Data Quality Assessment – License Finder

1. Relevance

- What is the total number of items relevant to non-degree credentials?
  32 variables for each license record. Information includes license titles, types, descriptions, states/territories, licensing agencies, application requirements, active, status, and NAICS and O*NET codes. The dataset also includes state FIPS, active status, and O*NET codes for 14 license compacts.

- What are the measures of NDC attainment like?
  The definition of an occupational license as used in the Current Population Survey and adopted by License Finder is that a license
  - Is a credential awarded by a governmental licensing agency based on pre-determined criteria
  - The criteria may include some combination of degree attainment, certifications, educational certificates, assessments (including state-administered exams), apprenticeship programs, or work experience
  - Conveys a legal authority to work in an occupation

  The following variables are relevant to measures of NDC attainment.
  - License type: Stand-alone license/Registry/Tied to business/Secondary license (other license is prerequisite)/Preliminary or temporary license/Undetermined
  - NAICS and O*NET codes
  - Application requirements:
    - Exams: No exam/State exam required/Third-party exam required/Both state and third-party exams required/Choice of state or third-party exam/Undetermined
    - Education: No education required/Specific course required/Degree required/Undetermined
    - Continuing education: No CE requirement/CE required/Undetermined
    - Certification: No mention of certification/Certification may substitute for license requirements/Certification required/Undetermined
    - Experience: No experience/Affidavit or referral/Experience/Current employment/Undetermined
    - Criminal record: No criminal record requirements/Specific type of conviction prohibited/Felony convictions prohibited/Any conviction prohibited/Background check required/Undetermined
    - Physical: No physical requirements/Vision test required/Physical exam/More significant physical requirements/Undetermined
    - Veteran preference: No veteran preference/Undetermined/A temporary license available to military and spouses until formal license approval/Advisers or additional guidance is available for military and spouses/Licensure by endorsement is available for military and spouses/Expedited processing is available for military and spouses/Expedited processing is available for military and spouses, and a separate licensure by endorsement process occurs/Expedited processing is available for military and spouses, and no background check is required/Expedited processing is available for military and spouses, with a temporary license available in the interim/Fees are reduced for military and spouses/Fees are reduced and a temporary license available in the

interim/Fees are reduced and expedited processing is available for military and spouses/Fees are reduced and expedited processing is available for military and spouses, with a temporary license available in the interim/Military and spouses are exempt from licensure requirements

- Active status: Active/No new licenses issued/Replaced/No longer licensed/Undetermined

- Are there any indicators related to NDC attainment that are unique to this dataset?
  Yes. It has a unique classification of license types and provides rich information on application requirements.

- Are there indicators of other phenomena that could be of sociological significance?
  No. The License Finder only contains data on licenses and the organizations that issue them.
- What is the purpose of the dataset, and how closely does that purpose align with the following use cases? (Evaluate as relevant or not relevant.)
  The License Finder is intended to help users find information about occupational licenses that states require for some jobs.
  a. Measuring the rate of attainment of non-degree credentials within the U.S. skilled technical workforce
     Not relevant. No data on the number of licensees.
  b. Measuring aggregate returns to non-degree credentials by credential type
     Not relevant. No data on returns of a license. Not linkable to individual income information b/c no data on licensees.
  c. Identifying disparities by race and gender in the attainment of non-degree credentials
     Not relevant. No data on the demographics of licensees.
  d. Identifying which non-degree credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Relevant. NAICS and O*NET codes are available.
  e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?
     Relevant. The dataset provides information on important regulatory issues such as the type, application requirements, and active status of a license (or a license compact).
  f. *Other examples we might add?*

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  The dataset attempts to cover all occupational licenses in the United States.

- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  10,715 license records in the dataset (before deduplication, as of 03/2022). We do not have an estimate for the total number of occupational licenses in the United States.

- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?

Analyst Resource Center (ARC), the data developer, collects occupational license data from each U.S. state and territory and combine it with additional data from federal agencies and web-scraping. The data goes through a central clean-up process that standardizes occupational coding, adds likely licenses, and ensures consistent structure.

ARC uses text parsing and information from the Center for State Occupational Regulation (CSOR), License2Work, the National Center for State Legislatures (NCSL), and the Military Spouse Portability Examination Report from UMN to obtain data for active status and application requirements. Data related to license compacts are obtained via internet search and is updated on an as-needed basis. Data related to industry is identified based on license description and details. When no industry is specified, it is assumed that the occupation requires a license across all industries.

ARC accepts requests for changes or additions to database through email (arc.deed@state.mn.us) or telephone (651-259-7398).

- Do cases that we believe should exist in the microdata actually exist in the data?
  ARC acknowledges that some states do not participate in the data collection effort and there is no way to guarantee that each state collects all of the license information for their state. In addition, inaccurate reporting also contributes to the missing data problem.

- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.)
  In the flat file (explained in "Accessibility"), the missing rate is 0.18% for O*NET codes and 14% for license types, but the missing rate for license types is zero in the relational file. For all other key variables, the missing rate is zero (as of 03/2022).

Summary assessment: Qualitative description of evidence of completeness and/or steps taken to ensure completeness. Describe percent of variables of interest with missing data, any patterns we can infer as to the distribution of missing data. Evaluate whether the dataset is sufficient (yes or no) for each use case.

