British Journal of Political Science (2022), **52**, 1398–1408
doi:10.1017/S0007123420000897

British Journal of Political Science

LETTER

The Comparative Legislators Database

Sascha Göbel[1]* and Simon Munzert[2]

[1]Faculty of Social Sciences, Goethe University Frankfurt am Main, Germany; and [2]Data Science Lab, Hertie School, Berlin, Germany
*Corresponding author. E-mail: sascha.goebel@soz.uni-frankfurt.de

(Received 7 June 2020; revised 12 November 2020; accepted 2 December 2020; first published online 23 February 2021)

Abstract
Knowledge about political representatives' behavior is crucial for a deeper understanding of politics and policy-making processes. Yet resources on legislative elites are scattered, often specialized, limited in scope or not always accessible. This article introduces the Comparative Legislators Database (CLD), which joins micro-data collection efforts on open-collaboration platforms and other sources, and integrates with renowned political science datasets. The CLD includes political, sociodemographic, career, online presence, public attention, and visual information for over 45,000 contemporary and historical politicians from ten countries. The authors provide a straightforward and open-source interface to the database through an R package, offering targeted, fast and analysis-ready access in formats familiar to social scientists and standardized across time and space. The data is verified against human-coded datasets, and its use for investigating legislator prominence and turnover is illustrated. The CLD contributes to a central hub for versatile information about legislators and their behavior, supporting individual-level comparative research over long periods.

Keywords: comparative; crowdsourcing; dataset; database; elites; legislators; parliament; politicians; political behavior; R; Wikidata; Wikipedia

Decades after the advent of the behavioral revolution in the study of politics, political representatives' behavior remains a key research subject in political science. Contemporary studies address topics such as political elites' strategic voting behavior (Brown and Goodliffe 2017), the microfoundations of elite decision making (Linde and Vis 2017), local interest representation by elites and its consequences (Rogers 2017), and elites' responsiveness to news and social media (Barberá et al. 2019). Systematic information about political elites is therefore frequently sought in the discipline (see Appendix Figure A1).

Despite this prolonged demand, there is a shortage of large-scale, cross-national and longitudinal sources of such information (Gerring et al. 2019). Consequences of this paucity include recurring and redundant data collection efforts, analyses limited to narrow time frames and single countries, varying data quality and evidence foundations, and adverse conditions for replication. Country- or topic-specific datasets often exist in isolation. Recently, scholars have thus called for more systematic datasets that bridge field-specific gaps in legislator data (Krcmaric, Nelson and Roberts 2020).

We approach this problem by unifying heterogeneous and collaborative micro-data collection efforts. A large volume of user-generated information about legislators is available on the web. We bring these sources together using the open-collaboration platforms Wikipedia and Wikidata to help overcome the limitations of previous projects and create a comprehensive data infrastructure. The Comparative Legislators Database (CLD) is a one-stop shop for rich, diverse and integrated individual-level data on national political representatives. Our focus on the national level is motivated by scholars' interest in (see Appendix Section A) and the extensive availability of crowdsourced information about this group of legislators.

The CLD currently covers 45,540 contemporary and historical politicians elected to 338 legislative sessions in ten national legislatures: the Austrian Nationalrat, the Canadian House of Commons, the Czech Poslanecká Sněmovna, the French Assemblée, the German Bundestag, the Irish Dáil, the Scottish Parliament, the Spanish Congreso de los Diputados, the United Kingdom House of Commons and the United States Congress. Features are recorded in linked content-specific datasets and include political, sociodemographic, career, online presence, public attention and visual information. The database gains substantive coverage through an integration with renowned political science datasets, including information such as votes, speeches and legislation. This bridges the unconnectedness of datasets, thus strengthening the consistency and diversity of data across countries. In addition, the architecture of our database makes it easy to connect other existing and future datasets, which will make it possible to further develop, step by step, a centralized data pool. We provide free and fast open-source software with targeted access to the CLD. We illustrate the utility of the database with two example applications including the study of dynamics of public attention to political representatives and a comparative analysis of legislative turnover over several decades. The conclusion discusses further potential areas of application.

Existing Projects and Demand For Data

Considerable efforts have been made to assemble information about elected officials. Data projects have emerged especially for US legislatures, which are frequently the subject of political science research (for example, Bonica 2016; CQ Press 2018; Inter-university Consortium for Political and Social Research and McKibbin 1997; Vote Smart 2018). Although less extensive, similar efforts have traveled beyond US borders, for instance for the UK House of Commons (Eggers and Spirling 2014) or the German Bundestag (Sieberer et al. 2020). While geographically confined and often equipped with a substantive and temporal focus, these single-country projects continue to assist research on political representatives.

In recent years, data collection efforts surrounding legislative elites have gained pace. Several projects have taken on the ambitious task of assembling information in a cross-national and longitudinal fashion (Azavea 2018; Bailer et al. 2018; Gerring et al. 2019; MySociety 2018; Wagner et al. 2017). The Parliamentary Careers in Comparison (Bailer et al. 2018) project, for example, collects fine-grained biographical and career-related data on parliamentarians in three European democracies since the Second World War. In an unprecedented effort, the Global Leadership Project (Gerring et al. 2019) relies on country experts to compile socio-demographic and political background information on over 38,000 politicians from 145 countries between 2010 and 2013.

