

# QWIPU Restructuring

---

## *Quarterly Workforce Indicators Public-Use Data (QWIPU) - Data Schema Modifications*

### Overview

The complete set of Quarterly Workforce Indicators are made available through a series of CSV (comma-separated values) files, which are provided to state partners and posted to the Cornell Virtual Research Data Center (VRDC, see <http://vrdc.cornell.edu/qwipu>) for general use.

The original set of tabulations contained 22 separate files for sex by age, with each file providing detail at a selected level of ownership, industry, and geography. This original set of 22 was tripled with the addition of race, ethnicity, and education tabulations. More recently, a new set of tabulations has been released to provide information on firm age and firm size, bringing the total number of files to 102. It has become increasingly difficult to adapt the original file organization and naming conventions to an expanded set of tables. For instance, in the original design each file had its own layout, including only variables relevant to the aggregation level provided on the file, resulting in 102 different data layouts.

In order to increase the usability of the files, we have developed a redesigned schema in order to make the data more accessible. The official release of the new layout will be in data produced for the second quarter of 2013. In the meantime, example QWIPU files in the new schema will soon be available through the VRDC and the complete set of QWI tabulations will be available with the Beta release of the LED Extraction Tool in late March. These Beta QWIPU data will be provided so users can make any necessary adaptations to existing databases to accommodate the new files when they become officially available.

## Summary of changes

---

### CSV File Naming Conventions

The original naming scheme developed for the QWI files could not be easily adapted to accommodate the range of new QWI offerings developed over the past few years, including new demographic and firm-level characteristics. One issue in particular that we have encountered is that the file names do not all have the same number of components delimited by the underscore, making name parsing difficult. Some of the names also were getting quite long, which caused problems for applications with shorter limits on table names (commonly 32 characters). With this release, we have shortened the names, while at the same time giving them all the same structure to facilitate automated file handling.

## Additional State-level Files for NAICS Sector

The original set of 22 files included four tables containing state-level totals only for NAICS3 and NAICS4 industries (both private-only and all ownership categories). While the information contained in these files is redundant to the grand totals on the corresponding files at the detailed geography levels, the smaller state-only files are very convenient for some purposes. However, the earlier schema did not include a separate file for the NAICS sector totals at the state level. This release will include state-level NAICS sector files to complete the set, bringing the total number of files to 108.

## Categorical Variables: Consistent file layout

Each data row in every file now contains an entry for the complete set of categorical variables (ownership, geography, industry, firm age/size, and demographics), yielding a consistent layout across all files. This will make for easier read-in for users who are interested in accessing more than one file.

## Categorical Variables: Geography and Industry Variables

The previous unique identifiers for geography and industry, `qwi_geocode` and `qwi_industry_code`, will be deleted from the tables, as will the `naicssec/naics3/naics4/county/metro/wib` variables. Instead, the NAICS and geography codes will now be reported in variables labeled *industry* and *geography*, with separate flags added to describe the geographic and industry level reported (*geo\_level* and *ind\_level*). QWI geography codes will now match standard Census geography codes.

## Categorical Variables: Re-indexing of Race and Ethnicity Categories

In order to improve the correspondence with LODES dataset used for OnTheMap, the ordering of the race and ethnicity categories has been modified. To differentiate this new classification, the leading character has been changed to "A" for both race and ethnicity.

## Categorical Variables: Additional Education Category for "Workers Under 25"

In the original specifications for the education tabulations, a filter was applied to exclude individuals under the age of 25. These individuals are not included in tabulations by education level because a large number within that age group have not completed their education. In deleting these jobs, the grand total for the sex by education distribution has been less than the same total for sex by age, or race by ethnicity. In order to reduce the confusion stemming from these differences, we will include a separate education tabulation category for those under 25. The level of education will not be available for those jobs, but the grand total will be consistent with the other tabulations. Note: This additional category will not be available in the Beta data release. Users can expect this new variable in the 2013 Quarter 2 release.

## Removal of Data Labels

The previous CSV files contained full data labels for every categorical variable, significantly increasing the size of the files with duplicative information. These labels have been removed from the QWIPU CSV files. In its place, we will be providing a separate “label” CSV file for each categorical variable describing the meaning of each code.

## SAS Read-in program

---

Historically, we have provided a program for users with access to SAS software to read-in the CSV files. However, because each table previously had its own layout, the complete read-in program was very extensive, as every table required its own step. Since the structure has been standardized in this release, the SAS program has been greatly simplified to combine all the input files in a single data step. Users can now easily modify the program to restrict the number of files that are being accessed, as desired. The program also generates SAS formats for all the categorical variables, and users will have the option of including the data labels as separate variables (this feature is commented out by default).

## Feedback

---

Please contact us at [ces.local.employment.dynamics@census.gov](mailto:ces.local.employment.dynamics@census.gov) with any questions or comments about the new file structure.