LEGDAT: A New Database for the Study of Legislative Processes and Lawmaking

Abstract

Legislatures are central to policy formulation, enactment, and oversight, yet comparative research has lacked data that systematically track how legislative proposals move through institutional stages and become law. This paper introduces LEGDAT, a new comparative database on legislative processes and lawmaking across 16 countries. Covering most cases from the 1990s through 2024, LEGDAT includes data on more than 350,000 bills and 60,000 enacted laws across diverse institutional settings. Bills are the primary unit of observation, and the database provides standardized information on their sponsors, the sponsors' institutional or partisan affiliations, bill status, committee involvement, procedural stages, final votes, passage or enactment dates, and time in parliament. It also includes the text of introduced bills and enacted laws and traces post-enactment modifications to measure legal durability and change. We illustrate its utility through an analysis of legislative production across electoral cycles in 16 countries and discuss additional applications for comparative legislative research. By providing standardized, scalable, and publicly available data, LEGDAT offers new resources for the comparative study of legislatures and public policy.

May 15, 2026

Mihaly Fazekas
Department of Public Policy, Central European University

Sebastian Thieme
Centre for European Studies and Comparative Politics (UMR CNRS 8239), Sciences Po
Sebastian.thieme@sciencespo.fr

Cyril Benoît
Centre for European Studies and Comparative Politics (UMR CNRS 8239), Sciences Po

Dominik Brenner
Department of International Relations, Central European University

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101132405. A previous version of this paper was circulated: Benoît, Cyril; Brenner, Dominik & Fazekas, Mihály (2024) "The LEGDAT Dataset: A Global Dataset on Legislative Processes, Outputs and Outcomes." CEU Global Corruption Observatory-Project Report. 2024 October. We are grateful to colleagues who offered their insights throughout the process, such as Matthew Stephenson and Jesper Johnson. We are also thankful for comments on earlier versions of the paper, especially from participants of the CEU DPP seminar (2022), Tag der Parlamentsforschung (2023), and Parliamentary Data for a Better Democracy (2022). This database was created thanks to numerous research assistants and computer scientists who contributed to the project. We are indebted to them.

Introduction

Legislatures are central venues for policy formulation, authorization, and oversight. They determine how issues reach the policy agenda, how proposals advance, and which become law. Yet comparative research has long lacked the tools to systematically analyze legislative processes across countries, constrained by the volume, complexity, and fragmentation of available data.

This paper introduces a new database – LEGDAT – that captures legislative processes across 16 countries, spanning, in most cases, from the 1990s to 2024. The database focuses on bills as the primary unit of observation, and includes over 350,000 bills and 60,000 laws. Moreover, it includes standardized information on key features of bill advancement such as the date of introduction, the status of a bill, the number of legislative stages to which a bill has advanced, potential dates of passage or enactment, and final votes on a bill. Our data also include key actors and committees associated with a bill's introduction or discussion. Crucially, we include the text of the initially introduced version of a bill and, if applicable, the enacted law. Moreover, we collect data on legislative stability by tracing modifications of laws associated with the bills in our sample. In sum, LEGDAT enables the comprehensive tracking of the policy process in and across national legislatures.

The scope of data collection – in terms of temporal depth, geographical coverage, and process-level detail – adds unique value to our contribution and allows comprehensive quantitative analyses of legislative processes and outputs in diverse institutional contexts. In what follows, we describe the data collection infrastructure, present summary statistics, and outline potential

applications for comparative legislative research. The data are made publicly available and regular updates are planned for the upcoming years.[1]

Motivation

The high availability of information on legislative websites (e.g., Inter-Parliamentary Union 2024) and advances in web scraping and big-data methods have enabled researchers to assemble large-scale and comparative datasets on legislatures and legislative processes.

For example, several multi-country datasets focus on the composition of legislatures, committees, or cabinets (e.g., Carnes et al. 2025; Döring and Manow 2024; Hellström et al. 2024, 2025; Turner-Zwinkels et al. 2022, Wäckerle et al. 2025). Other datasets are centered around legislative speech, assembling millions of speeches across predominantly European or United States legislatures (e.g., Gentzkow et al. 2018; Sylvester and Greene 2023; Erjavec et al. 2024; Rauh and Schwalbach 2020). A related line of work has collected legislative voting records, usually focused on individual countries (e.g., Eggers and Spirling 2014; Lewis et al. 2025; Louwerse et al. 2018).

In terms of legislative process and outputs, there is a small but growing number of projects enabling comparative research that are more closely related to our database. The Comparative Agendas Project (CAP) adopts a harmonized and structured approach, providing legislative text from around 25 countries, U.S states, and the European Union. It follows a shared data collection strategy that encompasses data on the executive, interest groups, and media (Baumgartner et al., 2019). While groundbreaking in the richness of the data and its extensive geographical scope, the database includes relatively few details on the legislative process (mainly year of introduction or passage) and most data points end in the mid 2010s.

[1] The link to the project website has been removed to preserve anonymity during peer review.

Furthermore, Frantzeskakis and Seeberg (2023) advance comparative legislative research by collecting data on bills or laws in 13 anglophone African countries. Another important source of information on laws and legislative stability is the FAOLEX database (FAO 2025), which collects laws and regulations, as well as information on legal changes in the areas of food, agriculture and natural resources management in over 200 countries. However, these datasets also contain few if any additional variables about the legislative process.

Most closely related to our work are datasets provided by ParlLawSpeech (Schwalbach et al. 2025) and König et al. (2022).[2] These datasets partially overlap with the data we collect, notably in bill and law text and dates of introduction and enactment and in some of the legislatures that are included.[3] At the same time, the projects differ in which bill-related data they collect additionally and their geographic scope. LEGDAT focuses largely on features of the lawmaking process such as who sponsored a bill, the stages at which a bill was scrutinized, including the committees involved, as well as modifications of laws after enactment. By comparison, ParlLawSpeech emphasizes legislative speech, and König et al. (2022) include a sectoral classification of bills. While there is an overlap in the geographic scope of the projects, notably in Europe and the United States, LEGDAT extends data collection beyond advanced capitalist democracies by including countries such as Colombia, India, Russia, and South Africa.

Our contribution with LEGDAT is thus two-fold. First, it permits extending analyses of the legislative agenda and legislative productivity by complementing existing data collection efforts. Specifically, it provides up-to-date data on proposed bills and enacted laws for

[2] In the U.S., projects with a similar focus have been undertaken by ProPublica at the federal level (https://projects.propublica.org/datastore/#congressional-data-bulk-legislation-bills) and Open States at the state level (https://open.pluralpolicy.com/data/).
[3] The Legislative Effectiveness Dataset takes a slightly different perspective by analyzing how many bills US legislators initiate (Volden & Wiseman 2014).

countries where such information was previously collected, as well as countries where no such efforts have been undertaken. Second, it expands on the existence of similar projects by focusing on additional features of the lawmaking process, from bill introduction through legislative deliberation to final enactment and post-enactment modification.

