

RegData India Technical Documentation

QuantGov

January 20, 2021

1 Purpose

RegData India is part of the RegData project that looks to quantify the regulation in a given jurisdiction's regulatory and legislative code and other regulatory documents. RegData India currently includes federal legislation promulgated in the form of Acts of Parliament by the Government of India, as well as all state legislation related to education promulgated by the relevant state authorities. The quantified legislative documents will provide accurate and informative regulatory data that can be used in a wide variety of research applications.

2 Content

RegData India currently quantifies the Acts of Parliament for the Government of India in force as of 2020, and all state legislation related to education in force as of 2020.

3 Technical Notes

Federal and State:

- The text included in this dataset was provided in its original form by India's Centre for Civil Society (CCS).
- Even though each dataset has gone through extensive quality assurance checks, the data is only as good as the source documents. Some

source documents could have been not uploaded by government officials, uploaded to incorrect areas, or scanned in a format unreadable by computers. While the data is highly accurate, it is always recommended to spot check source documents when the data is not being used in aggregate.

- Original data consists entirely of documents stored as PDFs, the majority of which contained an embedded text layer from which we gathered the text, but a number of which needed the application of optical character recognition (OCR) technology to recognize text from scanned images. A number of the PDFs also suffered quality issues in terms of encoding errors occurring when attempting to collect text from the PDF’s embedded text layer, resulting in distorted text.
 - As a result of the above quality issues, the India dataset was aggregated through processing all of the original documents in two ways, one where all documents have OCR applied to them, and one where only documents that do not have embedded text layers have OCR applied. We then compare which version of the cleaned document text has the higher restriction count and select that version.
 - This process increased total restriction count over our normal procedure by a small amount, reflecting better results of applying OCR to some documents that we normally would have simply the embedded text layer from. This new process underwent internal QA to ensure that these increased restrictions were not false positives.
 - We manually removed Hindi language content from one document after cleaning, as it decoded into garbled text that artificially inflated the word count of that document. There may be other documents with minor amounts of Hindi in the title or preface of the page, but these incidents were not severe enough to be detected by two separate language detection algorithms applied to each document.
- Note: This branch of RegData includes a set of unique restrictive terms that were identified by the team at CCS. The methodology for deriving these particular terms can be found [here](#).

- Due to the conceptual nature of these unique terms, which searches not only for terms that restrict private action like ours, but also for terms to do with punishment or penalizing individuals, we will divide these terms and terms used for other RegData projects into two series.
- The first series contains only the restrictive terms used in other jurisdictions covered by the RegData projects, while a separate data series contains the particular terms used by the team at CCS, which is labeled “Restrictions and Punishments.” Analysts seeking to make direct comparisons between India and other jurisdictions should use the standard restrictive terms for their analysis. See the variable descriptions section for a complete list of terms used in each data series.

4 Citation

If you use this data, please cite:

McLaughlin, Patrick A. and Walter Stover. RegData India (dataset). QuantGov, Mercatus Center at George Mason University, Arlington, VA, 2020. <https://quantgov.org/regdata-india>.

Current Version: 1.0