**Appendix: Database description**

1. Archival sources

The archival sources are alumni lists, professional association membership lists, and corporate directories. The list of schools includes all major mining and metallurgical engineering programs in the United States with nearly comprehensive coverage from 1870-1915, as well as several major school programs in Mexico, England, Germany, and France. The two corporate directories include all significant mining and metallurgical firms in the Anglo-American sphere, operating in mining districts around the world. Full bibliographic citations are available on request. Note: we could include a much abbreviated version of this, with the full data appendix available by request.

**Alumni lists**

University of Arizona, 1895-1915

Camborne School of Mines (UK), 1898-1940

Colorado School of Mines, 1883-1905

Columbia University, School of Mines, 1867-1929

Escuela Nacional Ingenieros (Mexico), 1868-1905

École National des Mines (France), 1871-1922

École Nationale des Mines de Saint-Étienne (France), 1871-1922

BergAkademie Freiberg (Germany), 1766-1939

Harvard University, 1811-1919

Lehigh University, 1874-1917

Michigan College of Mines, 1888-1909

Missouri School of Mines, 1874-1915

MIT, 1861-1940

Montana School of Mines, 1903-1925

Pennsylvania State College, School of Mines, 1861-1916

Royal School of Mines (UK), 1812-1920

Stanford University, 1892-1899

University of Minnesota, Minnesota School of Mines, 1902-1924

University of Nevada, Mackay School of Mines, 1876-1917

University of New Mexico, 1890-1921

University of South Dakota, 1888-1915

University of West Virginia, 1905-1919/20

Worcester Polytechnic, 1871-1926

Yale University, Sheffield Scientific School (Mining and Metallurgy),

Total number of cells: 1,980,712

Total number of individuals: 66591

**Professional Association – AIME**

American Institute of Mining Engineers. [members & associates lists]

 1873-1896, 1896, 1898-99, 1901-03, 1906, 1912

Total number of cells: 289,548

**Corporate Directories**

American Mining Manual (AMM; sometimes International Mining Manual), embracing the principal operating metal mines, mills, smelting & refining plants of the United States, Mexico and Canada and coal mines of the western states, Mexico and Canada. Alexander R. Dunbar, ed. Denver: Western Mining Directory Co.: 1907, 1911, 1916, 1919, 1920, 1922

Mining Year Book (MYB) containing full particulars of mining companies. Walter R. Skinner. London: The Financial Times: 1887-1889, 1891, 1893-1895, 1903, 1908-1923

The Mexican Magazine. México, D.F. 1926

Total number of cells: 15,141,420

1. Extraction, cleaning, and name parsing

The extraction process follows the digitized text documents generated through Hathitrust or through Adobe, file by file. The source format and the accuracy of the OCR require a specific combination of automated extraction, visual classification, and data validation. The cleaning and classification process used regular expressions in tidyverse,[[1]](#endnote-1) and several iterations of parsing through openrefine.[[2]](#endnote-2) See our code (alumni.R and corpis.R)[[3]](#endnote-3) for the specific steps used for the different types of sources; all codes are available on request.

Name parsing for individuals was elaborated through a simplification of the name. Names appeared in the sources in several forms, from full names with titles to abbreviations with last names and initials. Some of them had different orthographies (as for Mc or Mac, or elimination of "von" or "de"), which further complicated the parsing. We simplified the orthographies and the structure into last name and initials to consolidate them. See the file combo.R for the specific coding steps of consolidation.

Name parsing for organizations is more complicated, as organizations changed their name, consolidated with others, were absorbed or simply registered as a subsidiary on some occasions.[[4]](#endnote-4) Naming authorities contain a small amount of information for these entities, as thousands of firms appeared and disappeared in a relatively short amount of time. Consequently, the name parsing was a combination of cleaning through regular expressions, searches through the Virtual International Authority File database,[[5]](#endnote-5) and use of openrefine. See the file corpis.R and combo.R for the specific criteria.

1. General characteristics of the database

The database was compiled using R software into a tidy format. Values, or the discrete pieces of information, are organized in variables, with all values with the same attribute and observations, and all values measured in the same unit.[[6]](#endnote-6) Throughout the cleaning process, we can identify three units that conform to the database: rows of text, individuals, and organizations.