Notwithstanding the contributions of these advances, they illustrate a pervasive data collection trade-off: an increase in spatial coverage comes at the cost of temporal scope and substantive depth, and vice versa. This compromise is rooted in gathering information mainly through primary data collection. Fielding expert surveys or employing human coders is expensive. Given limited project budgets, high development costs naturally require a trade-off and restrict a project's scope or substantive focus.

This trade-off has implications for the applicability of existing projects. Substantive problems and data requirements differ vastly. Projects with broad geographic coverage enable a comparative perspective, yet may lack detailed information required for more specific questions. Conversely, specialized projects may not be applicable to the countries at hand. Complementing data with the means of integrating into other projects or domain-specific datasets can go a long way. However, existing data projects usually do not provide such data linkage.

Finally, the increased costs of data collection have consequences for the public availability and lifespan of data projects. Resource-intensive approaches generate incentives to put data under

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press

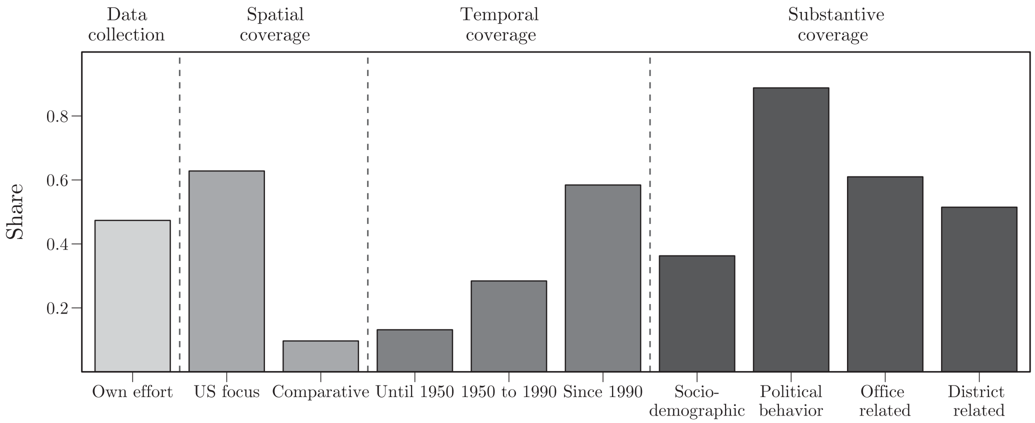


Figure 1. Data on national legislators in 209 articles from five political science journals

embargo and to retain the first-user publication advantage. Limited financial-planning horizons also make it difficult to keep projects alive in the long run.

To provide a picture of the demand side for data on political representatives, we conducted a survey of articles published in five top political science journals between 2009 and 2018 (see Figure 1 and Appendix Section A). The results echo the landscape of existing projects and the consequences of the discussed trade-off.

Almost half of the surveyed studies engaged in original data collection (47 per cent). A majority of studies working with legislator data focus on the United States (63 per cent). Research comparing representatives from different countries is quite rare (10 per cent). The substantive demand is largely oriented at political behavior variables, such as roll-call votes or ideology estimates, reported in prominent single-country datasets. Studies usually also cover rather short time windows (median = 10 years) and focus primarily on recent sessions. We suspect a regional bias in the availability of single-country data, and restrictions in the applicability and accessibility of existing cross-national and longitudinal projects as the main causes of these patterns.

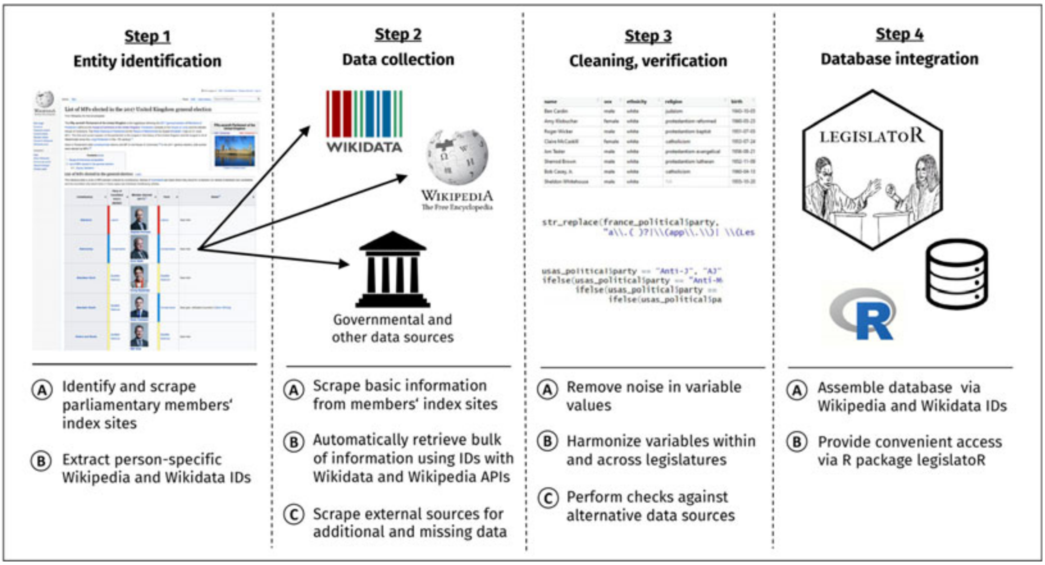
We respond to these findings by shifting the task from primary data collection to bringing together collaborative micro-data collection efforts and integrating existing projects. To achieve these objectives, we rely primarily on the open-collaboration platforms Wikipedia and Wikidata. These platforms provide a steady and mostly standardized flow of information. Our approach requires minimal employment of manual labor and financial resources, allowing the CLD to grow in all directions and making it sustainable.