To our knowledge, no other comparative dataset provides comprehensive, process-oriented data that systematically map how legislation advances through distinct institutional stages. This gap has constrained efforts to test theoretical claims, identify cross-national patterns, and develop a comparative empirical research agenda on legislative behavior and policymaking. Formal theories of legislative organization, for example, often rely on assumptions about how institutions shape legislative behavior, processes, and outcomes (e.g., Cox and McCubbins 2012). Yet without systematic evidence on how legislators and parties behave across institutional contexts, these theories remain only partially tested. For scholars of policymaking, understanding legislatures is essential: agenda-setting analyses show which issues reach the legislative docket (Baumgartner et al. 2019), but without data on deliberation, enactment, and post-enactment modification, our understanding of policy development remains incomplete. The study of legislative capacity and performance (Arter 2007) is another domain in which progress has been limited by the absence of comprehensive and comparable data.

LEGDAT - Data Collection Framework and Database

LEGDAT offers a detailed, standardized, and scalable database that enables comparative analysis of legislatures and their role in shaping public policy. Below, we first describe the data collection process. Then, we provide a set of descriptive statistics to present the data contained in the first version of the database (LEGDAT 1.1).

Data Collection Framework

The case selection of 16 countries was driven by a number of factors, including the objectives of geographical breadth and having countries with a diverse set of institutional and political characteristics. The countries currently included are Australia, Bulgaria, Brazil, Chile, Colombia, France, Germany, Hungary, India, Jordan, Poland, Portugal, Russia, South Africa, the United Kingdom, and the United States.[4]

The basis for the standardized data collection in each country is a table of variables, where the variables are selected to comprehensively map the legislative process. In addition to information about the originators of legislative proposals and their partisan or institutional affiliations, LEGDAT reports both the status of a bill, the modifications it has encountered (captured by the size of the bill and, if applicable, the size of the law), various information pertaining to committee involvement and votes, and, if applicable, the number of days from introduction to passage. Table 1 provides an overview of the variables in the first version of the LEGDAT database. Please note that for variables that assume enactment into law (e.g., technical law) the values for non-enacted bills are coded as missing.

[4] The goal for this ongoing project is to continuously increase coverage of all countries where parliaments are key to lawmaking and where technical constraints do not make data collection infeasible.

TABLE 1. Summary of the main variables

Variable name	Description	Type
record_id	Unique bill ID assigned by the research team	string
bill_id	Unique source ID for the passed law/bill	string
bill_title	Short title of the passed bill/law	string
origin_type	Originator of a bill proposal: individual MP, government	categorical
originator_name	names of the MPs or government body/bodies	list
originator_affiliation	Institutional/party affiliation of the originator	list
bill_status	Passed/rejected by the legislature or ongoing (bills merged into other bills coded as rejected)	categorical
affecting_laws_count	Count of laws which modify the enacted law	numeric
affecting_laws_first_date	Date when the enacted law was first modified	date
date_introduction	Date of introduction to parliament	date
date_passing	Date of passing the law/enactment of the law	date
time_in_parl_days	Number of days a law spent in the legislature	numeric
bill_size	Size of the bill upon introduction	numeric
bill_text	Full text of initially introduced bill	string
law_size	Size of the enacted law	numeric
law_text	Full text of the enacted law	string
stages_count	Total number of legislative stages	numeric
committee_count	Count of committees involved with the bill	numeric
final_vote_in_favour	Number of votes in legislature in favor of a bill	numeric
final_vote_against	Number of votes in legislature against of a bill	numeric
final_vote_abst	Number of final absentee votes	numeric
annual_average_number_of_modifications	Annual average number of modification that a law received	numeric
standard_law	TRUE = standard law, FALSE = technical law (less than 4,000 characters)	categorical
omnibus_indicator	Laws larger than 2 sd from country-specific law text mean	logical
fast_tracked_procedure_indicator	Laws passed within 10 days after introduction	logical
intiator_indicator	Non-government initiated laws	logical
modification_indicator	Laws modified more than once a year on average, after enactment	logical

The data collection process can be divided into 4 stages: 1) source identification, 2) source annotation to provide instructions for web scraping and parsing of data, 3) web scraping and parsing of data, 4) data evaluation, with the last two or three stages typically repeated until the quality of the data is sufficiently high.

In the first stage we identify the relevant sources for the data. Depending on the national context, information may be distributed across different websites in the same country. Identifying and evaluating the target legislative data source includes assessing its data content and technical features such as the format of text publications (pdf, html, etc).

In the second stage, the detailed annotation of the source requires that each variable in our variable template is mapped onto specific locations on the target website. This process is aided by a variable guide with example screenshots of the location of each variable on the website, an explanation on how to reach this location (e.g., a URL) and a description of potential issues with scraping the variable (e.g., pop-up windows or linking variables across multiple websites). They also include instructions on how to recode the scraped data to conform to our variable template. These tasks were completed by research assistants with knowledge of the country's language and familiarity with legislative processes under the supervision of the authors.

After these preparations, the third stage is to implement collection of the data. This includes writing algorithms to automatically scrape or download relevant information from the target websites. Further, the semi-structured data are coded into standard data tables, following our data standard (i.e., relational database). To address technical constraints and speed up the process of data collection, we rely on the services of an IT company to implement the third stage. To reduce complications in the technical implementation, we used regular meetings and ongoing communication to coordinate the work of research assistants and the IT company. The

algorithms for scraping and parsing the data will be made publicly available to increase transparency and replicability.

Fourth, we thoroughly check data quality, both in terms of scope and accuracy. Scope is checked by comparing data record counts against website record counts and by detecting comparatively low numbers of observations in a given year. Data accuracy is algorithmically checked by reviewing missing rates and zero-values rates for each variable. We use high or systematic missing rates as a basis for investigating the data collection for errors. Moreover, randomly selected samples of records are manually checked by research assistants, variable by variable, against the source data.[5] This process is used to create a list of data collection errors. The errors are then fed back to stages 2 and 3 to correct either the instructions for or the implementation of web scraping and data parsing. We cycle through data collection and evaluation until we do not identify further errors in the data fields or data fixes are not currently feasible, for example, when laws are provided as scanned pdf documents rather than machine-readable text.[6]

Descriptive Statistics

In this section, we first describe the coverage of the first version of the LEGDAT dataset and show key summary statistics for our main variables. Table 2 provides an overview of the coverage of the database by country. Overall, our data on 16 countries span multiple decades and include over 350,000 observations on bills and laws. To count a bill or law as an

[5] To obtain relevant samples across different sets of variables, the process is implemented at the levels of proposed bills, enacted laws, and, if applicable, different types of legislative procedures.
[6] In the Supplemental Appendix, we validate our data collection effects by comparisons of bill and law counts to CAP (Baumgartner et al., 2019) and ParlLawSpeech (Schwalbach et al. 2025), and via a systematic assessment of completeness and accuracy of the variables. For any remaining gaps in coverage and accuracy, pending further updates to the data, this transparency allows researchers to make informed decisions when using the data for their research.

observation, we require either a substantive title or the text of the bill.[7] LEGDAT's coverage for most countries goes back at least to the 1990s and generally extends to 2024.[8] While the data for a number of countries include observations from earlier years, here, we count as the start of our coverage the first year in a given country where the number of observations is greater than 25 percent of the median number of observations.

