Data Sources and Maintenance Notes

For questions about the portal, contact Jeremy Hodgson at [jeremy.hodgson@wgu.edu](mailto:jeremy.hodgson@wgu.edu).

## Institutions Pages

All of the data which backs the Institutions (Main) and Institutions (Specific) pages comes from IPEDS Access databases, available for download at: <https://nces.ed.gov/ipeds/use-the-data/download-access-database>. Each year has its own database, complete with many tables and an overwhelming volume of information. The querying/wrangling of the data into tables was done with much thought into the evolution of the platform and the needs of its users. The 4 tables which back the page include: ipeds\_static.csv, ipeds\_dynamic.csv, degrees\_main.csv, and degrees\_institution.csv.

**ipeds\_static.csv** is the simplest of the tables as it contains only data from the most recent year’s Access database (2020-21 at the time of the portal’s release) and all fields included in the query come from the same table. Using the Query Wizard in the create tab, create a table from HD20XX which includes: UNITID, INSTNM,STABBR, ADDR, CITY, ZIP, WEBADDR, OBEREG, SECTOR, ICLEVEL, CONTRL, HLOFFER, GROFFER, DEGGRANT, HBCU, MEDICAL, TRIBAL, LANDGRNT, INSTCAT, UGOFFER, CCBASIC, C18ENPRF, C18IPUG, C18IPGRD, C18UGPRF, LONGITUD, and LATITUDE. Rename LONGITUD and LATITUD as longitude and latitude either before or after you export the file as a CSV. This table needs to be replaced yearly with the one from the most recent Access database.

**ipeds\_dynamic.csv** was created by linking tables using UNITID as a key and then querying a variety of variables from each of them. Each year included in the dataset was its own table, so they had to be concatenated at the end, adding an additional year column to differentiate between them. This allows the different variables to be trended over time in the portal’s interactive visualizations. The code which was used to do this can be found in the Institution Page section of data\_processing.ipynb. You won’t have to go through this process again.

Only new data needs to be appended.Here are the table and field names within each table necessary to use the Query Wizard to create the table for a new year to append to the current dataset:

| **Table Name** | **Fields to Include** |
| --- | --- |
| HD20XX | UNITID |
| C20XX\_C | AWLEVELC, CSTOTLT, CSTOTLM, CSTOTLW, CSAIANT, CSASIAT, CSBKAAT, CSHISPT, CSNHPIT, CSWHITT, CS2MORT, CSUNKNT, CSNRALT, CSUND18, CS18\_24, CS25\_39, CSABV40, CSUNKN |
| DRVC20XX | ASCDEG, BASDEG, MASDEG, DOCDEGRS, DOCDEGPP, DOCDEGOT, CERT1, CERT2, CERT4, PBACERT, PMACERT |
| DRVIC20XX | TUFEYR3, CINSON, COTSON, CINSOFF, COTSOFF, CINSFAM, COTSFAM |
| DRVEF20XX | ENRTOT, PCTENRWH, PCTENRBK, PCTENRHS, PCTENRAP, PCTENRAS, PCTENRNH, PCTENRAN, PCTENR2M, PCTENRUN, PCTENRNR, PCTENRW, EFUG, EFGRAD, DVEF02, DVEF03, PCTDEEXC, PCTDESOM, PCUDEEXC, PCUDESOM, PCGDEEXC, PCGDESOM |
| IC20XX | PEO1ISTR, PEO2ISTR, PEO3ISTR, PEO4ISTR, PEO5ISTR, CREDITS1, CREDITS2, CREDITS3, CREDITS4, DISTNCED, DSTNCED1, DSTNCED2, DSTNCED3 |

Not all of these fields are used in the portal as of its release, but they were included to allow for evolution of its features without the labor of preparing the entire data set from scratch again.