Data Collection

Figure 2 illustrates our data collection and processing workflow, broken down into four steps. Step 1 (entity identification) uses Wikipedia as a starting point for three reasons. First, Wikipedia is one of the largest data sources on politicians in the world. Secondly, the micro-data collection efforts of the many volunteers on this platform allow us to crowdsource the primary data collection part. Thirdly, Wikipedia provides information on elected officials in a largely consistent structure across countries and over time. Legislators are commonly listed on parliament- and session-specific pages.[1]

[1] These pages are identified and scraped as ‘index sites’ to gather links to legislators’ articles, which we use to extract person-specific Wikipedia and Wikidata IDs. For instance, https://en.wikipedia.org/wiki/List_of_MPs_elected_in_the_2017_

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press


Step 1
Entity identification
Ⓐ Identify and scrape parliamentary members' index sites
Ⓑ Extract person-specific Wikipedia and Wikidata IDs
Step 2
Data collection
WIKIDATA
WIKIPEDIA The Free Encyclopedia
Governmental and other data sources
Ⓐ Scrape basic information from members' index sites
Ⓑ Automatically retrieve bulk of information using IDs with Wikidata and Wikipedia APIs
Ⓒ Scrape external sources for additional and missing data
Step 3
Cleaning, verification
Ⓐ Remove noise in variable values
Ⓑ Harmonize variables within and across legislatures
Ⓒ Perform checks against alternative data sources
Step 4
Database integration
LEGISLATOR
R
Ⓐ Assemble database via Wikipedia and Wikidata IDs
Ⓑ Provide convenient access via R package legislatoR

```
str_replace(france_political$party,
              "a\\.( )?|\\(app\\.\\)| \\(Les
```

```
usas_political$party == "Anti-J", "AJ"
ifelse(usas_political$party == "Anti-M
       ifelse(usas_political$party ==
              ifelse(usas_political$pa
```

Figure 2. Data collection and processing workflow

In Step 2 (data collection), we scraped basic information from the index sites and used the gained IDs to query Wikipedia and Wikidata's Application Programing Interfaces (APIs) for various legislator features and external identifiers. Based partly on this information, we located and consulted other databases, such as parliamentary websites, to reduce the amount of missing information.

Data were cleaned, harmonized and verified in Step 3. Much of the data came from Wikidata, which is pre-structured, so little cleaning was necessary. Variables were harmonized across legislators to create consistent value schemes. To supervise data quality, checks against alternative data sources were performed for sampled entries.

In Step 4 (database integration), the legislature- and content-specific datasets were integrated into a global database. A consistent architecture across all tables was established and made openly accessible via the legislatoR R package. More details on individual steps are reported in Appendix Section B.

Content, Architecture and Access

The CLD consists of nine tables for each country (see Figure 3). A table with socio-demographic data (*Core*) sits at the center. It includes a legislator's name, sex, date of birth and death, ethnicity, religious affiliation, and geographic coordinates for place of birth and death. Political information is stored in a second table (*Political*). It provides the legislative period a representative was elected to, when that session started and ended, party affiliation, constituency, duration in office, government membership and leader indicators for each session.

Two more tables deliver information about legislators' public offices (*Offices*) and professions (*Professions*). Both contain descriptions with Boolean values indicating membership. In accordance with growing interest in elite behavior on the internet, another table includes social media handles and URLs of personal websites (*Social*). To support image-based analyses, the CLD also provides a table that stores URLs of legislators' portraits (*Portraits*).

For every legislator's article, Wikipedia documents the process of information generation and its consumption. This metadata is stored in two additional tables. Revision records are reported

United_Kingdom_general_election lists all MPs that were members of the 57th UK House of Commons and links them to their own entry on Wikipedia (for example, https://en.wikipedia.org/wiki/Stephen_Kinnock).