[7] If there is no law text, we also require that the value of the bill_status variable is equal to "PASS".
[8] The year of the observations is taken from the date of introduction (bills) or the date of passage or enactment (laws). In some cases, where these dates are missing, the year was obtained from the title.

TABLE 2. Coverage of the Dataset by Country

Country	Coverage (Years)	n (obs)	n (bills)	n (laws)
Australia	1997-2024	6,038	6,038	4,001
Bulgaria	2001-2024	5,675	5,675	1,851
Brazil	2001-2024	72,857	72,046	2,887
Chile	1900-2024	25,630	24,528	14,251
Colombia	1998-2024	10,924	10,924	1,678
France	2009-2024	1,721	1,721	504
Germany	1973-2024	9,683	9,683	6,096
Hungary	1998-2024	6,103	6,103	3,401
India	1952-2024	9,598	9,598	3,639
Jordan	1923-2022	2,064	–	2,064
Poland	1998-2024	5,155	–	5,155
Portugal	1977-2024	12,290	12,290	4,917
Russia	1996-2022	19,796	19,796	6,305
South Africa	2006-2024	692	692	557
United Kingdom	2006-2024	3,465	3,465	608
United States	1993-2024	161,846	161,846	5,355
Total		353,537	344,405	63,269

The sample frame includes as observations all bills that lead to changes in law. For enacted bills, we also collect information on law-specific variables. In Jordan and Poland, the data currently include enacted legislation, but not ongoing or rejected bills. Moreover, in Brazil and Chile, we have information about laws, but not bills more generally, at the beginning of the coverage period, resulting in a higher number of observations than bills.

The table shows a lot of cross-country variation in the proportion of bills that become laws. For example, in Australia or Germany the proportion is approximately two out of three, compared to less than five percent in Brazil or the United States. Such variability is not surprising, given that bills are proposed not only to make law but also to take positions and send signals about the policies they would like to implement given sufficient majorities (e.g., Schiller 1995).

Next, Table 3 reports summary statistics for the main variables in LEGDAT. For numeric variables, we present the median, mean, and standard deviation. For nominal variables (*originator_name, originator_affiliation, bill_status, standard_law*) we report the mode, and for date variables, we provide the median date.

The results highlight several features of the dataset. First, approximately half of all bills were introduced before May 2012, reflecting both the historical coverage of the database and the growing volume of legislation in more recent decades. Second, the size of bills and laws is markedly right-skewed: the majority are relatively short, under 4,000 characters, but a minority of very large legislative texts raise the mean and standard deviation considerably. This distribution suggests that most bills are concise, while a small set of omnibus or highly detailed laws account for the upper tail.

Third, the statistics on committee_count indicate that the median bill is referred to at least one committee, emphasizing the central role of committees in legislative scrutiny across diverse institutional contexts. Finally, the modal categories of the nominal variables provide additional insight into recurring institutional patterns, such as the predominance of particular types of originators, the most frequent bill statuses, and the typical classification of laws.

TABLE 3. Summary Statistics for the Main Variables

Variable name	Key statistics
origin_type	Mode: Non-Government (86.4%)
originator_name	Mode: Câmara dos Deputados
originator_affiliation	Mode: Democrats
bill_status	Mode: Reject (62.17%)
affecting_laws_count	Median: 0; Mean: 0.95; St.Dev: 5.40
affecting_laws_first_date	Median: 2006-06-22
date_introduction	Median: 2012-05-07
date_passing	Median: 2003-07-02
time_in_parl_days	Median: 141; Mean: 249.52; St.Dev: 378.08
bill_size	Median: 3,804; Mean: 14,908.83; St.Dev: 83,658.79
law_size	Median: 3152; Mean: 21,334.04; St.Dev: 99,859.62
stages_count	Median: 2; Mean: 2.53; St.Dev: 2.72
committee_count	Median: 1; Mean: 1.42; St.Dev: 1.49
final_vote_in_favour	Median: 349; Mean: 306.99; St.Dev: 121.42
final_vote_against	Median: 1; Mean: 30.51; St.Dev: 60.02
final_vote_abst	Median: 1; Mean: 10.92; St.Dev: 27.92
Annual_average_number_of_ modifications	Median: 0; Mean: 0.08; St.Dev: 0.77
standard_law	Mode: Standard Law (90.45%)
omnibus_indicator	Median: 0; Mean: 0.03; St.Dev: 0.18
fast_tracked_procedure_indicator	Median: 0; Mean: 0.04; St.Dev: 0.19
intiator_indicator	Median: 0; Mean: 0.34; St.Dev: 0.47
modifications_indicator	Median: 0; Mean: 0.01; St.Dev: 0.12

Due to the large number of observations from a few countries, such as Brazil and the United States, some of these statistics are not representative across all countries. To illustrate variation across the countries in LEGDAT, we present radar charts with summary statistics in Supplemental Appendix A2.

Discussion of Potential Applications

LEGDAT provides a significant new foundation for comparative research on legislatures by offering structured and temporally extensive data on legislative processes across diverse institutional settings. In the remainder of this paper, we would like to emphasize its potential to address longstanding questions in legislative studies and public policy, as well as to suggest a few applications.

