

State RegData Technical Documentation

Purpose:

State RegData extends the logic of the RegData US project to the American states. Like RegData US, State RegData datasets employ the QuantGov platform to download and analyze both state regulations and statutes, turning thousands of pages of dense regulatory and statutory text into datasets. Each state dataset contains the following data outputs: general metadata, restriction counts, word counts, and industry relevance. The production of each State RegData dataset requires unique code development. Interact with State RegData and compare states using the new [RegCensus Explorer interactive](#). A bulk download version of State RegData 3.0 can be found [here](#).

Content:

As of July 31, 2021, there are 46 quantified state regulatory codes in the State RegData project plus the District of Columbia. Regulations for Vermont, New Jersey, Arkansas, and Hawaii are all currently not quantified due to not having a regulatory code, having a non-useable website, or paywalls (or a combination of the three). In addition, Connecticut and Alaska were not quantifiable for State RegData 2.0 and Utah was not quantifiable for State RegData 3.0, so there are regulations data for only 43 states in version 3.0.

For the first time, version 3.0 will also include data for state statutory codes in addition to the regulatory data. There are 43 quantified state statutory codes in the State RegData project plus the District of Columbia. Statutes for Alaska, Arkansas, Colorado, Georgia, Mississippi, New Mexico, and Tennessee are all currently not quantified due to having a non-useable website or paywalls (or a combination of the two).

All these state datasets include a count of the total number of words and restrictions in each part of the state's regulatory or statutory code. It also includes complexity measures of each state's code, such as Shannon entropy and sentence length. Finally, version 3.0 regulations data includes metadata that identifies the agency or department responsible for the regulation (not available for statutes because they are passed by the legislature, not written by agencies).

Machine learning metrics are also included in State RegData 3.0. Each state dataset includes industry level data that lists all of its state's distinct regulatory code parts alongside a 3-digit NAICS industry classification with the probability that the regulatory code part in question pertains to the NAICS industry classification in question.

Technical Notes:

- The development of State RegData 1.0 stretched over 3 years due to the amount of text that needed to be collected. Therefore, while states can be compared with

each other, every state dataset was not collected in the same year. Refer to the dataset's "date collected" for information on when the dataset was collected. Starting with State RegData 2.0 data, the annual collection was condensed to a few months (primarily between March and June 2020), greatly reducing this potential problem.

- State RegData 1.0 used two different machine learning classifiers over its three-year development period. While both are informative, the 3.0 classifier is an upgrade over the 2.2 classifier, and it is highly suggested to not compare industry data from states that use two different classifiers. Starting with version 2.0, State RegData uses the 3.0 classifier for all states.
- In an attempt to rectify the large differences in which states publish regulatory codes, the State RegData project attempted to aggregate each state's code to a unit that is roughly similar. For example, while California organizes their code all the way to the rule level, we aggregated at the chapter level. This allows for easier comparison in both the metadata and the industry relevance results. The general benchmark per unit of analysis is a median of 3,000 words and a mean of 12,000.
- The State RegData 1.0 data for a few states were updated due to webscraping errors in their development. Make sure to update these states with the new data if you are using data from version 1.0 for a current project. These states are Maine, North Dakota, Texas, and Wisconsin.
- A few states saw some big differences between versions 1.0 and 2.0 that cannot not be explained by an ordinary regulatory process. Idaho repealed and replaced their entire regulatory code, resulting in a nearly 40 percent decrease in restrictions. Michigan and Wyoming both repealed regulations relating to OSHA requirements and replaced them with incorporation by reference. Kentucky and Missouri both underwent significant red tape reduction efforts, resulting in significant decreases in their total restriction counts.

Current Version: 3.0

Release Date: 7-31-2021