State RegData Definitive Edition

QuantGov

July 26, 2022

1 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. Interact with State RegData and compare states using the RegCensus Explorer interactive. A bulk download version of State RegData Definitive Edition can be found here.

2 Content

As of July 26, 2022, there are 49 states and the District of Columbia in the State RegData project, which includes both state regulatory and statutory codes. Arkansas is the only state whose regulations we are unable to quantify due to the state not having a regulatory code (but we do have data for their statutes).

The state datasets include a count of the total number of words and restrictions in each part of the state's regulatory code. It should be noted that metadata formatting between the states is inconsistent since every state has a different way of organizing its regulatory code. Michigan's code, for example, has a single level of hierarchy, while Washington breaks its code down by title and chapter.

Each state-level industry dataset lists all of its state's distinct regulatory code parts alongside the number of industry relevant restrictions pertaining to the 3-digit NAICS industry classification, which is calculated by multiplying the probability that the regulatory code part in question pertains to the NAICS industry classification in question by the number of restrictions in that part.

For the first time, State RegData Definitive Edition includes cleaned agency and "cluster" data for state regulations and "cluster" data for state statutes. The agency data refers to the agency responsible for writing or enforcing the regulation, while the clusters attempt to group the agencies into categories. Clusters enable the user to better compare topic-specific regulations or statutes between states, in a way that is more difficult to do with only agency-level data. Clusters were throughtfully determined with a lot of time and effort by the QuantGov team.

3 How to Access

3.1 Exploring and Visualizing the Data

For users that simply want to explore the data we have already made, we created several tools on `quantgov.org` that allow users to visualize the data created from the State RegData project.

The first tool is the State RegData Definitive Dashboard. The dashboard allows the user to compare restriction counts across every state for the year 2022. By clicking into each state, one can see more data containing our analysis of the state's regulatory and statutory codes, including word and restriction counts, top regulated industries, and agency-level data.

The second tool is the RegCensus Explorer. This tool provides a map to explore the state-level data. The tool provides comparisons, rankings, and other useful information. You can even compare the U.S. state data with the states and provinces from Australia and Canada.

Another tool that allows users to explore our data and document is RegHub. Using the search tool at `https://reghub.ai/search` (requires free login), users can search for keywords in all documents in the QuantGov corpus, with the ability to filter by jurisdiction, year, topic, and document type (i.e. regulations or statutes). The full text of the search results can then by downloaded by supplying an email address.

3.2 Downloading the Data

In addition to exploring the data using the QuantGov tools, there are several avenues for downloading the data to explore or analyze for yourself. These tools include bulk downloads, an interactive downloader (for both data and documents), and an API (either direct calls or through a Python library).

Bulk downloads can be found at https://reghub.ai/data/bulk. The downloads are organized by data project. The datasets for State RegData Definitive Edition can be found close to the top, just under the download for RegData U.S. Federal 4.1. The downloads are split into regulations and statutes to minimize the mixing of the two data sources. Each download contains document-level data, including word and restriction counts along with agency and cluster data. A separate dataset contains industry data which can be merged into the main dataset using the "document_id" column. These document-level bulk downloads are intended for researchers looking to explore or analyze our data in detail.

The interactive data downloader can be found at https://reghub.ai/data/custom. The custom data downloader allows the user to download both summary and document-level data for State RegData and other data projects. The data downloader allows the user to select one or more jurisdictions, document type (i.e. regulations or statutes), summary or document-level data, the series (i.e. restriction counts, word counts, industry probabilities, etc.), and years. The interactive downloader is intended to provide smaller, more specific datasets than the bulk downloads, without requiring the user to have knowledge about Python or the API.

The document downloader can be found at https://reghub.ai/documents (requires free login). The docucument data downloader gives the user access to pre-compiled sets of documents containing machine-readable text that can be used in text analysis research. This downloader is intended for computational social science, natural language processing, and other technical researchers who would like to do research on raw, machine-readable text.

The Python library for the API can be downloaded using `pip install regcensus` or directly from GitHub. See the README on the GitHub page for more details about how to use the package, or watch the video tutorial at https://www.quantgov.org/quantgov-api. The Python library is intended for users with experience using Python who would like to access more specific data not available through our bulk downloads or interactive downloader tool.

4 Technical Notes

- 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.

- For about half the states, we switched our source of text from the official website to the repository of regulatory and stautory codes at Casetext, which often provided cleaner, more consistent text and metadata, and a smoother webscraping experience. In cases where we found errors or inaccuracies in the Casetext repository, we stuck with the official government website as our source.

- Data from previous iterations of State RegData were updated for a few states due to webscraping errors in their development. Make sure to update these states with the new data if you are using data from previous versions for a current project. These are Delaware regulations, Maine regulations, Massachuetts regulations, Minnesota regulations, Missouri regulations, New Hampshire regulations, New York regulations, South Carolina regulations, and Vermont statutes. In addition, a few states' 2021 data were removed due to quality issues with the scrapings. These include Alabama statutes, Louisiana regulations, Maine statutes, Michigan statutes, New York statutes, and Pennsylvania statutes.

- The "cluster" data introduced in this version of State RegData was created by manually sorting agencies into clusters. While a lot of thoughtful time and effort was put in to ensure the accuracy of this sorting, judgement calls were made in rare cases where it was difficult to assign an agency to a cluster.

5 Contact

If you have any questions concerning our documents, data, methodology, or anything else about State RegData Definitive Edition, please email us at info@quantgov.org.

6 Citation

If you use this data, please cite:

McLaughlin, Patrick A. and Jonathan Nelson. State RegData Definitive Edition (dataset). QuantGov, Mercatus Center at George Mason University, Arlington, VA, 2021.

Current Version: 4.0