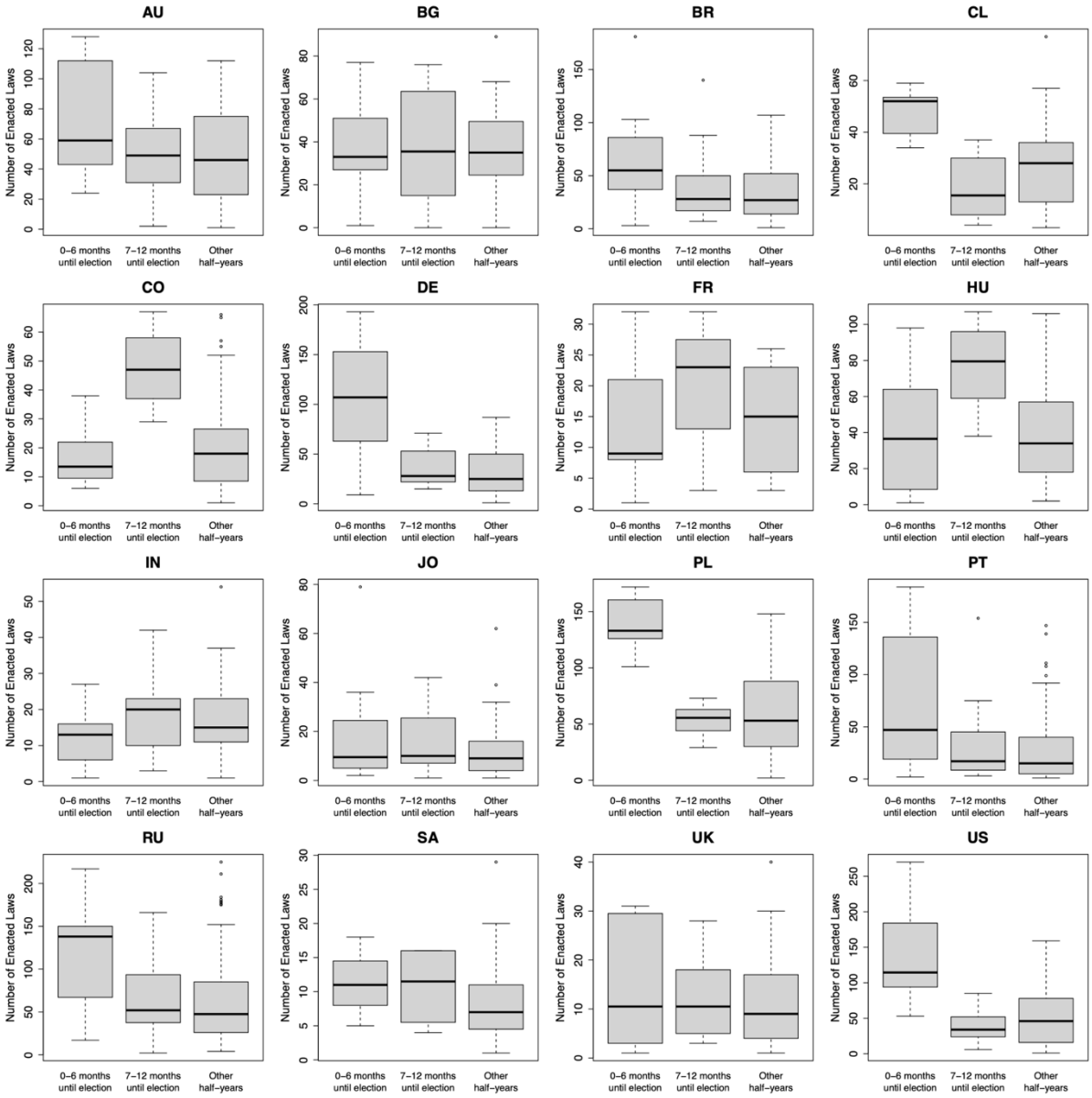
First, LEGDAT creates new opportunities to evaluate core institutional theories in legislative studies, by including detailed and structured data on bill initiation, procedural advancement, and final passage. These features facilitate disaggregated analyses of legislative throughput, bottlenecks, and delay, as well as direct tests of classic institutional theories regarding bicameralism, executive-legislative balance (Döring 1995), or veto players (Tsebelis 2002) and governments' capacity for policy change (Angelova et al. 2018).

Second, LEGDAT facilitates the study of political dynamics *within* legislatures, including agenda control, gatekeeping, legislative conflict, coalition management, and opposition behavior. These questions are central to formal models and empirical studies of legislative organization (Bäck et al. 2022; Cox and McCubbins 2012). By identifying bill originators, distinguishing government from non-government initiatives, and recording fast-tracked procedures and omnibus bills, LEGDAT allows scholars to examine how institutional rules and legislative roles structure the flow of legislation across systems that vary in party discipline, institutional centralization, and executive dominance. Measures of vote margins and sponsor affiliation provide further leverage on party cohesion, legislative strategy, coalition governance, and legislative productivity. Researchers can assess whether coalition partners support one another's bills, whether coalition agreements or bargaining duration condition legislative productivity, and whether voting behavior changes across legislative stages or

government cycles. In multiparty systems, LEGDAT enables analyses of opposition coordination, obstruction, and policy signaling (Martin and Vanberg 2011; Carroll and Cox 2012; Kam 2009), particularly when combined with external datasets on coalition formation, governance, and termination (Hellström et al. 2024). In presidential systems, it supports disaggregated analyses of executive-legislative relations, including whether executive-sponsored bills are more likely to pass and whether partisan alignment between branches increases legislative success (Mainwaring and Shugart 1997; Cheibub, Przeworski, and Saiegh 2004). LEGDAT's standardized structure allows these analyses to travel across regime types, party systems, constitutional structures, and time.

Third, LEGDAT advances the study of legislative outcomes, legal stability, and democratic responsiveness. For example, the data can be used to detect and analyze political legislation cycles (e.g., Padovano and Sy 2025). Evidence for increases in the number of enacted laws leading up to elections have been linked to legislators' motivations for signaling competence (Padovano and Petrarca 2012) and obscuring the consequences of legislation (e.g., Gratton et al. 2021). In the panels of Figure 1, we show the number of laws passed across different legislative election cycles in each country. Specifically, we aggregate the number of laws into half-year intervals from each election and plot the distribution of counts across electoral cycles for different groups of half-years (one, two, or more before an election). The data provide indicative evidence for the existence of political legislation cycles in most of the countries in the sample, including in non-democratic Russia.

FIGURE 1. Legislative Production Across Electoral Cycles of Legislatures in 16 Countries



In addition, a distinctive contribution of LEGDAT is its focus on legislative durability and change. It records whether an enacted law is subsequently modified, how soon the first change occurs, how many modifying laws it accumulates, and how frequently it is updated on average. These features allow scholars to examine the long-term trajectory of legislation and create new opportunities for testing theories of policy feedback, punctuated equilibrium (Baumgartner et al. 2019) and policy revision cycles (Maltzman and Shipan 2008). For example, researchers

might explore whether laws passed under accelerated procedures are more prone to revision or whether legislative instability varies across political systems and time periods.

In sum, LEGDAT provides infrastructure for legislative research at breadth and depth, enabling the empirical testing of long-standing theoretical claims and the development of new research agendas. Importantly, the structure and content of LEGDAT also make it highly amenable to integration with other datasets. Its standardized format across countries and time periods supports multi-level research designs that combine legislative data with sources on party systems, public opinion, cabinet composition, or media coverage. Moreover, the inclusion of full legal texts opens the door to natural language processing applications, such as tracking rhetorical strategies, detecting policy content, or classifying laws by policy domain or ideological tone.

