The Comparative Legislators Database*

Sascha Göbel, University of Konstanz[†]
Simon Munzert, Hertie School[‡]

This Version: June 5, 2020

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. We introduce 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. We 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. We verify the data against human-coded datasets and illustrate its use for investigating legislator prominence and turnover. 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, Crowd-Sourcing, Dataset, Database, Elites, Legislators, Parliament, Politicians, Political Behavior, R, Wikidata, Wikipedia

Word count. 3,998

*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"). 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, and participants of the Conference on Comparative Research on Political Elites at the Hertie School for their helpful feedback.

[†]Ph.D. candidate in Political Science, Graduate School of Decision Sciences, University of Konstanz. Address: Universitätsstraße 10, 78464 Konstanz, Germany. Email: sascha.goebel@uni-konstanz.de.

[‡]Assistant Professor of Data Science and Public Policy, Hertie School. Address: Friedrichstr. 180, 10117 Berlin, Germany. Email: munzert@hertie-school.org.

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

Despite this prolonged demand, there is a shortage of large-scale, cross-national, and longitudinal sources for such information (Gerring et al., 2019). Consequences of this paucity are: 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.

We approach this problem by unifying heterogeneous and collaborative micro-data collection efforts. User-generated information about legislators is available in plenty on the Web. By bringing these sources together, especially through the open-collaboration platforms Wikipedia and Wikidata, we overcome limitations of existing projects and create a comprehensive data infrastructure. The Comparative Legislators Database (CLD) serves as 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 scholar's interest in (see Section A in the SI) 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, or legislation. 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 towards political representatives and a comparative analysis of legislative turnover over several decades. Further potential areas of application are discussed in the conclusion.

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 that are frequently the subject of political science research (e.g. [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 boundaries, for instance for the UK House of Commons ([Eggers and Spirling, 2014](#)) or the German Bundestag ([Sieberer et al., Forthcoming](#)). 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 ([Bailer et al., 2018](#); [Wagner et al., 2017](#); [Azavea, 2018](#); [MySociety, 2018](#); [Gerring et al., 2019](#)). 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 the sociodemographic and political background for over 38,000 politicians from 145 countries between 2010 and 2013.

Notwithstanding the contribution 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 means high costs. In light of limited budgets, high development costs naturally require a trade-off and restrict a project’s scope or substantive focus.

This trade-off has implications regarding 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, increased costs of data collection have consequences for the public availability and lifespan of data projects. Resource-intensive approaches yield incentives to put data under embargo and to retain the first user publication advantage. In addition, limited financial-planning horizons 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 Section A in the SI). The results echo the landscape of existing projects and the consequences of the discussed trade-off.

Despite a considerable number of existing data projects, almost half of the surveyed studies engaged in original data collection (47 percent). A majority of studies working with legislator data focus on the US (63 percent). By contrast, research comparing representatives from different countries is quite rare (10 percent). 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 for these patterns.

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

We respond to this by shifting the task from primary data collection to bringing together collaborative micro-data collection efforts and integrating existing projects. For this 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. In Step 1: entity identification, Wikipedia is used as a starting point. First, this is because Wikipedia represents one of the largest data sources on politicians worldwide. Second, the micro-data collection efforts of the many volunteers on this platform allow us to crowd-source the primary data collection part. Third, Wikipedia provides information on elected officials in a largely consistent structure across countries and over time. Commonly, legislators are listed on parliament- and session-specific pages. These pages are identified and scraped as “index sites” to gather links to legislators’ articles, which we use to extract person-specific Wikipedia and

Figure 2: Data collection and processing workflow.

Wikidata IDs. For instance, https://en.wikipedia.org/wiki/List_of_MPs_elected_in_the_2017_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 (e.g. https://en.wikipedia.org/wiki/Stephen_Kinnock).

In Step 2: data collection, we scrape basic information from the index sites and use the gained IDs to query Wikipedia and Wikidata’s Application Programming Interfaces (APIs) for various legislator features and external identifiers. In part, based on this information, we locate and consult other databases, such as parliamentary websites, to reduce the amount of missing information.

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

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

Content, Architecture, and Access

The CLD consists of nine tables for each country (see Figure 3). A table with sociodemographic data (*Core*) sits at the center. It includes a legislators’ 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 hold 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 both the process of information generation and its consumption. This metadata is stored in two additional tables. Revision records are reported 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., Forthcoming](#)), and plenary speeches held in Austria, the Czech Republic, Germany, Ireland, Spain, and the UK ([Herzog and Mikhaylov, 2017](#); [Rauh,](#)

Figure 3: Structure of the database.

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. Instructions for installing and working with the package are offered in section D in the SI.