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press

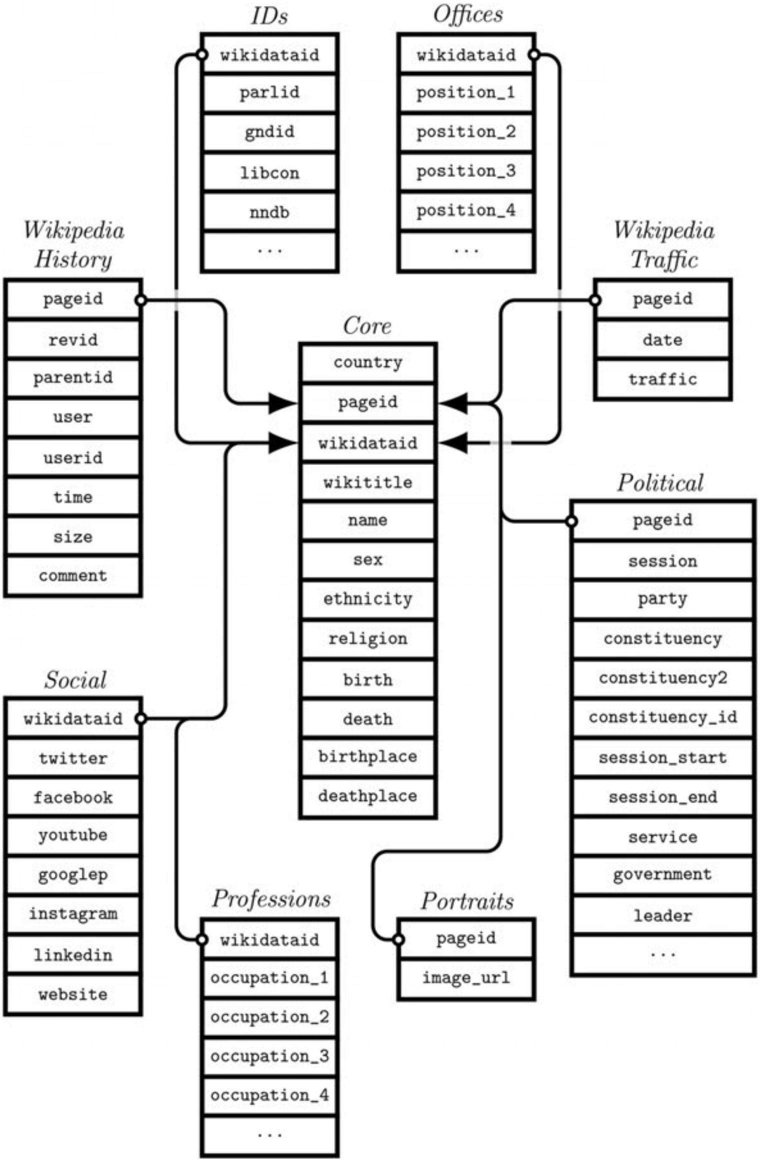


Figure 3. Structure of the database

with an identifier that locates information directly on Wikipedia (*Wikipedia History*). User names, (internet protocol addresses) IDs of (non-registered) editors, the date, time, size and comments of revisions are stored too. Long-term records of the daily views received by each page are reported since 2007 (*Wikipedia Traffic*).

A central feature of the CLD is its integration with other projects. A final table includes several individual identifiers, for instance, for linkage to official parliamentary websites or political science datasets (*IDs*). We have integrated the CLD with bills, votes and ideology estimates in the US Congress (Adler and Wilkerson 2018; Lewis et al. 2019), roll-call votes from the German and UK legislatures (Eggers and Spirling 2014; Sieberer et al. 2020), and plenary speeches held in Austria, the Czech Republic, Germany, Ireland, Spain and the UK (Herzog and Mikhaylov 2017; Rauh, De

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press

Wilde and Schwalbach 2017). This integration offers the means to discover new avenues for research and perspectives on long-standing questions.

All tables come in a format familiar to social scientists: every row represents a politician, every column a variable. This makes it easy to extract, filter, transform, sort, aggregate and visualize information. To promote usability, tables are arranged in a relational fashion. Two legislator-specific keys, the Wikipedia page and the Wikidata ID, link all tables to the *Core*. This offers an intuitive organization that facilitates targeted access.

The CLD is hosted online and access is managed via R, an open-source software environment for statistical computing. We opted for R because it is one of the most popular data analysis programs used in the social sciences. We programmed an open R package that implements fast, targeted and memory-efficient access without having to download the full database. Appendix Section F contains instructions for installing and working with the package.

Application 1: Tracking Public Attention Paid to Legislators with Wikipedia Page Views

We present two applications that highlight the three core strengths of the CLD in comparison with existing projects. First, it facilitates the exploration of new and partly untapped data, such as metadata from Wikipedia. Secondly, it makes it easy to conduct comparative research on legislative elites. Thirdly, it helps evaluate long-term trends in elite behavior.[2]

