Data Quality Assessment – Certification Finder

Rubric for Assessing Dataset Quality

Each dataset subject to our data quality assessment will be evaluated according to a rubric as defined below. The answers to these questions will be used to create and publish profiles of each dataset, which will be available on the website of the GW Program on Skills, Credentials, and Workforce Policy.

1. Relevance

- What is the total number of items relevant to non-degree credentials?
  47 variables in use, including certification name, acronym, type, training/education/exam/renewal requirements, demand/accreditation status, NAICS and O*NET codes, and contact information of the certifying organization.

- What are the measures of NDC attainment like?
  The data developer defines a certification as "an award you earn to show that you have specific skills or knowledge in an occupation, industry, or technology". The following variables are relevant to measures of NDC attainment.
  - CERT_TYPE (certification type): core, advanced, skill, specialty or product/equipment (see here for definitions)
  - Requirements:
    - TRAINING: Is significant education or training is needed? Yes, No, or Unknown
    - EXPERIENCE: Is significant work experience is needed? Yes, No, or Unknown
    - EXAM: Is an exam required? Yes, No, or Unknown
    - RENEWAL: How many years until certification must be renewed?
    - CEU: Can it be renewed with Continuing Education Units? Yes, No, or Unknown
    - REEXAM: Can it be renewed with re-examination? Yes, No, or Unknown
    - CPD: Can it be renewed with Continuing Professional Development? Yes, No, or Unknown
    - CERT_ANY: Can it be renewed in multiple ways. Yes, No, or Unknown
  - ACCRED_ID (demand/accreditation status)
    - High demand: frequently mentioned in online job postings
    - Industry endorsed: endorsed by a major industry association that is not itself the developer of the certification
    - Related to the Job Corps training program
    - Identified in military Credentialing Opportunities On-Line (COOL) sites
    - Accredited by the American National Standards Institute (ANSI)
    - Accredited by the National Commission for Certifying Agencies (NCCA)
  - NAICSCODE and ONETCODE

- How much data is provided on each NDC described in this dataset? Are there any indicators related to NDC attainment that are unique to this dataset?
  Yes. It provides rich information on the requirements and demand/accreditation status of a certification. It also has a unique classification of certification types.

- 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 Certification Finder is intended to help people identify professional certifications that may be useful for current or future employees.
  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 certified individuals.
  b. Measuring aggregate returns to non-degree credentials by credential type
     Not relevant. No data on returns of a certification. Not linkable to individual income information b/c no data on certified individuals.
  c. Identifying disparities by race and gender in the attainment of non-degree credentials
     Not relevant. No data on the demographics of certified individuals.
  d. Identifying which non-degree credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Relevant. Data available on demand/accreditation status as well as NAICS/O*NET codes.
  e. Evaluating the effectiveness of public policies that support the attainment of non-degree credentials?
     Relevant. Data available on a certification's relation to Job Corps (a public training program), COOL (a military credentialing assistance programs), and two non-profit accreditation agencies, ANSI and NCCA.
  f. *Other examples we might add?*

2. Non-response

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  The data developer does not explicitly state the frame of reference/population of the dataset, but it seems that the dataset attempts to cover all professional certifications in the United States.

- What is the number of cases, and how does that number compare to known estimates of the relevant population?

  11,543 certifications in the dataset (before deduplication, as of 03/02/2022). We do not have an estimate for the total number of professional certifications in the United States.

- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  We did not find information about the publisher's data collection practices.

- Do cases that we believe should exist in the microdata actually exist in the data?
  We do not see any sign of missing cases in our assessment of the data.

Summary assessment: Did the data publisher make adequate efforts to avoid non-response? Describe any efforts to avoid non-response and any evidence that non-response has been minimized.

3. Coverage

- What does the organization that creates or maintains this dataset do to minimize missing data?
  We did not see evidence of a specific strategy to minimize missing data.

- What percent of cases lack data for key variables of interest? (Direct assessment if not in metadata.)
  Coverage rate for key variables ranges from 24% to 99%. See below for details. (11,543 records as of 03/02/2022)

Variable	Obs.	Coverage
TRAINING	8,597	74%
EXPERIENCE	8,415	73%
EXAM	11,450	99%
RENEWAL	6313	55%
CEU	6,040	52%
REEXAM	5,685	49%
CPD	5,294	46%
CERT_ANY	6,536	57%
CERT_TYPE	9,526	83%
ACCRED_ID	2,743	24%
NAICSCODE	8,921	77%
ONETCODE	8,748	76%

- What percent of the population of interest is in the dataset (if the size of the overall population is known or can be estimated)?
  We do not have an estimate for the population of professional certifications in the United States.