References

Angelova, Mariyana, Hanna Bäck, Wolfgang Müller, and Daniel Strobl (2018) "Veto Player Theory and Reform Making in Western Europe," *European Journal of Political Research*, 57: 282-307.

Arter, David (2007) "Introduction: Comparing the Legislative Performance of Legislatures," *The Journal of Legislative Studies* 12: 245-257.

Bäck, Hanna, Wolfgang Müller, Mariyana Angelova, and Daniel Strobl (2022) "Ministerial Autonomy, Parliamentary Scrutiny and Government Reform Output in Parliamentary Democracies," *Comparative Political Studies* 55(2): 254-286.

Baumgartner, Frank.R., Breunig, Christian, and Grossman, Emiliano (2019) *Comparative Policy Agendas: Theory, Tools, Data*. Oxford University Press.

Carnes, Nicholas, Jushua Ferrer, Miriam Golden, Esme Lillywhite, Noam Lupu, and Eugenia Nazrullaeva (2025) "The global legislators database: Characteristics of national legislators in the world's democracies," *British Journal of Political Science 55*: e27.

Caroll, Royce, and Gary Cox (2007) "The Logic of Gamson's Law: Pre-election Coalitions and Portfolio Allocations," *American Journal of Political Science* 51(2): 300-313.

Cheibub, José A., Adam Przeworski and Sebastian M. Saiegh (2004) "Government Coalitions and Legislative Success Under Presidentialism and Parliamentarism," *British Journal of Political Science* 34(4): 565-587.

Cox, Gary, and Matthew McCubbins (2012) *Legislative Leviathan: Party Government in the House*. 2nd ed. New York: Cambridge University Press.

Döring, Herbert (ed.) (1995) *Parliaments and majority rule in Western Europe*, New York: St Martin's Press.

Döring, Holger, and Philip Manow (2024) Parliaments and governments database (ParlGov): Information on parties, elections and cabinets in established democracies. Available on: https://www.parlgov.org. Date of access: September 24, 2025.

Eggers, Andrew, and Arthur Spirling (2014) "Electoral security as a determinant of legislator activity, 1832–1918: New data and methods for analyzing British political development," *Legislative Studies Quarterly* 39(4): 593-620.

Erjavec, Tomaž, Matyáš Kopp, Nikola Ljubešić, et al. (2024) "ParlaMint II: advancing comparable parliamentary corpora across Europe," *Lang Resources & Evaluation* 59: 2071-2102.

FAO. FAOLEX Database. Complete Collection. License: CC BY-NC-SA 3.0 IGO. Extracted from: www.fao.org/faolex/opendata. Date of Access: May 13, 2025.

Frantzeskakis, Nikolaos, and Henrik Bech Seeberg (2023) "The legislative agenda in 13 African countries: A comprehensive database," *Legislative Studies Quarterly* 48(3): 623-655.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16. Available on: https://data.stanford.edu/congress_text. Date of access: September 24, 2025.

Gratton, Gabriele, Luigi Guiso, Claudio Michelacci, and Massimo Morelli (2021) "From Weber to Kafka: Political instability and the overproduction of laws," *American Economic Review 111*(9): 2964-3003.

Hellström, Johan, Torbjörn Bergman, Jonas Lindahl, Hanna Bäck, Gabriella Ilonszki, Wolfgang C. Müller, and Kaare Strøm (2024) Party Government in Europe Database (PAGED) – Coalition Governance in Central Eastern and Western Europe Dataset, Version 2024.12. Available on https://repdem.org. Date of access: September 24, 2025.

Hellström, Johan, Torbjörn Bergman, Jonas Lindahl, and Elsa Karlsson Gustafsson (2025). The Representative Democracy Data Archive (REPDEM) - Basic dataset, Version 2025.03. Available on https://repdem.org.

Inter-Parliamentary Union (2024) World E-Parliament Report 2024. Available on: https://www.ipu.org/resources/publications/reports/2024-10/world-e-parliament-report-2024. Date of access: September 24, 2025.

Kam, Christopher (2009) *Party Discipline and Parliamentary Politics*. New York, Cambridge University Press.

König, Thomas, Nick Lin, Xiao Lu, Thiago N. Silva, Nikoleta Yordanova, and Galina Zudenkova (2022) "Agenda control and timing of bill initiation: A temporal perspective on coalition governance in parliamentary democracies," *American Political Science Review*, 116(1), 231-248.

Lewis, Jeffrey, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet (2025). Voteview: Congressional Roll-Call Votes Database. Available on: https://voteview.com/. Date of access: September 24, 2025.

Louwerse, Tom, Simon Otjes, and Cynthia van Vonno (2018) "The Dutch Parliamentary Behaviour Dataset," *Acta Politica* 53: 149-166.

Mainwaring, Scott, and Matthew S. Shugart (1997) *Presidentialism and Democracy in Latin America*. New York: Cambridge University Press.

Maltzman, Forrest, and Charles R. Shipan (2008) "Change, Continuity, and the Evolution of the Law," *American Journal of Political Science* 52(2): 252-267.

Martin, Lanny W., and Georg Vanberg (2011) *Parliaments and coalitions: The role of legislative institutions in multiparty governance.* Oxford University Press.

Padovano, Fabio, and Ilaria Petrarca (2012) “A theory of cyclical production of laws and decrees,” *Centre Condorcet for Political Economy Working Paper*, (9).

Padovano, Fabio, and Youssoufa Sy (2025) “Conditional political legislation cycles,” *International Review of Law and Economics*, 106305.

Rauh, Christian, and Jan Schwalbach (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. Available on: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/L4OAKN. Date of access: September 24, 2025.

Schiller, Wendy (1995) “Senators as political entrepreneurs: Using bill sponsorship to shape legislative agendas,” *American Journal of Political Science* 39(1): 186-203.

Schwalbach, Jan, Lukas Hetzer, Sven-Oliver Proksch, Christian Rauh, and Miklós Sebők (2025). “ParlLawSpeech”; GESIS, Cologne. Available on: https://parllawspeech.org. Date of access: September 24, 2025.

Sylvester, Christine, and Zachary Greene (2023). "ParlEE plenary speeches V3 data set: Annotated full-text of 10 million sentence-level plenary speeches of six EU legislative chambers". Available on: https://doi.org/10.7910/DVN/IETNNB (Harvard Dataverse, V1). Date of access: September 24, 2025.

Tsebelis, George (2002) *Veto Players: How Political Institutions Work*. Princeton University Press.

Turner-Zwinkels, Tomas, Oliver Huwyler, Elena Frech, Philip Manow, Stefanie Bailer, Niels D. Goet, Simon Hug (2022) “Parliaments Day-by-Day: A New Open Source Database to Answer the Question of Who Was in What Parliament, Party, and Party-group, and When,” *Legislative Studies Quarterly* 47(3): 761–784.

Volden, Craig, and Alan Wiseman (2014) *Legislative Effectiveness in the United States Congress. The Lawmakers*. Cambridge University Press.