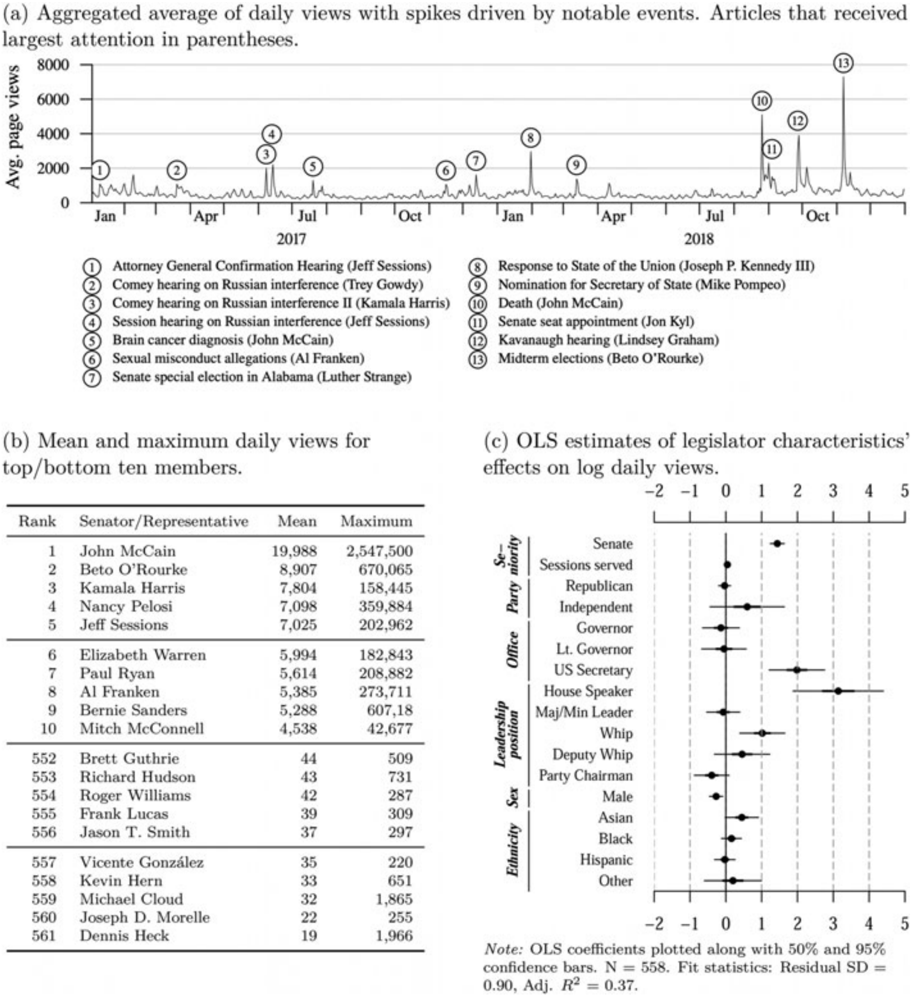
A central motivation to study political elites is their key role in shaping public discourse. However, politicians differ vastly in their factual power and the attention they attract from the public and the media. Scholars have quantified these traits using newspaper data, for instance (Ban et al. 2019). Alternatively, we can use Wikipedia page traffic – that is, how often an entry has been accessed on a given day by ordinary users of the encyclopedia (Munzert 2018). More prominent politicians should receive more extensive attention. Figure 4 shows analyses of page traffic data for members of the 115th Congress between 2017 and 2018. Even in a sample of articles on high-profile political elites like these, there is considerable variation in attention. Panel 4b reports the ten top and bottom members by average daily views. Top-ranking politicians, with high name recognition and media presence, receive thousands of daily page views on average, whereas rank-and-file members get no more than a few dozen. This corroborates the face validity of the measure for public attention.

In addition to baseline differences, Panel 4a shows how offline events correspond to spikes in the time series of aggregated average daily views. While some events are directly linked to the legislator covered in the article (such as news on John McCain's brain cancer diagnosis and his death about one year later), others demonstrate how legislator activity can raise awareness in the public. Both Trey Gowdy and Kamala Harris's public attention benefited from their performance in the Comey hearings, and Joe Kennedy III attracted nationwide interest when he offered his response to the State of the Union in January 2018.

The ordinary least squares model in Panel 4c investigates whether attention to legislative elites in terms of Wikipedia article views is largely idiosyncratic and driven by exogenous shocks, or whether systematic factors have any predictive value. On average, senators are substantively more prominent than house members (corresponding to 370 per cent higher attention in terms of article views), and members with US secretary appointments receive more attention than former state officials (680 per cent). However, while prominent leadership positions, such as Speaker of the House, are also reflected in page views (2,480 per cent), the ranking of party leader positions does not translate into received attention. More attention is also paid to female

[2] In Appendix Section E we present an additional comparative application using the Wikipedia data that is linked in the CLD to track the substantive representation of women in parliaments.

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press



(a) Aggregated average of daily views with spikes driven by notable events. Articles that received largest attention in parentheses.

(b) Mean and maximum daily views for top/bottom ten members.

