V4.0.1 R2015Q2 qwipu_mo_20150601_1902
```

The fifth component (V4.0.1) identifies the version of the schema being used. Thus, all value labels, the naming and structure of the files, the geographic and industry coding vintages, etc. can be deduced from the information available in the V4.0.1 directory.

Names of data files follow certain rules, which are documented in the file "lehd_csv_naming". (Note: all references below are relative to http://lehd.ces.census.gov/data/schema/)

- In the example above, the file can be found at V4.0.1/lehd_csv_naming.html.
- The latest version of "lehd_csv_naming" can be found at latest/lehd_csv_naming.html.

For each **identifier variable** on the data file, a set of allowable values is defined. Definitions of allowable values are provided as CSV files, with headers. Available **indicator variables** are defined, and labels provided. These definitions are summarized in the file "lehd_public_use_schema" (formerly named "QWIPU_Data_Schema.pdf").

- For instance, in the example above, the file can be found at http://lehd.ces.census.gov/data/schema/V4.0.1/lehd_public_use_schema.html
- (QWI) **indicators** are named and listed in the section "Indicators" in V4.0.1/lehd_public_use_schema.html in human-readable form, and as machine-readable CSV files at V4.0.1/variables_qwipu.csv .
- **Identifiers** for the QWI files are listed in the section "Identifiers" in machine-readable form, and as CSV files at V4.0.1/lehd_identifiers_qwi.csv (note: different files may have different identifiers).
- One of the available **identifiers** is "agegrp", for which the allowable values are listed in the "agegrp" section and provided in machine-readable form at V4.0.1/label_agegrp.csv
- The latest version of "lehd_public_use_schema" can be found at latest/lehd_public_use_schema.html

Versioning

Versioning rules follow Semantic Versioning V2.0.0, which states that

> Given a version number MAJOR.MINOR.PATCH, increment the:
>
> - MAJOR version when you make incompatible API changes,
> - MINOR version when you add functionality in a backwards-compatible manner, and
> - PATCH version when you make backwards-compatible bug fixes.

In practice,

- LEHD increments the major number when a new data format is used that would break import procedures by outside systems (variables are dropped, are in a different order, existing variables change names; file naming conventions change for existing files)
- LEHD increments the minor number when
  - variables are added, without changing order of existing variables
  - new types of data are added (e.g., J2J, LODES) without changing existing files
  - changes in values are of a "significant" nature
  - changes to the structure of the schema documentation are made
- LEHD increments the "patch" number when changes are made to existing codes that do not break import of data, or change the interpretation of the data in a significant way
  - a description is corrected
  - a set of value labels is changed in a minimal way

- LEHD does not increment the version number when corrections to the human readable schema documentation itself are made, but does indicate such changes in the CHANGE section with the calendar date of the revision.

Examples of "patch"-level changes are:

- updated geography definitions (changes in state-specific geographies impacting a small set of areas, for instance a Workforce Investment Board (WIB) or a small number of counties) (see CHANGES in V4.0.1, V4.0.2, V4.0.3 for examples)
- change in North American Industry Coding System (NAICS) coding affecting only a small number of industries (see CHANGES in V4.0.2 for an example).

Switching from Standard Industrial Classification (SIC) to NAICS would have been a *major* version number change, changing from NAICS 1997 to 2007 - which had more significant changes, but did not fundamentally change the way the data are read in - would have been a *minor* version number change.

Additional revisions within a "patch"-level schema will be identified in the CHANGES.txt by date, but will not otherwise carry a different version number. Revisions are only used to correct for bugs, and to improve documentation of the schema itself, but not to change the schema.

Draft Versions

LEHD will publish a draft version of minor or major schema changes, in order to be able to allow for comments by the community. A draft schema may also accompany *beta* data products, where both schema and data are published to elicit comments from the public. Draft versions do not necessarily lead to a final specification, and should be treated as work in progress.

Most Current Version

For convenience, the latest non-draft version is accessible at http://lehd.ces.census.gov/data/schema/latest/. However, users should note that at any point in time, data published by LEHD may reference an older schema, as noted in the Usage section above. Users are strongly encouraged to reference a well-specified revision number in their programs, derived from the "version*txt" file provided with each data release.

Curation

LEHD commits to keeping a public record of all major, minor, and patch versions of the schema in an accessible, public location (currently, at http://lehd.ces.census.gov/data/schema/). Additional revisions are stored internally in code versioning systems, and can be provided upon request.

Upcoming Version

The next minor version as of the writing of this paper will be V4.1. Draft versions of V4.1 were circulated and made available since 2015 (see V4.1-draft, V4.1b-draft, V4.1c-draft, and V4.1d-draft.) Notable changes in V4.1 are the addition of J2J schema details, the expansion of QWI schema details to account for National QWI and federal worker data, the addition of variability measures (available initially in the beta National QWI data distribution only), and the addition of shape (SHP) files. Of course, these changes will be explicitly confirmed in the V4.1/CHANGES.txt file once it is finalized. Note that the availability of *metadata* on a particular data product does not convey about availability of the actual data product. In fact, in general, we strive to release the specifications of upcoming data products before the data product itself is available, to allow the user community to prepare for it.

Outlook

Currently, the schema covers QWI and J2J data publications. We continue to work on expanding the schema to cover additional LEHD data publications, such as the LEHD Origin-Destination Employment Statistics (LODES).

Looking further ahead, the current schema is in format that is unique to LEHD. We consider the current schema structure to be an intermediate step towards a standards-compliant schema. LEHD continues to work towards creating a schema that is fully compliant with open data standards.