4. 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?
  There are five types of credentials, seven types of indicators of demand/accreditation status, seven trinary (yes/no/unknown) variables for training/education/exam/renewal requirements, and a continuous variable for the maximum years between two renewals. The NAICS code is at the six-digit level and the O*NET code is at the eight-digit level.

Summary assessment: Is this data granular enough (yes or no) to perform analyses for each of the use cases identified under "relevance"? How do we rate the overall granularity of the data (high, medium, low)?

The data is granular enough for analyses related to industry and occupations. It is not granular enough for analyses related to training/education/exam/renewal requirements as most variables of interest are only trinary (yes/no/unknown).

5. Timeliness

- How often is the dataset updated?
  According to the Data Source page on CareerOneStop, the dataset is updated on a rolling basis, though the technical document of the dataset notes that data is updated biannually.
  The most recent downloadable version of the dataset is dated 03022022, and a 09032021 version, previously downloaded, is also available, so it seems that the update schedule is now on a rolling basis.

- Is data collected continuously or in waves? If in waves, what is the duration of those waves? Are some variables more frequently updated than others?
  Continuously. No information on whether some variables are more frequently updated than others.

- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?

  How soon a certification will be added to the dataset after its creation is unknown, but the dates when a certification is added to the dataset and when it is last updated are both available.

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

The dataset is updated on a rolling basis. We do not know how soon a certification will be added to the dataset after its creation. Researchers can assess the recency of a record through the dates when it is added and when it is last updated.

6. Integrity

- What are the risks to the integrity of this dataset?
  CareerOneStop, the data developer, is sponsored by the Employment and Training Administration (ETA) of the U.S. Department of Labor. We are not able to identify any special ties between a specific certifying organization/industry group and CareerOneStop/ETA.

- How are data outliers handled? (May be available from published documentation if not metadata.)
  It seems that data is recorded as-is. Most key variables are categorical. The only key variable that is continuous is the maximum years between two renewals, which ranges from 0 to 10.
  We do not know if the data developer tries to handle outliers in a certification's frequency of being mentioned in online job postings. This data is used to determine whether a certification is in high demand but is not disclosed in the dataset.

- 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.

7. Accessibility

- How do researchers access this dataset?
  The whole dataset is available for download. Researchers can also download part of the dataset by browsing specific occupations/industries or using the keyword search feature on the website.
- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  Yes. SUPPRESS indicates whether a record contains confidential data that must be suppressed for public use. However, the type of suppressed information is unknown. DELETED indicates whether a record is deleted from the website but still exists in the database.
- 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 folder size is 38MB and the dataset can be opened with Microsoft Access or SQL.

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 or SQL.

8. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?
  Yes, it is CERT_ID (numerical). The variable is created exclusively for this dataset, but researchers may link records to other datasets through CERT_NAME (text), which is not unique to all records and indicates the existence of duplicates.
- 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).
  Certification names should be common and largely consistent among datasets of certifications.
- 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 possible through CERT_NAME, though deduplicating records and comparing them with certification names in other datasets can be demanding. Linkage to occupation and industry coding schemes is straightforward as both O*NET and NAICS codes are available. Records can also be linked to job training and accreditation programs like Job Corps, the military COOL, ANSI, and NCCA

through corresponding indicators. The name and contact information (phone, email, address) of the certifying organization is also provided so linkage to certifying organizations is also possible.

9. 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 (03022022 and 09032021). According to the technical document, three variables (EITHER, NSSB_URL, and CERT_URL) are not in use anymore.

- 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.
  CERT_ID, consistent over time, can be used to link records of the same certification in different versions.

- How far back do administrative records from this dataset go?
  2002, according to the technical document.

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 Certification Finder has been collected for 20 years and key variables seems to be consistent over time. Only the most recent version is open to download from the website, adding to the difficulty of doing longitudinal research with this dataset. Researchers can link records of the same certification in different versions using CERT_ID, unique to each record and consistent over time.