Rank	Senator/Representative	Mean	Maximum
1	John McCain	19,988	2,547,500
2	Beto O'Rourke	8,907	670,065
3	Kamala Harris	7,804	158,445
4	Nancy Pelosi	7,098	359,884
5	Jeff Sessions	7,025	202,962
6	Elizabeth Warren	5,994	182,843
7	Paul Ryan	5,614	208,882
8	Al Franken	5,385	273,711
9	Bernie Sanders	5,288	607,18
10	Mitch McConnell	4,538	42,677
552	Brett Guthrie	44	509
553	Richard Hudson	43	731
554	Roger Williams	42	287
555	Frank Lucas	39	309
556	Jason T. Smith	37	297
557	Vicente González	35	220
558	Kevin Hern	33	651
559	Michael Cloud	32	1,865
560	Joseph D. Morelle	22	255
561	Dennis Heck	19	1,966

(c) OLS estimates of legislator characteristics' effects on log daily views.

Note: OLS coefficients plotted along with 50% and 95% confidence bars. N = 558. Fit statistics: Residual SD = 0.90, Adj. R^2 = 0.37.

Figure 4. Descriptive statistics and predictive model of Wikipedia page views of members of the 115th US Congress

legislators (24 per cent) and, in part, ethnic minorities (Asian 60 per cent). Overall, the structural factors predict a modest portion of the variance in logged daily views (adj. $R^2 = 0.37$).

Appendix Figures D1 to D8 provide replications of the page view analyses for the other legislatures represented in the CLD. Across all settings, spikes in aggregated viewership can be linked to significant events, the rankings of politicians by average page views have high face validity, and legislator characteristics predict a fair share of the variance in logged daily views. While more research is needed to evaluate the use of Wikipedia viewership metadata for analyzing political elites, we have shown that the data provides meaningful signals on the relative public attention legislators receive. Furthermore, the fact that the measure provides a metric that is comparable across contexts is beneficial for comparative research.

Application 2: Investigating Legislative Alternation and Renewal

Legislative turnover, the share of new (as opposed to re-elected) legislators, is an important measure of elite circulation. It serves as a diagnostic tool for assessing the representativeness and functioning of parliaments (Putnam 1976). Turnover can be categorized as either alternation or

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press

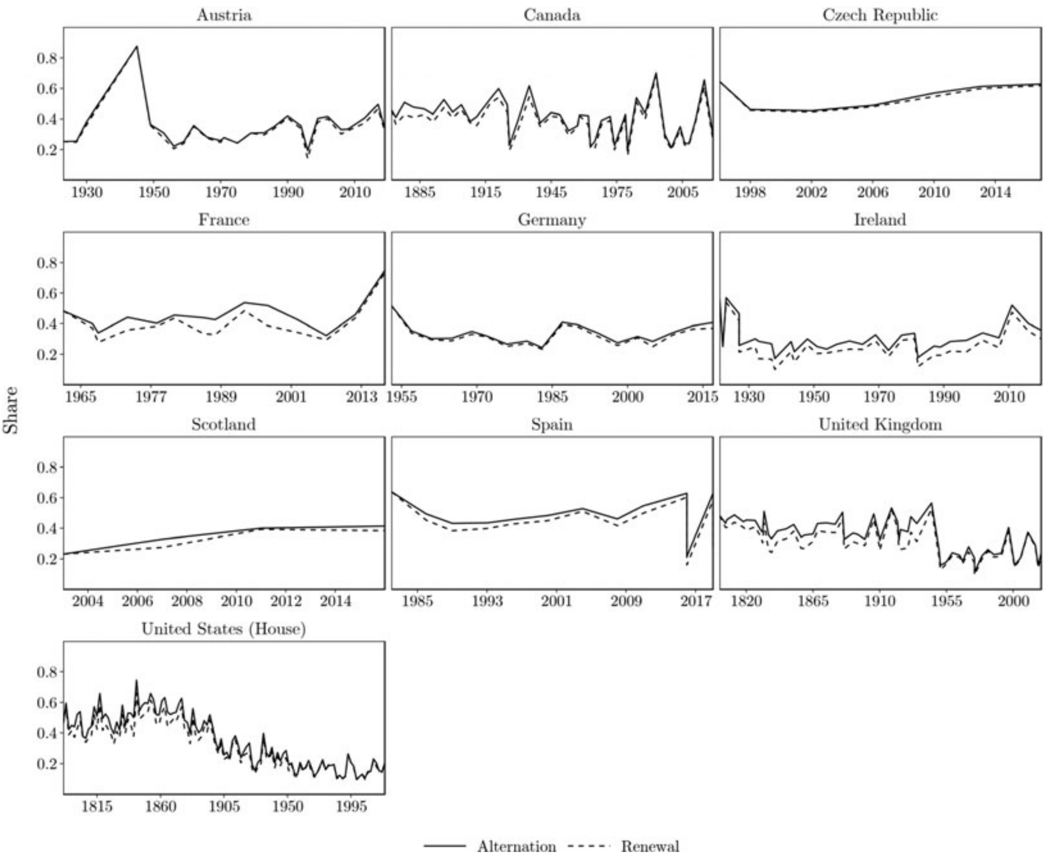


Figure 5. Legislative turnover in ten legislatures

renewal (François and Grossman 2015). Alternation is an aggregate measure of newly incoming legislators that lumps together politicians who are entering parliament for the first time and past members who, after having lost their mandate at some point, are returning to parliament. Renewal refers to only new, first-term legislators. Prior comparative research focuses on alternation and is restricted to a few decades or, at best, the post-World War II period (Gouglas, Maddens and Brans 2018; Matland and Studlar 2004).