Though a full coverage of all occupational licenses in the U.S. cannot be guaranteed, the data developer has made adequate effort to ensure that data missed from state submissions are filled with information from various sources. The missing rate for key variables is substantially zero, and the dataset is sufficient for each use case identified as relevant in "Relevance".

3. Granularity

- How granular (i.e., how many different categories exist, if not continuous) is data for key variables of interest (attainment, field of study, income)? What about for different levels of aggregation researchers might consider, such as geography, age, and race?
  Variables for license types, application requirements, and active status, as listed in Section 1, have a high level of granularity. Both two-digit and six-digit NAICS codes are available, and the O*NET code is at the eight-digit level.

- Is this data granular enough (yes or no) to perform analyses for each of the use cases identified under "relevance"?
  The data is granular enough to perform analyses for each of the use cases identified as relevant.

Summary assessment: How do we rate the overall granularity of the data (high, medium, low)?

High.

4. Timeliness

- How often is the dataset updated?
  Officially every four to six months.
- Is data collected continuously or in waves? If in waves, what is the duration of those waves? Are some variables/records more frequently updated than others?
  License information is collected from each state by the Analyst Resource Center and available for download on CareerOneStop. States are expected to submit revisions every two years, and new information is updated on CareerOneStop every four to six months, typically in March, July, and November, as new information is received. Some records may be more frequently updated than others as the update frequency varies with states and occupational licenses.
- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  The maximum time lag between when a state submits new information and when the information is updated in the dataset seems to be six months.

Summary assessment: What is the length of the field period and the time between field and the availability of data to researchers?

The length of the field period is four to six months. The maximum time lag between data collection and data availability seems to be six months.

5. Integrity

- What are the risks to the integrity of this dataset?
  Most of the data is reported by state agencies, which may have different levels of motivation in reporting data accurately and timely.
- How are data outliers handled? (May be available from published documentation if not metadata.)
  It seems that data is recorded as-is. Outliers are not a significant issue in this dataset as all key variables are categorical.
- Were there changes to the dataset that may have resulted from political influence? If so, do those changes threaten the overall quality of the data?
  None that we are able to identify.

Summary assessment: Describe any known risks to integrity we are able to determine from our research.

We do not identify significant risks to integrity.

6. Accessibility

- How do researchers access this dataset?
  The whole dataset is available for download in Microsoft Access in two different file types, relational and flat. The flat file contains all available information in one table, but because many licenses are coded to multiple O*NET codes, there are duplicates.

- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  The dataset contains a table of the number of licenses awarded for a selected occupation in each state, but currently there is no records in that table in the published version of the dataset.

- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  Downloading the dataset is free. Currently the file size is 26MB for the relational file and 137MB for the flat file. The data developer recommends using the dataset in Microsoft Access because when exporting data in other formats, such as Microsoft Excel, the length of the description fields may cause formatting issues.

Summary assessment: Is the data available to researchers? How do the hurdles to accessing data compare to other datasets we evaluate? Is the data access procedure consistent for all parts of the dataset, or are there pieces of the data that are more or less accessible?

Yes, researchers can download the whole dataset for free and open it easily on most devices equipped with Microsoft Access.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?
  No, the license ID in this dataset is not a unique identifier, nor is the combination of license ID and state FIPS. Inaccurate reporting of state agencies adds to the difficulty of data cleaning and is a source of duplicates in the dataset.

- Is it common? (e.g., a SSN might be higher value than an address, though even name/address might be sufficient to match in some cases).
  No UID found.

- Do occupation and industry coding schemes correspond to commonly used frameworks such as O*Net and NAICS? If not, are they well documented in metadata?
  Yes, both O*NET and NAICS codes are available.

Summary assessment: Are linkages possible on key variables or individual cases (yes or no)? Rate the potential for establishing meaningful data linkages for each use case (good, fair, poor).

Linkage on individual cases is poor as there is no unique identifier, but still possible through license names. Linkage to occupation and industry coding schemes is straightforward as both O*NET and NAICS codes are available.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  Yes for key variables, based on the two versions available (03/2022 and 11/2021).
- What is the construction of the dataset like? Is the microdata organized in "waves"? Can multiple observations for the same unit of analysis (i.e., person) over time be easily linked together?
  For each downloadable version, the data is cross-sectional, and only the most recent version is open to download from the website. Linkage among the records for same license over time is difficult as no unique identifier exists.
- How far back do administrative records from this dataset go?
  According to the dataset's technical document, data collection began in 1997. However, in a 2019 quality review, the reviewer notes that available prior versions date back to 2018. The reviewers acknowledged the complexity of collecting longitudinal data but also expressed interest in creating a time series.

Summary assessment: Identify the length of time covered by the dataset (and the consistency of data collection over time) and rate as shorter or longer than other datasets. Objectively assess fit between time covered by data and time period of interest for each use case.

Data for License Finder has been collected for 25 years and key variables seems to be consistent over time. Only the most recent version is open to download from the website and no unique identifier is available, adding to the difficulty of linking records from different time periods and doing longitudinal research with this dataset.