Wäckerle, Jens, Bruno Castanho Silva, Danielle Pullan, and Firuze Taner (2025) The Committee Membership Dataset (CMD) *(Version 1.0.0) [Data set]. GESIS, Cologne. https://doi.org/10.7802/2940*

LEGDAT: A New Database for the Study of Legislative Processes and Lawmaking

Supplemental Appendix

Supplemental Appendix A1: Validation of Data Collection Efforts

This section provides additional information on the level of completeness and data quality and accuracy of the current version of LEGDAT (1.1). We first compare the bill and law counts in LEGDAT to CAP (Baumgartner et al., 2019) and ParlLawSpeech (Schwalbach et al. 2025). Then, we provide evidence of the extent to which values are missing in our data. Lastly, we provide an assessment of the quality of the data in LEGDAT compared to the availability of the data at the source.

Supplemental Appendix A1.1: Comparing LEGDAT to existing datasets.

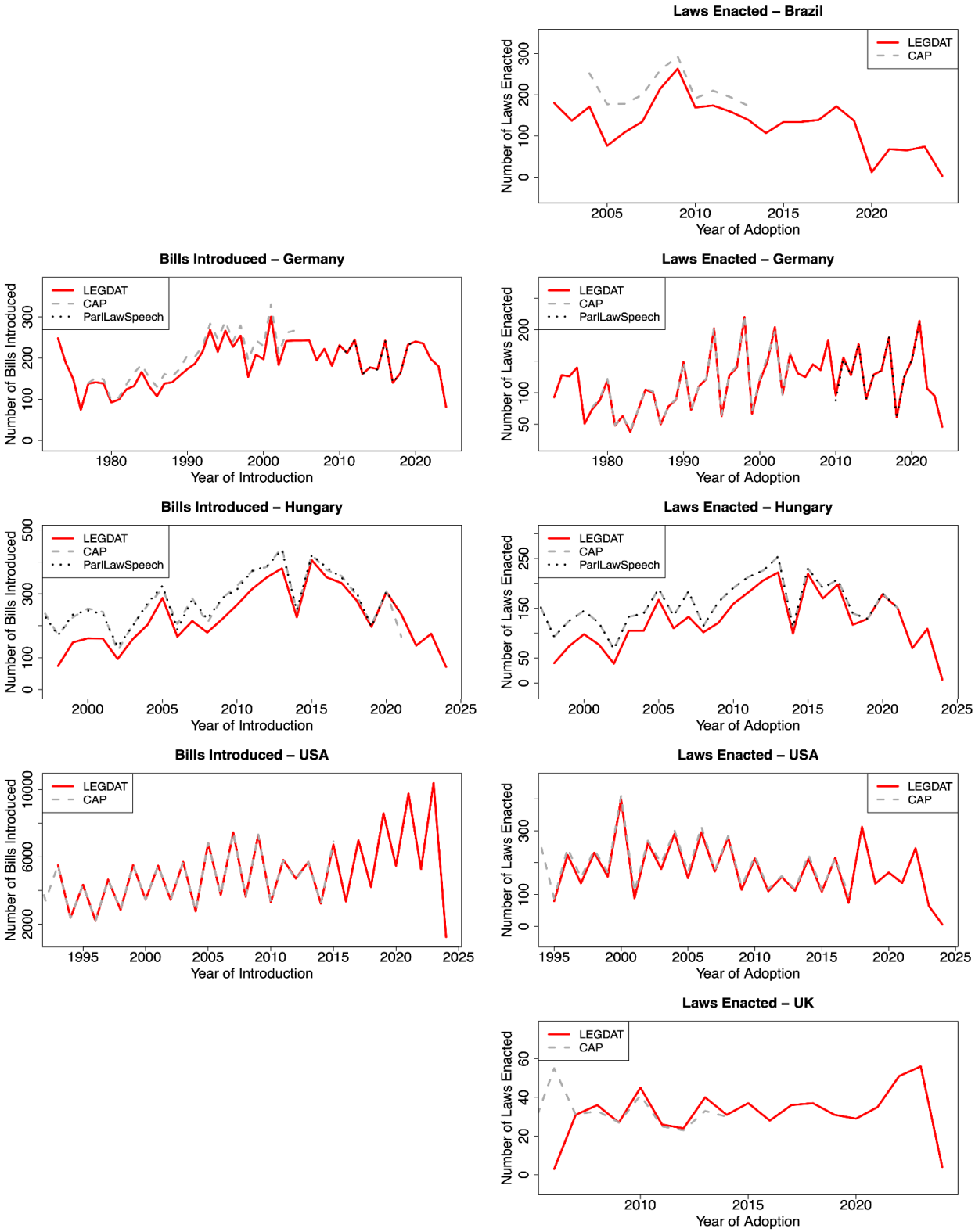
To validate our data collection efforts, we compare the number of introduced bills per year and the number of enacted laws per year in LEGDAT to the respective numbers in two comparable projects, CAP (Baumgartner et al., 2019) and ParlLawSpeech (PLS) (Schwalbach et al. 2025). Both projects have partial overlaps with LEGDAT in country-year coverage, though for some countries, CAP only collected laws. Specifically, information on laws is available for Brazil (CAP), Germany (CAP and PLS), Hungary (CAP and PLS), the United Kingdom (CAP), and the United States (CAP). In addition, both CAP and PLS have information on bills in Germany and Hungary, while CAP collected data on bills in the United States.

In comparing the numbers, we focus on the date of introduction for bills and the year of adoption for laws.[1] To make the data more comparable, we exclude the first and last years of the data series for each compared project, as data collection in a given project may not have started with the beginning of the first calendar year and ended with the end of the last calendar year. In addition, we exclude simple resolutions from CAP data in the United States, as LEGDAT is focused on bills that can change laws. The comparisons are shown in the panels of Figure 1, with bills in the left column and laws in the right column.

Overall, we see a strong correspondence between bills and laws in LEGDAT and the other two projects for the country-years where there is overlap. For some countries and years, there appear to be fewer observations in LEGDAT, for example in Hungary. To the extent that this is reflective of missed observations of bills and laws we intend to capture, subsequent iterations of LEGDAT will seek to expand data collection to ensure comprehensive coverage. The higher number of laws in LEGDAT compared to CAP for the United Kingdom can be attributed to LEGDAT's inclusion of private bills.

[1] Depending on the country, this variable currently captures the date of final passage in the legislature or the date of enactment into law.

FIGURE A1. Comparing LEGDAT to the Comparative Agendas Project and ParlLawSpeech



Supplemental Appendix A1.2: Overview of missing values.