**degrees\_institution.csv** is created by querying UNITID, AWLEVEL, CIPCODE, and CTOTALT from table C20XX\_A and concatening the resulting tables after adding a year column to differentiate between them. The degree category column is also added during cleaning. Data processing for both degrees\_institution.csv and ipeds\_dynamic.csv will need to be done using Python in a Jupyter Notebook or some other IDE as the files are much too large for Excel to handle. The code necessary to wrangle/clean it into its final form can be found in data\_processing.ipynb. Also, note that the cleaning code drops columns containing the demographic information of degree earners. These fields were included in initial queries, but dropped afterwards to significantly reduce the size of the dataset. Future queries don’t need to include these fields.

Finally, **degrees\_main.csv** is derived from degrees\_institution.csv. UNITID’s are dropped and the data is grouped by CIPCODE so that the focus is on what majors are earned in aggregate instead of at specific schools. The code for producing this data set can be found in data\_processing.ipynb

Within the zip file containing each Access database, there is an Excel spreadsheet which assigns comprehensible names to the table and variable abbreviations for that year. These mappings informed the creation of decode.py, which maps integer values in columns to specific strings within portal pages. The IPEDS Dictionary page of the portal provides a simple interface for searching and referencing this table.

As IPEDS releases a new database each year, each of these tables will need to be updated annually.

## Education Page

The education section of the portal allows users to explore educational attainment and educational requirements for different occupations. It is built using three datasets: ed\_attainment\_by\_field.csv, edlevel\_by\_occupation.csv, and requireded\_by\_occupation.csv.

The data in **ed\_attainment\_by\_field.csv** comes from Census American Community Survey table S1502, “Field of Bachelor’s Degree for First Major” (available here: [https://data.census.gov/cedsci/table?tid=ACSST5Y2020.S1502](https://data.census.gov/cedsci/table?q=Educational%20Attainment&g=0100000US_0400000US01,02,04,05,06,08,09,10,12,13,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,44,45,46,47,48,49,50,51,53,54,55,56&tid=ACSST5Y2020.S1502&moe=false&tp=true)). A significant amount of wrangling was done manually inside of an Excel workbook to prepare this dataset. Ignore the breakdown by sex and look only at the total column. Mirror the format of the current file when copying and pasting values. Since this data is based on 5 year estimates, it won’t need to be updated until 2025.

**edlevel\_by\_occupation.csv** comes from BLS Employment Projections Table 5.3, “Educational attainment distribution for workers 25 years and older by detailed occupation, 2018-19”, available at <https://bls.gov/emp/tables/educational-atainment.htm>. Click the “XLSX” hyperlink below the title to download it as an Excel spreadsheet. Delete the row which contains the title of the table so that the column names are the first row, then export the dataset as a CSV file. This table won’t need to be updated until 2030.

**requireded\_by\_occupation.csv** comes from BLS Employment Projections Table 5.4, “Education and training assignments by detailed occupation” (available at: [https://bls.gov/emp/tables/education-and-training-by-occupation.htm](http://bls.gov/emp/tables/education-and-training-by-occupation.htm)). Delete the row which contains the title of the table so that the column names are the first row, then export the dataset as a CSV file. This table also won’t need to be updated until 2030.

## Employment Page

The employment section of the portal allows users to explore employment projections as well as historical employment counts and wage data for different occupations. It is built from two datasets: employment\_projections\_by occupation.csv and oesm\_nat.csv.

**employment\_projections\_by\_occupation.csv** comes from BLS “Table 1.2 Employment by detailed occupation, 2020 and projected 2030”, available here: <https://bls.gov/emp/tables/emp-by-detailed-occupation.htm>. Under the title, where it says “Other available formats: ”, there is a hyperlink which will allow you to download the table as an Excel file. From there you can convert it into a CSV. The only cleaning required is deleting the row from the top which contains the title of the table so that the column headers are the first row. This table won’t need to be updated until 2030.

The data contained in **oesm\_nat.csv** comes from BLS’s Occupational Employment and Wage Statistics here: <https://www.bls.gov/oes/tables.htm>. New data is released in May of each year, so the table will need to be updated yearly. The current file contains data from 2015 onwards. Click the hyperlink “XLS” next to “National” to download the data as an Excel file. The code found in the Employment page section of data\_procssing.ipynb will help you transform and concatenate it with the rest of the data. Pay special attention to the names of the columns. They have changed over the years. Make sure you update them to match the oesm\_nat.csv ones before concatenating. Again, the code in the data processing notebook will make this simple.

