LEHD Public Use Data Schema Versioning

Contents

(Printable version)

> **!** **Important**
> This document is not an official Census Bureau publication. It is compiled from publicly accessible information by Lars Vilhuber (Labor Dynamics Institute, Cornell University). Feedback is welcome. Please write us at lars.vilhuber@cornell.edu.

1 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 schema. 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.

For each variable, a set of allowable values is defined. Definitions of allowable values are provided as CSV files, with header variable definitions. The naming conventions of the data files is documented separately, in a similar fashion.

2 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 and flexible 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 this server.

3 Usage

Each data release is accompanied by a file specifying a compact notation for metadata. For instance, the R2014Q4 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: where the fifth component (V4.0) identifies 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 directory.

4 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 drop import 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 WIB or a small number of counties) (see CHANGES in V4.0.1, V4.0.2, V4.0.3 for examples) * change in NAICS coding affecting only a small number of industries (see CHANGES in V4.0.2 for an example). Switching from 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.

5   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.

6   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.

7   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.

8 Changes

This section is reserved for documentation of changes to the versioning schema itself, as well as for description of major version changes related to the schema documentation.

- link:v3.5 First documented schema

- link:V4.0 Second documented schema, change in file naming conventions; added and dropped variables.

- link:V4.0.1 First formally structured schema documentation of V4 schema.

- V4.1 Additional files and variables (not finalized yet)

This revision: Tue Mar 15 17:09:19 EDT 2016