Table A1 presents the percentage of missing values for each country and variable in our sample frame. As the values of some variables are only meant to be collected for bills that become law, we provide both the percentage of missing values in the entire sample (top value) and the percentage of values in the sample of bills which passed the legislative process. We note that the missing values are calculated based on the observations in our sample. As highlighted in the paper, upcoming data collection efforts will seek to not only improve the quality and completeness of values for observations in the samples but also to obtain bills and laws that are still missing from our samples during the respective coverage periods. This includes the following types of bills: “Acto Legislativo” bills in Colombia, “Loi Constitutionelle" and "Loi Organique" bills in France, rejected and ongoing bills in Poland, “Projeto de Revisao Constitutional” in Portugal, and joint resolutions (“House Joint Resolution” and “Senate Joint Resolution”) in the United States. Similarly, we will address the issue of duplicate bills in France that result from the re-initiation of previously introduced legislation.

Table A1. Percent of missing observations by, variable and country

	AU	BG	BR	CL	CO	DE	FR	HU	IN	JO	PL	PT	RU	SA	UK	US
record_id	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bill_id	0	0	0	45	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	81	0	0	1	0	0	0	0	0	0	0	0	0
bill_title	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
origin_type	0	24	6	52	1	0	0	0	0	100	0	0	0	1	0	0
	0	8	1	85	2	0	0	0	0	100	0	0	0	0	0	0
originator	10	58	6	54	1	1	0	0	47	100	81	20	50	0	1	0
	9	76	1	93	1	0	0	0	100	100	81	41	66	0	5	0
originator_affiliation	18	58	11	96	2	1	69	50	19	100	31	20	100	10	1	0
	9	76	51	93	3	0	70	82	29	100	31	41	100	3	5	0
bill_status	0	39	4	0	0	0	0	0	4	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
affecting_laws_count	37	99	96	45	85	84	100	91	0	28	0	98	37	82	92	100
	5	98	21	1	0	75	100	84	0	28	0	94	37	78	55	91
affecting_laws_first_date	81	99	99	93	98	84	100	91	98	91	71	98	98	83	92	100
	71	98	90	87	84	75	100	84	95	91	71	94	97	79	55	94
date_introduction	3	0	2	45	1	0	0	0	0	0	0	5	0	2	0	0
	2	0	32	81	4	0	2	0	0	0	0	13	0	3	0	0
date_passing	35	67	96	45	85	37	72	44	62	0	0	64	67	33	83	97
	3	0	29	1	0	0	0	0	2	0	0	0	0	16	2	1

Table A1 (continued). Percent of missing observations, by variable and country

	AU	BG	BR	CL	CO	DE	FR	HU	IN	JO	PL	PT	RU	SA	UK	US
time_in_parliament_days	35	67	98	90	85	37	72	44	63	100	1	65	68	34	83	97
	3	0	62	82	4	0	2	0	6	100	1	13	0	18	2	1
bill_size	11	67	1	52	18	38	3	9	37	100	15	32	27	1	37	12
	8	0	20	84	17	47	3	8	39	100	15	44	20	0	13	10
bill_text	11	67	1	52	18	38	3	9	37	100	15	32	27	1	37	12
	8	0	20	84	17	47	3	8	39	100	15	44	20	0	13	10
law_size_char	37	67	96	45	85	38	71	77	92	0	0	73	70	21	85	97
	5	0	25	1	0	1	0	59	80	0	0	34	5	2	14	21
law_text	37	67	96	45	85	48	71	77	92	1	0	73	70	21	85	97
	5	0	25	1	1	18	0	59	80	1	0	34	5	2	14	21
stages_count	0	43	83	45	0	0	0	1	0	100	0	0	0	0	0	4
	0	0	33	81	0	0	0	0	0	100	0	0	0	0	0	5
committee_count	0	0	0	45	0	0	0	2	100	100	2	0	0	0	100	0
	0	0	0	81	0	0	0	0	100	100	2	0	0	0	100	0
final_vote_for	94	89	100	100	98	98	100	43	100	100	1	37	76	100	100	99
	93	67	97	100	89	97	100	0	100	100	1	6	26	100	100	67
final_vote_against	94	89	100	100	98	98	100	43	100	100	1	60	76	100	100	99
	93	67	97	100	89	97	100	0	100	100	1	64	26	100	100	67
final_vote_abst	100	89	100	100	98	98	100	43	100	100	1	69	76	100	100	99
	100	67	97	100	89	97	100	0	100	100	1	63	26	100	100	68
average_annual_nr_of_modifications	37	99	96	45	85	84	100	91	62	28	0	98	79	84	92	100
	5	98	29	1	0	75	100	84	2	28	0	94	37	80	56	91

Table A1 (continued). Percent of missing observations, by variable and country

	AU	BG	BR	CL	CO	DE	FR	HU	IN	JO	PL	PT	RU	SA	UK	US
standard_law	37	67	96	95	85	48	100	77	92	100	0	73	70	21	85	97
	5	0	25	92	1	18	100	59	80	100	0	34	5	2	14	21
omnibus_indicator	37	67	96	45	85	64	71	77	93	1	12	73	70	34	85	97
	5	0	32	1	1	42	0	59	82	1	12	34	5	17	16	21
fast_tracked_procedure_indicator	35	67	98	90	85	37	72	44	62	100	1	65	68	34	83	97
	3	0	60	82	4	0	2	0	3	100	1	13	0	18	2	1
initiator_indicator	35	70	96	92	85	37	71	44	62	100	1	60	68	32	83	97
	3	8	29	85	2	0	0	0	3	100	1	0	0	16	2	1
modifications_indicator	35	67	96	45	85	37	71	91	62	0	0	98	68	32	92	97
	3	0	29	1	0	0	0	84	2	0	0	94	0	16	56	1

Note: The table presents the percentage of missing values for each country and variable in the respective samples. As the values of some variables are only meant to be collected for bills that become law, we provide both the percentage of missing values in the entire sample (top value) and the percentage of values in the sample of bills which passed the legislative process. For bill and law text lengths, zero-valued observations are considered missing. These assessments refer to the current version of the data (LEGDAT 1.1).

Supplemental Appendix A1.3: Overview of data quality compared to source values.

Table A2 provides an overview of the current data quality and accuracy of the observations in our samples compared to data available at the respective sources. This includes whether we are accurately capturing the fact that data are missing at the source. The assessments are provided in an abridged format, as "+", "+/-", and "-", indicating different levels of quality and caution that researchers should apply when using the variables. "+" indicates an approximate accuracy level of 95%-100% given the available values on the website for the bills in our sample. "+/-" indicates an approximate accuracy level of 70%-95% and/or that variables should be used or relied upon with caution. "-" indicates an approximate accuracy level below 70% and/or that variables should only be used or relied upon with great caution. Additional information on sources of error are provided in data documentation files.