## Industry Page

The industry section of the portal allows users to explore industry prevalence and trends both nationally and at a state level. It uses three tables, each of which correspond to a section of the page. These are: industry\_by\_state.csv, industry\_by\_region.csv, and industry\_over\_time.csv.

The **industry\_over\_time.csv** dataset was built from BLS data available at <https://download.bls.gov/pub/time.series/ce/>. Cleaning and preparation of the data was done manually, downloading the appropriate tables and then combining desired parts through copying and pasting into a final spreadsheet and converting to a CSV file. For each industry (ignoring files at the top and starting with mining and logging), data comes from the AllEmployeeHoursAndEarnings file.

The industry column contains the name of the industry the dataset describes. The segment column currently only contains the string “total”. The original tables break down each industry into extremely specific sub groups (i.e. bowling center managers), so many that it made more sense to only include the industry category overview. Should the portal ever evolve to include this higher level of detail, this column would contain the name of the specific sub industry. Year comes directly from the data, beginning 2006 onward. Period on the website maps to month in the data set. Values are formatted as “M1”, “M2”, “M3” etc with the number 1-12 representing the month number. The date column is self created. The month and year are used as given with the day assumed to be the first of the month. This date column was important for making the x-axis of the graphs in the portal functional.

Employees, hours, and earnings come from the value column on the site. To differentiate which values are what, it is necessary to look up the series id in the first column. Look for the series which contain averages for the industry segment at its broadest level. <https://econstats.com/blsce/ces10m.htm> contains table names which are helpful for doing so. Conveniently, the ones used in the portal’s current data set are the first three series on each page. Data is released monthly. As such, the backing CSV should be updated regularly.

The **industry\_by\_state.csv** and **industry\_by\_region.csv** are the same data set except industry\_by\_region.csv has a Code column which maps the region column’s name to a state abbreviation. The states dictionary in decode.py can be used to append this additional column which is necessary for the geographic visualizations in the portal. The data comes from Census American Community Survey table S2405: Industry By Occupation For the Civilian Employed Population 16 Years and Over (<https://data.census.gov/cedsci/table?q=S2405&tid=ACSST5Y2020.S2405>). Include states in Geos. Ignore the margin of error column and percent allocated row at the bottom of the table. Copy and paste data in a spreadsheet to match the format of industry\_by\_region.csv.

## IPEDS Dictionary Page

Allows users to look up IPEDS variables and tables using their name, acronym, and other identifying features. Uses two tables, **ipeds\_tables20.csv** and **ipeds\_vartable20.csv**. Both of these files can be exported from the IPEDS access documentation, which can be found at <https://nces.ed.gov/ipeds/use-the-data/download-access-database>. As of 8/18/2022, IPEDS202021Tablesdoc.xlsx contains the vartable and tables sheets used in the portal. The backing tables should be replaced annually following the release of new IPEDS data. After exporting the specific sheets from the Excel workbook, no additional data cleaning is required.

## Summary of Required Updates

| **Dataset Name(s)** | **Portal Section** | **Update Frequency** |
| --- | --- | --- |
| industry\_over\_time.csv | Industry | New data released monthly |
| oesm\_nat.csv | Employment | Annually, data released in May |
| ipeds\_static.csv, ipeds\_dynamic.csv,  degrees\_main.csv,  degrees\_institution.csv | Institutions (Both Main and Specific pages) | Annually, released provisionally in July |
| ipeds\_tables20.csv,  ipeds\_vartable20.csv | IPEDS Dictionary | Annually, released in July; update year in table name (ie. ipeds\_vartableXX) |
| ed\_attainment\_by\_field.csv | Education | Sometime in 2025 |
| industry\_by\_state.csv,  industry\_by\_region.csv | Industry | Sometime in 2025 |
| edlevel\_by\_occupation.csv,  requireded\_by\_occupation.csv | Education | Sometime in 2030 |
| employment\_projections\_by\_occupation.csv | Employment | Sometime in 2030 |