We use the CLD to trace both alternation and renewal during the entire history of ten legislatures (yielding more than double the number of sessions previously investigated). The solid line in Figure 5 represents the results for legislative alternation. For Germany, Ireland, Spain, the UK and the United States, we find averages of alternation that are 5 to 20 per cent larger than previously reported (Matland and Studlar 2004). This may be partly due to the prior omission of parliamentary foundation periods, which are characterized by greater instability and shortsightedness. For instance, the UK is commonly viewed as a consistently low-turnover country with an upward trend (Gouglas, Maddens and Brans 2018; Matland and Studlar 2004). Yet, we find turnover fluctuations of around 40 to 50 per cent. A large drop is only apparent after the Second World War, when it moved to around 20 per cent and has not displayed a clear direction since.

Regarding renewal, we corroborate prior findings showing a substantive number of returning politicians among incoming legislators in France (François and Grossman 2015). Measuring only alternation can indeed hide potential vulnerabilities in the electoral connection, due to the retention of power and the manifestation of a ruling class. However, we observe similar patterns only

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press

for Ireland and Spain. For other countries, alternation is either congruent with renewal or has become so during the last few decades. Especially for the UK and the US, a drop in alternation rates is slightly compensated by the absence of returning politicians. That said, turnover reached alarmingly low levels, especially in the United States.

The applications illustrate the CLD's potential to examine diverse questions. Future research can build on these analyses and use the CLD to gain further insight into the historical development of individual career choices or how individuals achieve prominence on their way from being a freshman to a figure of importance in national politics.

Discussion

As with any large-scale data collection effort, the CLD comes with limitations and potential biases. In this section, we put them into perspective and point to future work on the database.

Coverage Bias

The collaborative micro-data collection efforts on Wikipedia and Wikidata are limited by the population of people contributing to these projects. In the CLD, this is reflected in an over-representation of legislatures from the Western hemisphere. We aim to address this issue by covering more legislatures over time. As full coverage of legislators was a key consideration in our choice of countries (see Appendix Section C for a comparison to existing, comprehensive lists of legislators), non-random missingness of elites within legislatures is currently not an issue. A more serious concern is item non-response – that is, non-randomly missing information on features. We checked missingness on various variables by legislatures (see Appendix Figure C1) and by core demographics (see Appendix Table C1). Religion, ethnicity and portrait data are generally sparsely populated, and recently active legislators are better covered than those from early sessions (recency bias). However, we do not find evidence of ethnic minority or gender biases (if anything, more missings occur for men, which is likely confounded with recency bias and women entering parliaments rather late).

Data Quality

We argue that the collaborative nature of Wikipedia and Wikidata works against measurement errors due to flawed variable codings. While anybody can edit information on these platforms at any time, Appendix Figure C2 shows that content creation on legislators' Wikipedia pages balances editor experience and collective intelligence. To check whether this control mechanism works as expected, we compared the CLD with hand-coded information from other projects. Appendix Figure C3 shows an almost 100 per cent agreement for every variable we checked. In the few cases of disagreement, the CLD was either on par with other projects or yielded the correct information more often. We additionally conducted a Google search for official information on 500 randomly drawn legislators and confirmed most of the information in the CLD (see Appendix Figure C4). Where we could not confirm information, this was mostly because information was not traceable. Only in rare cases could we actually disconfirm information in the CLD. In any case, the CLD is updated frequently, so its quality and exhaustiveness should further increase over time.

Breadth of Coverage

The CLD does not claim to be complete in terms of features. While it is strong on both breadth and depth, the focus is on providing a consistent set of features over all covered legislatures and allowing extensibility by linking to other, more specialized data sources.

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press

Conclusions

Learning about political representatives requires a diverse array of information. This includes behavioral data and metadata, as much as static characteristics and historical information. The CLD provides a central hub for these data and a starting point for students of legislative elites.

On a practical note, the CLD affords a first overview of potentially important data. This is followed by the swift and easy access to and integration of a variety of different types of data. Together, this helps researchers manage the focus of their project in two ways. First, it steers data collection and management efforts to where they are actually required. Secondly, it facilitates quick exploratory analyses to discover and sound out questions.

The CLD can enable substantive contributions to several topics. For instance, the *Wikipedia Traffic* table, combined with integrated roll-call or speech data, may be used to study the consequences of legislative behavior by analyzing whether it predicts shifts in public attention. Similarly, the *Portraits* table could be deployed in studies that investigate the behavioral consequences of facial features. In addition, the *Professions* table can serve as the basis for categorizing politicians in blue- and white-collar backgrounds. Combined with social media accounts in the *Social* table or integrated speech data, this facilitates the study of class-based differences in representatives' communication. Finally, revision records of legislators' personal Wikipedia biographies, in the *Wikipedia History* table, offer an opportunity to analyze the dissemination, targets and potential sources of political disinformation on Wikipedia. As the CLD grows in the years to come, its practical and substantive potential will expand further.