 The units as rows of text are the original form of extraction on the database. The classification process then identifies the type of information through regular expressions but keeping a particular row ID (which allows us to trace back the data mining process). A single row can then contain an observation of the pattern that is the focus of the original source, either an individual-centered list or an organizational-centered list. In many cases, a combination between several rows appears. Non-classified information and trash were eliminated from this original text-centered dataset in the cleaning process, and the total number of values in this dataset is 18,210,654.

The classification process then proceeds to identify the patterns of occurrence of information in every row and the relationship with other rows. While keeping the row ID, the classification assigns characteristics from other rows into a single unit. This transforms the database from one centered on rows of text, in which the observation is a line, into two types of datasets: one centered around individuals (in which individuals are the units) and one centered around organizations/companies (in which companies are the units). The properties of the different rows are classified into attributes of organizations (capital, incorporation, technology, employees, etc.) and individuals' characteristics (place of origin, education, employment, etc.).

The dataset centered around individuals contains 315,137 observations of 15 variables, corresponding to 124,451 unique individuals. The dataset centered around organizations contains 116,780 observations of 16 variables, corresponding to 52,206 unique organizations. Both datasets can be described as unbalanced data panels, like other datasets around firms and individuals, with a *long* or *stacked* structure. The voluntary reporting, the transformations of corporate names and government agencies, and the sole existence of new actors (individuals, universities, and corporations, among others) appearing in the mining world make the construction of a complete dataset impossible. [[7]](#endnote-7) Nevertheless, we have no information to suspect that the missing observations are nonrandom, as incentives to report did not change during the period. Plot 1 shows the temporal density of individuals in the database, and Plot 2 shows the temporal density of companies in the database.

It is difficult to know from the dataset the gender composition, but we used the genderdata package to make historical predictions of sex assigned at birth. We predicted the sex assigned at birth of 70,834 individuals, almost 55% of the sample: from them, around 2.6% might have been female, and 97.4% might have been male.[[8]](#endnote-8) The geographical distribution of the database is heavily concentrated in the North Atlantic for both the individual-centered dataset and the organization-centered dataset. We can identify 242,629 geolocations on the individual-centered dataset, corresponding to 23,644 different locations corresponding to 98,501 different individuals. We can identify 563,693 geolocations on the organization-centered dataset, corresponding to 3,105 locations and 21,063 different organizations.[[9]](#endnote-9)

Plot 1. Total number of individuals in the individual-centered dataset. 

Plot 2. Total number of organizations in the organization-centered dataset.



Plot 3. Geolocation of individuals in the individual-centered dataset



Plot 4. Geolocation of organizations in the organization-centered dataset



1. Validation

There are four levels of possible errors in the final database: errors in the extraction process through the OCR; errors in the cleaning process; errors in classification between different kinds of information; and errors in parsing, both for individuals and organizations.

 We conducted three assessments of the types of mistakes present in the database in 2020 and 2021 that measured the most common errors and the overall accuracy of the methodology. We separated the text-centered dataset into five categories according to the quantity of data. Then, a randomized representative sample at 95% significance of the rows of the database was evaluated by a researcher, comparing them with the original source and making informed decisions about the mistakes in the process. See the files alumni.R, corpis.R and test.xlxs for the specific algorithms used for randomizing the evaluation and the test sample. The main results are in the following table.

Table 1. Validation of representative samples of the dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  Item | MYB (n=409) | AMM (n=384) | MMM (n=368) | ALUMNI (n=384) | AIME (n=382) |
| Company/Organization | 43.77% | 73.84% | 92.18% | 72.62% | 89.73% |
| Capital | 99.27% | 98.78% | 96.82% |   |   |
| Year Registered | 99.27% | 98.29% | 99.76% | 79.95% |   |
| Location | 93.46% | 96.50% | 89.00% |   | 92.42% |
| Last name | 88.75% | 83.13% | 74.08% | 94.62% | 86.55% |
| First name | 94.13% | 97.56% | 94.62% | 93.89% | 86.55% |
| Middle name | 91.69% | 87.78% | 76.53% | 66.99% | 86.55% |
| Role | 96.58% | 98.29% | 96.33% | 76.77% | 95.35% |
| Nationality |   |   |   |   | 99.51% |
| Education |   |   |   |   | 100.00% |
| Overall (n=1927) | 88.36% | 91.77% | 89.91% | 80.81% | 92.08% |

Although the overall accuracy was 88.59%, the disparity of the results with the original sources in essential items ("Company" in the corporate lists and alumni list, and the role in the alumni lists) motivated a general revision of the different steps of cleaning, classifying, and parsing in the complete dataset. After the database extension and the modification of the steps, the dataset was consolidated, and a second revision of the accuracy was made. Again, a randomized representative sample at a 95% confidence level was completed. (comboprueba.xlsx) The company name for corporate sources parsing doubled to 84.09%, and the identification of "middle name" in alumni lists improved to 96.92%. These are the results of the accuracy of the new integrated dataset (combo4.csv and crp.csv).