Table A2. General assessment of current data quality compared to data available at source, by variable and country

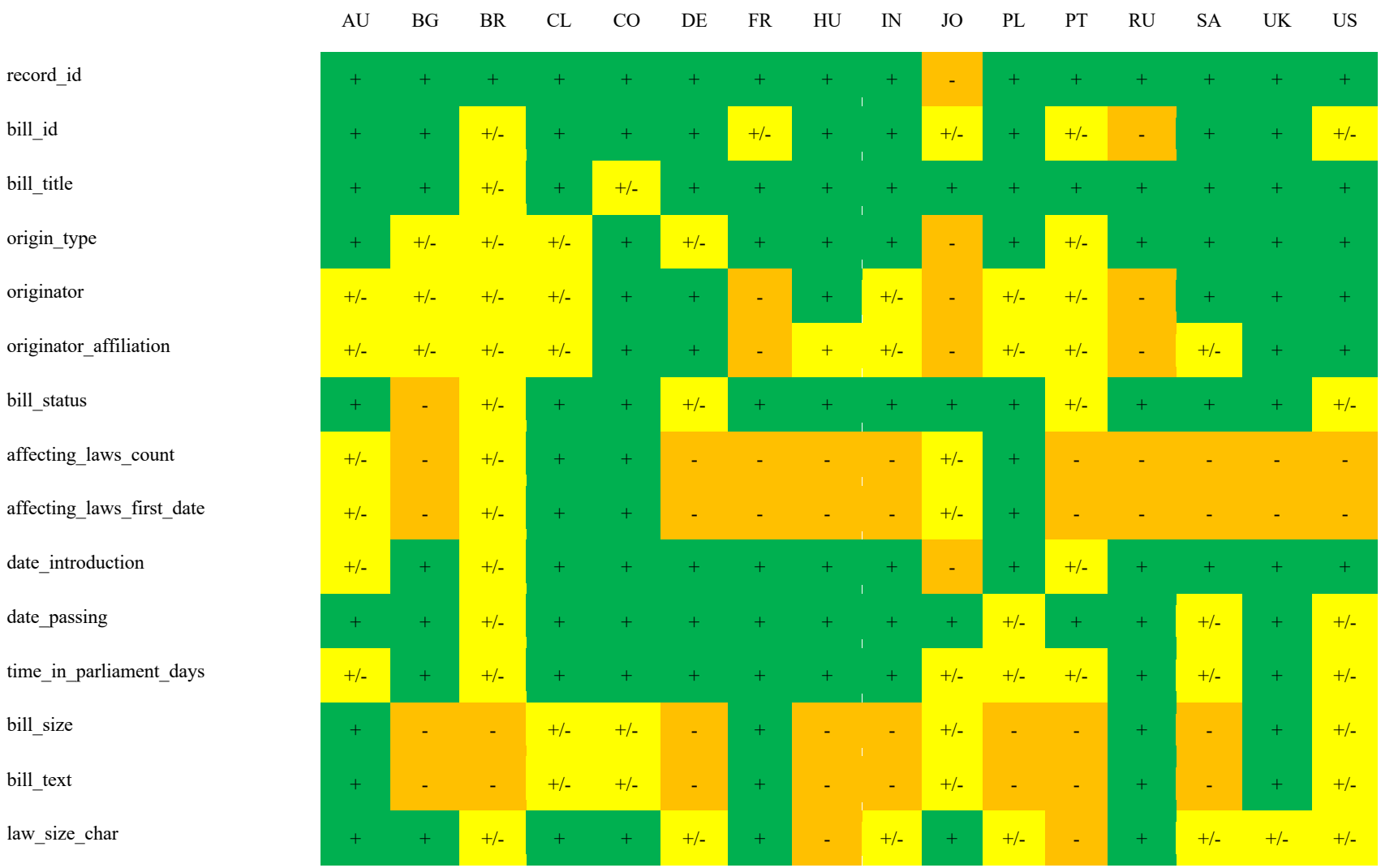


Table A2 (continued). General assessment of current data quality compared to data available at source, by variable and country

	AU	BG	BR	CL	CO	DE	FR	HU	IN	JO	PL	PT	RU	SA	UK	US
law_text	+	+	+/-	+	+	+/-	+	-	-	+	+/-	-	+	+	+/-	+/-
stages_count	-	-	-	+	+/-	+/-	+/-	+/-	+	+/-	+	+	-	+	-	-
committee_count	-	+	-	-	-	+	-	+	-	+/-	+	+/-	-	-	-	+/-
final_vote_for	-	+	-	-	+/-	-	-	+	-	+/-	+	+/-	-	-	-	+/-
final_vote_against	-	+	-	-	+/-	-	-	+	-	+/-	+	+/-	-	-	-	+/-
final_vote_abst	-	+	-	-	+/-	-	-	+	-	+/-	+	+/-	-	-	-	+/-
Average_annual_nr_of_modifications	+/-	-	+/-	+	+	-	-	-	-	+/-	+	-	-	-	-	-
standard_law	+	+	+/-	+	+	+/-	+	-	+/-	-	+/-	-	-	+/-	+/-	+/-
omnibus_indicator	+	+	+/-	+	+	+/-	+	-	+/-	+/-	+/-	-	+	+/-	+/-	+/-
fast_tracked_procedure_indicator	+/-	+	+/-	+	+	+	+	+	+	+/-	+	+/-	+	+/-	+	+/-
initiator_indicator	+	+/-	+/-	+/-	+	+/-	+	+	+	-	+	+/-	+	+	+	+/-
modifications_indicator	+/-	-	+/-	+	+	-	-	-	-	-	+	-	-	-	-	-

Note: Symbols in the table can be interpreted as follows: “+”: an approximate accuracy level of 95%-100% given the available values on the website for the bills in our sample; “+/-”: an approximate accuracy level of 70%-95% and/or that variables should be used with caution; “-”: an approximate accuracy level below 70% and/or that variables should only be used with great caution. These assessments refer to the current version of the data (LEGDAT 1.1).

Supplemental Appendix A2: Cross-Country Descriptive Statistics

To illustrate variation across the countries in LEGDAT, we present radar charts with summary statistics for 13 of the main variables in Figure A2. In panels 1-3, we show country-specific medians for the bill-specific variables stages_count, committee_count, origin_type, bill_status, and bill_size. Panels 3-6 show cross-country variation in the law-specific variables law_size, standard_law, time_in_parl_days, and annual_average_number_of_modifications. Panels 7 and 8 show the proportions of laws in each country that are classified as having been passed using different forms of irregular legislative procedural discretion (omnibus_indicator, fast_tracked_procedure_indicator, and initiator_indicator)[2] or are subject to irregular lawmaking processes regarding the number of modifications (modifications_indicator). Each panel shows the value for the summary statistic across all countries at 0 degrees.

[2] As only legislators may introduce bills in the United States, the variable initiator_indicator is constant in this case.

FIGURE A2. Visualizations of Descriptive Statistics by Country and Across All Countries