Supplementary material. Online appendices are available at https://doi.org/10.1017/S0007123420000897.

Acknowledgements. We would like to thank Lada Rudnitckaia, Johana Sperlova, Angela Peckham, Lorenzo Andreoli, Maximilian Bayer, and Toma Pavlov for excellent research assistance. We would also like to thank Christian Breunig, Benjamin Guinaudeau, Peter Selb, participants of the Center for Data and Methods Colloquium at the University of Konstanz, participants of the Conference on Comparative Research on Political Elites at the Hertie School, and the anonymous reviewers as well as the editor for their helpful feedback.

Data availability. Data replication sets are available in Harvard Dataverse at: https://doi.org/10.7910/DVN/GYSEGP.

Financial support. This research was supported by a grant of the Excellence Initiative of the German federal and state governments (GSC 1019) and by the Daimler and Benz Foundation (Funding period 2017/18; project 'Citizen and Elite Activity on the Wikipedia Market Place of Political Information').

References

Adler ES and Wilkerson J (2018) Congressional bills project. NSF 00880066 and 00880061.

Azavea (2018) Cicero. Available from https://www.cicerodata.com.

Bailer S et al. (2018) Parliamentary Careers in Comparison. Available from http://parliamentarycareersincomparison.org.

Ban P et al. (2019) How newspapers reveal political power. *Political Science Research and Methods* **7**(4), 661–678.

Barberá P et al. (2019) Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review* **113**(4), 883–901.

Bonica A (2016) A data-driven voter guide for U.S. elections. Adapting quantitative measures of the preferences and priorities of political elites to help voters learn about candidates. *Russell Sage Foundation Journal of the Social Sciences* **2**(7), 11–32.

Brown AR and Goodliffe J (2017) Why do legislators skip votes? Position taking versus policy influence. *Political Behavior* **39**(2), 425–455.

CQ Press (2018) *Congress Collection*. Thousand Oaks, CA: Sage.

Eggers A and Spirling A (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.

François A and Grossman E (2015) How to define legislative turnover? The incidence of measures of renewal and levels of analysis. *Journal of Legislative Studies* **21**(4), 457–475.

Gerring J et al. (2019) Who rules the world? A portrait of the global leadership class. *Perspectives on Politics* **17**(4), 1079–1097.

Göbel S and Munzert S (2020) Replication data for: The comparative legislators database. https://doi.org/10.7910/DVN/GYSEGP, Harvard Dataverse, V1. 6:oJ0ylHRq2POvWyRdNJDogA==.

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press

Gouglas A, Maddens B and Brans M (2018) Determinants of legislative turnover in Western Europe, 1945–2015. *European Journal of Political Research* **57**(3), 637–661.

Herzog A and Mikhaylov SJ (2017) Database of Parliamentary Speeches in Ireland, 1919–2013. IEEE Proceedings of the 2017 International Conference on the Frontiers and Advances in Data Science, pp. 29–34.

Inter-university Consortium for Political and Social Research and McKibbin C (1997) *Roster of United States Congressional Officeholders and Biographical Characteristics of Members of the United States Congress, 1789–1996. Merged Data*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Krcmaric D, Nelson SC and Roberts A (2020) Studying leaders and elites. The personal biography approach. *Annual Review of Political Science* **23**, 133–151.

Lewis JB et al. (2019) *Voteview. Congressional roll-call votes database*. Available from https://voteview.com/data.

Linde J and Vis B (2017) Do politicians take risks like the rest of us? An experimental test of prospect theory under MPs. *Political Psychology* **38**(1), 101–117.

Matland RE and Studlar DT (2004) Determinants of legislative turnover. A cross-national analysis. *British Journal of Political Science* **34**(1), 87–108.

Munzert S (2018) Measuring the importance of political elites. Available from https://osf.io/preprints/socarxiv/t8gs5.

MySociety (2018) EveryPolitician. Available from https://everypolitician.org.

Putnam RD (1976) *The Comparative Study of Political Elites*. Englewood Cliffs, NJ: Prentice-Hall.

Rauh C, De Wilde P and Schwalbach J (2017) The ParlSpeech data set. Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states. Harvard Dataverse V1. Available from https://doi.org/10.7910/DVN/E4RSP9.

Rogers S (2017) Electoral accountability for state legislative roll calls and ideological representation. *American Political Science Review* **111**(3), 555–571.

Sieberer U et al. (2020) Roll-call votes in the German Bundestag. A new dataset, 1949–2013. *British Journal of Political Science* **50**(3), 1137–1145.

Vote Smart (2018) Project Vote Smart. Facts matter. Available from https://votesmart.org.

Wagner C et al. (2017) *Politicians on Wikipedia and DBpedia*. Köln: GESIS.

Cite this article: Göbel S, Munzert S (2022). The Comparative Legislators Database. *British Journal of Political Science* **52**, 1398–1408. https://doi.org/10.1017/S0007123420000897

https://doi.org/10.1017/S0007123420000897 Published online by Cambridge University Press