Data Quality Assessment – National Labor Exchange (NLx) Research Hub

Rubric for Assessing Dataset Quality

1. **Relevance**

- What is the total number of items relevant to credentials?
  There are 51 variables in the dataset, those most relevant to NDC attainment include job ID, job title, job description, job URL, O*Net code, NAICS code, number of positions available, job schedule, job shift, expected number of hours per week, salary unit, salary minimum and maximum, minimum education required, minimum experience required, license requirements, training requirements, application methods, job posting location, date and time the job was first acquired by DirectEmployers, date of most recent update, date of expiration; company name, contact, location Federal Contractor status, and Federal Employer Identification Number.

- What are the measures of credential attainment like?
  Degree, license, and training programs.

- Are there any indicators related to education attainment that are unique to this dataset?
  Job salary and hours information.

- Are there indicators of other phenomena that could be of sociological significance?
  There is potential to draw out a huge amount of data from the job posting descriptions. One could theoretically mine data on certifications and other credentials required, skill requirements, experience requirements, benefits, and company policies, among other factors.

- 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.)
  A partnership between the National Association of State Workforce Agencies (NASWA) and DirectEmployers, NLx Research Hub is a cloud-hosted warehouse of jobs data sourced from the National Labor Exchange. NLx Research Hub is created to increase the amount of labor market information in the U.S. to facilitate the recruitment, hiring, and training opportunities of U.S. workers and deepen partnerships between industry, government, and academia by enhancing the infrastructure to support the convergence of research, education, and talent pipelines.
  a. Measuring the rate of attainment of credentials within the U.S. skilled technical workforce
     Not relevant. The dataset does not include individual or aggregate level information of employees or jobseekers.
  b. Measuring aggregate returns to credentials by credential type
     Somewhat relevant. Researchers can combine credential requirement and salary information to assess the aggregate returns to a credential.
  c. Identifying disparities by race and gender in the attainment of credentials
     Not relevant. The dataset does not include demographic information.
  d. Identifying which credentials are associated with the strongest labor market returns for individuals in the skilled technical workforce
     Somewhat relevant. Researchers can combine job posting counts, credential requirement and salary information to assess the employability and expected salary a credential.
  e. Evaluating the effectiveness of public policies that support the attainment of credentials?

Somewhat relevant. Researchers can use the dataset to inform policymakers of the most needed credentials.

2. Coverage

- What is the frame of reference for the dataset, what population does the dataset attempt to cover?
  NLx attempts to cover all real job postings in the United States.
- What is the number of cases, and how does that number compare to known estimates of the relevant population?
  The database includes approximately 300,000 employers, 4 million daily job postings, and 75 million historical job postings, including openings from small- and medium-sized employers.
- How does the publisher of the data ensure that data is collected for cases that should be in the dataset?
  All 50 state workforce agencies, plus District of Columbia, Guam, Puerto Rico, and the U.S. Virgin Islands participate in NLx. NLx uses indexing (a.k.a. scraping or spidering) to extract job postings from the career sites of over 18,000 employers. The indexed employer community includes both DirectEmployers member companies and nonmembers who would like their jobs to appear in the NLx. State workforce agencies can help increase the total number of NLx job openings by identifying indexable corporate sites and notifying the NLx operations team.

  NLx gathers currently available and unduplicated job opportunities and takes validation seriously to ensure only real jobs from verified employers are made available. Jobs with requirements to purchase training or equipment to secure employment, unpaid positions, commission-only positions, multi-level marketing positions, franchise opportunities, and companies currently involved in a strike or labor dispute or of questionable ethics or illegal are not included.

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?
  - Job posting location: zip code, city, state, country
  - Company location: address, zip code, city, state, country
  - Job schedule: part-time/full-time/flexible
  - Shift: day shift/night shift/swing shift
  - O*Net code:
  - NAICS code
- Is this data granular enough (yes or no) to perform analyses for each of the use cases identified under “relevance”?
  Yes, though the data may need serious cleaning and/or organizing.

4. Timeliness

- How often is the dataset updated?
  Daily.

- 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?
  Continuously.

- What is the time lag between when an event occurs, when it is recorded, and when that data is available to researchers?
  The NLx feed is refreshed on a daily basis through a "kill and fill" process. When a job is taken off a corporate website, state job bank, or USAjobs.gov, it will no longer be made available for viewing on NLx. This usually happens within one day of the job being removed from the source site.

5. Integrity

- What are the risks to the integrity of this dataset?
  None that we are able to identify. NLx takes validation process to ensure that scams and illegal jobs are excluded from the database.

- How are data outliers handled? (May be available from published documentation if not metadata.)
  There is no effort to exclude outliers.

- 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.

6. Accessibility

- How do researchers access this dataset?
  Researchers need to complete and submit a data trust request form to request data from NLx.

- Are any variables or cases withheld from researchers? If so, does that withholding or censoring affect researchers' ability to use the data?
  Company's name, Federal Employer Identification Number, address line 1 and O*Net code of the job posting are restricted in the dataset. This limits researcher's ability to link job postings to companies and occupations.

- Are there direct costs or indirect costs (e.g., training, resources) associated with accessing and using the data?
  Researchers need to fill in the request form, provide supporting documents, and wait for approval from the NLx. Denied request cannot be resubmitted unless modified to account for reasons for denial.

7. Interoperability

- Is there a unique identifier for individual cases? If so, is it one that can be found in other datasets?

Yes. There are two unique job IDs, one utilized by the Data Warehouse and another assigned by DirectEmployers. Linkages to industry, occupation, and company are possible but can be limited by data restrictions described in Section 6.

- 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.
- 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, NAICS and O*Net codes are available, though O*Net code is restricted.

8. Suitability for Longitudinal Research?

- Is the metadata consistent over time, at least for key variables?
  Yes
- 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?
  It is continuous. There are no "waves." One can look back in time and download data extracts referring to specific time periods in specific places, though if analyzing the data offline there may be computational challenges.
- How far back do administrative records from this dataset go?
  Approximately 2010.
-