Building Indian Parliamentary Datasets

Komal Preet Kaur[1], Vraj Patel[2]

Description

India's parliamentary system is one of the largest and most complex democratic institutions in the world. Yet, much of its legislative record remains inaccessible to researchers in a structured, analyzable format. This lack of accessible data has left major gaps in our understanding of how Parliament functions, how representatives behave, and how institutions evolve over time. Without structured data, it is difficult to trace patterns in party dynamics, legislative responsiveness, or the representation of marginalized groups. This limits the ability of scholars to test theories of democratic accountability, institutional performance, and political behavior in one of the world's most important democracies. It also restricts comparative research, making it harder to place India within broader global frameworks of legislative studies. In this proposal, we aim to build structured, ready-to-analyze datasets from India's Parliament. By building comprehensive, analyzable, publicly-accessible datasets from parliamentary records, we hope to enable researchers to advance the theories of political science and our understanding of Indian politics.

Currently, the only structured and analyzable parliamentary data available is of Question Time in the Lower House from 1999 to 2019 (Bhogale 2019). The vast majority of parliamentary proceedings such as debates, questions posed in the Upper House, committee compositions, speeches made by prime ministers, budget speeches, bills introduced, and other parliamentary activities, exist in disparate formats, including scanned PDFs, printed documents, and fragmented HTML pages, along with inconsistent metadata and inconsistent language. We propose to webscrape, digitize, translate, and structure Indian parliamentary data spanning multiple decades into analyzable formats. The goal is to transform these materials into digitized and standardized formats so that researchers can conduct rigorous, data-driven studies of legislative behavior, party dynamics, and agenda-setting to test and refine theories of representation, democratic accountability, and governance.

We believe this project aligns closely with the mission of the DDSS. Using computationally-intensive methods, we plan to produce a suite of open-access, structured datasets that can serve as foundational research tools. With the support of DDSS, structuring Indian parliamentary data into an analyzable format will open many avenues for scholars to study legislative behavior and institutional dynamics. Finally, as we plan to train four undergraduate and graduate students to carry out this project, they will receive hands-on experience in webscraping, data cleaning, natural language processing, and database design. The grant funding from DDSS will not only support the project's development but also build the technical capacity of students.

Current Progress

To date, we have webscraped the published English parliamentary debates from the eighth Lok Sabha session through the current session, spanning 1985 to 2024 in the Lower House. We are currently verifying, cleaning, annotating, and structuring this data into both an SQLite database and accessible dataframes. Our current repository contains approximately 47 GB of PDF-formatted data, with

[1]kk0014@princeton.edu. Postdoctoral Research Associate, Princeton School of Public and International Affairs.
[2]vrpa3077@colorado.edu. Institute of Behavioral Sciences, University of Colorado Boulder.

comprehensive metadata (date, file size, session number of the parliament, relative path, ...) stored in an SQLite database. In parallel, we have webscraped detailed records from question hour proceedings covering Lower House sessions from 1999 to 2024, including metadata on total questions submitted per session, ministry response rates, and the rotation and grouping of ministries for question-level data critical for understanding executive-legislative accountability. Grant support from DDSS will enable us to complete the ongoing work and expand our scope to include additional parliamentary records, such as Zero Hour interventions, prime ministerial speeches, budget presentations, no-confidence motions, and the complete corpus of questions posed and answered in both houses. With our current two-person team, this support is essential for scaling our data processing capabilities and ensuring speedy project completion.

Proposed Methods

We will employ Python-based web scraping frameworks, specifically Selenium for JavaScript content, and BeautifulSoup for static HTML. Given that legacy documents such as PDFs exhibit minor but inconsistent layout variations across parliamentary terms and statement types, we will implement specific structural logic across the extracted text. Optical character recognition (OCR) will be conducted primarily using Adobe Acrobat Pro's OCR tools to digitize scanned documents. Additionally, we plan to use deep learning OCR techniques and tools such as a finetuned Tesseract model applied with Stroke Width Transform (SWT) to enhance accuracy on low-quality scans. This hybrid OCR pipeline mitigates typical errors from legacy sources and improves text fidelity necessary for valid data. All raw extracted data, including OCR outputs, will be stored in a repository with integrated version control to ensure traceability and reproducibility throughout dataset construction. While we are still in the process of finalizing the key storage platform (e.g., GitHub), we hope to engage with experts at DDSS to learn about their suggestions and guidance.

To facilitate broad usability, cleaned and standardized datasets will be provided in supplementary Excel spreadsheets, formatted for ease of use for future analysis work. However, these Excel files will serve as secondary outputs; PostgreSQL will be utilized as the primary data storage and management system, supporting consistency, querying, and longitudinal tracking. We plan to develop a normalized relational database schema in PostgreSQL to organize the varied and hierarchical data structures inherent in parliamentary records: parliamentary terms, multiple sessions within each term, diverse statement types per session, and corresponding metadata. PostgreSQL is chosen for its complex relational queries and text search capabilities, which are great for handling multi-dimensional legislative data. The schema will include entities representing sessions, statements, ministries, members, and ministerial tenures. Unique identifiers will be assigned to parliamentarians, political parties, and legislative items to enable precise longitudinal analyses. Ministry and ministerial information will form linked reference data, allowing intelligent space search reductions in storage and enabling automated fuzzy string matching corrections (using the FuzzyWuzzy Python library) on text inconsistencies from OCR. This combination ensures both data completeness and quality with minimal manual correction burden. Finally, the codebase for scraping, OCR, cleaning, and database population will be fully open-source, modular, and parameterized to support reproducibility and future extensions.

References

Bhogale. 2019. "TPCD-IPD: TCPD Indian Parliament Dataset (Question Hour) 1.0.". Trivedi Centre for Political Data.

Budget

We request a USD 10,000 grant to recruit and train a team of four research assistants, who will gain hands-on experience and build a range of skills for their professional development. Their contributions will be central to accelerating the structuring of parliamentary records and ensuring the quality and usability of the final datasets in this project.

Table 1: Proposed Budget

Cost category	Amount (in USD)
Research assistance (4 RAs, USD 25/hour for 20 weeks)	10,000
Total	10,000