Table 2. Second validation of representative sample

|  |  |
| --- | --- |
| Item | Accuracy |
| Last name | 94.01% |
| First name | 93.75% |
| Middle name | 93.75% |
| Education | 100.00% |
| Company/organization | 94.53% |
| Location | 94.01% |
| Role | 100.00% |
| Overall (n=384) | 95.72% |

A third assessment of the database looked at the matching of names across different sources, pairing on one side names from the MYB list and, on the other, names in the whole database (including MYB but from different years/sections). This third assessment focused on the probability of false positives or locating spurious matches of information from different sources through a common simplified name. While false negatives in matching (by minor variations on orthography or variations in the reporting of middle names) are possible in the database, false positivity is a more significant concern as it might overestimate the connectivity of individuals and organizations in the dataset. In short, type I errors are considered more costly for this analysis than the existence of type II errors and, therefore, the minimization of false positives is privileged.

 A spurious match could be caused by: convergent errors in the OCR; convergent errors in the cleaning and classification process; coincidence in the last name, the first letter of the first name, and the first letter of the middle name. We generated a randomized representative sample at 95% significance for matching information through a simplified name. Then a researcher validated the parsing as true or false positives by looking at the two original sources and the contextual information. From all the above causes, false positives represented just 1.83% of the test sample (n=382). See the files combo.R and testmat.xlsx for the complete information on this test. After this process, a visual examination of the most prominent 2000 organizations in the database was conducted and manually corrected for the errors described above. Most of the corrections involved not the misidentification of the company but a spelling mistake that was carried out from the OCR. No further accuracy assessment has been made on the manually corrected datasets (combo5\_1.csv and crp3.csv).

1. OCR extraction through Hathitrust .txt file, Adobe PDF OCR and tesseract. Wickham et al., “Welcome to the Tidyverse”; Ooms, “Tesseract: Open Source OCR Engine.” [↑](#endnote-ref-1)
2. See an evaluation of this open software in Petrova-Antonova and Tancheva, “Data Cleaning: A Case Study with OpenRefine and Trifacta Wrangler.” [↑](#endnote-ref-2)
3. R scripts and csvs will be available after the peer review process at github.com/xxxxxx/invisibleredux.git [↑](#endnote-ref-3)
4. See a longer discussion of the problems of name parsing for organizations in Song et al., “Research on Organization Name Matching Based on Word Vector.” [↑](#endnote-ref-4)
5. Schneider, “Viafr: Interface to the ‘VIAF’ ('Virtual International Authority File’).” On the limitations of the VIAF authority file methodology see Bianchini, Bargioni, and di San Girolamo, “Beyond VIAF.” [↑](#endnote-ref-5)
6. Wickham, “Tidy Data”; Silge and Robinson, “Tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” [↑](#endnote-ref-6)
7. See some of the implications of missing data in panel analysis with similar datasets (individuals that self report in waves and firms in the market) in: Baltagi and Chang, “Incomplete Panels: A Comparative Study of Alternative Estimators for the Unbalanced One-Way Error Component Regression Model”; Baltagi and Song, “Unbalanced Panel Data: A Survey”; Young and Johnson, “Handling Missing Values in Longitudinal Panel Data with Multiple Imputation.” [↑](#endnote-ref-7)
8. Mullen, “Gender: Predict Gender from Names Using Historical Data”; Blevins and Mullen, “Jane, John... Leslie? A Historical Method for Algorithmic Gender Prediction.” [↑](#endnote-ref-8)
9. Using the Google API for geolocation, the ggmap package and the countrycode package for mapping. Kahle and Wickham, “Ggmap: Spatial Visualization with Ggplot2.”; Arel-Bundock, Enevoldsen, and Yetman, “Countrycode: An R Package to Convert Country Names and Country Codes”; South, “Rworldmap: A New R Package for Mapping Global Data.” [↑](#endnote-ref-9)