Application 1: Capturing Legislators' Prominence with Wikipedia Page Views

We present two applications that highlight the three core strengths of the CLD in comparison with existing projects: First, it allows an exploration of new and partly untapped data, such as metadata from Wikipedia. Second, it makes it easy to conduct comparative research on legislative elites. Third, it allows studying long-term trends in elite behavior.

A central motivation to study political elites is their key role in shaping public discourse. However, politicians differ vastly in their factual power and perceived prominence by the public. Scholars have quantified these traits using, for instance, newspaper data (Ban et al., 2019). Alternatively, we can make use of Wikipedia page traffic, i.e. how often an entry has been accessed on a given day by ordinary users of the encyclopedia (Munzert, 2019).

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 a large 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

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

(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. $R^2 = 0.37$.

members get no more than a few dozen. This corroborates the face validity of the measure for public prominence.

In addition to baseline differences, Panel 4a shows how offline events correspond with 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 giving his response to the State of the Union in January 2018.

The OLS model in Panel 4c investigates whether attention towards 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% higher attention in terms of article views), and members with US secretary appointments receive more attention than former state officials (680%). However, while prominent leadership positions, such as Speaker of the House, are also reflected in page views (2,480%), the ranking of party leader positions does not show in received attention. More attention also goes to female legislators (24%) and, in part, ethnic minorities (Asian 60%). Overall, the structural factors predict a modest part of the variance in logged daily views (adj. $R^2 = 0.37$).

Application 2: Investigating Legislative Alternation and Renewal

Legislative turnover, the share of incoming legislators, is an important measure of elite circulation and a diagnostic tool for the assessment of the representativeness and functioning of parliaments (Putnam, 1976). Turnover can be distinguished into alternation, the overall share of incoming legislators, and renewal, the share of first-term members (François and Grossman, 2015). Existing 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 over the whole history of ten legislatures (yielding more than double the amount of sessions previously inves-

Figure 5: Legislative turnover in ten legislatures.

tigated). The solid line in Figure 5 represents results for legislative alternation. For Germany, Ireland, Spain, the UK, and the US, we find averages of alternation that are 5 to 20 percent larger than previously reported (Matland and Studlar, 2004). In part, this may be due to prior omission of parliamentary foundation periods, characterized by higher 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 percent. A large drop is only apparent after the Second World War, with turnover moving to around 20 percent without a clear direction since.

As regards renewal, we corroborate 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 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 an absence of returning politicians. That said, turnover reached alarmingly low levels, especially in the US.

The applications illustrate the potential of the CLD for studying diverse questions. Future research can build on these analyses and use the CLD to gain further insight about the historical development of individual career choices or on 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, the CLD comes with limitations and potential biases. In the following, 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, non-random missingness of elites within legislatures is currently not an issue. A more serious concern is item non-response, i.e. information on features missing not at random. We checked missingness on various variables by legislatures (see Figure C1 in the SI) and by core demographics (see Table C1 in the SI). Religion, ethnicity, and portrait data are generally sparsely populated and recently-active legislators are better covered than legislators from early sessions (recency bias). However, we do not

find evidence for ethnic minority or gender biases (if anything, more missings occur for men, which is likely confounded with recency bias and woman 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, Figure C2 in the SI shows that content creation on legislator’s 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. Figure C3 in the SI shows an almost 100% 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 Figure C4 in the SI). 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 being 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.

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 and integration of a variety of different data types. Together, this supports researchers in managing the focus of their project. For one, by steering data collection and management efforts where they are actually required. Second, by facilitating quick exploratory analyses to discover and sound out questions.

We envision several topics where the CLD can prove helpful in enabling substantive contributions. For instance, the *Wikipedia Traffic* table, combined with integrated roll call or speech data, may be used to study 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. For another example, 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 allows for 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, this practical and substantive potential will expand further.

References

- Adler, E Scott and John Wilkerson. 2018. “Congressional bills project.” *NSF* 880066(00880061):213–227.
- Azavea. 2018. *Cicero*.
URL: <https://www.cicerodata.com>
- Bailer, Stefanie, Philip Manow, Simon Hug, Tomas Turner-Zwinkels, Niels Goet, Oliver Huwyler and Elena Frech. 2018. *Parliamentary Careers in Comparison*.
URL: <http://parliamentarycareersincomparison.org/>
- Ban, Pamela, Alexander Fournaies, Andrew B. Hall and James M. Snyder. 2019. “How Newspapers Reveal Political Power.” *Political Science Research and Methods* 7(4):661–678.
- Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost and Joshua A. Tucker. 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, Adam. 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, Adam R. and Jay Goodliffe. 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, 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.

- François, Abel and Emiliano Grossman. 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, John, Erzen Oncel, Kevin Morrison and Daniel Pemstein. 2019. “Who rules the world? A portrait of the global leadership class.” *Perspectives on Politics* 17(4):1079–1097.
- Gouglas, Athanassios, Bart Maddens and Marleen Brans. 2018. “Determinants of legislative turnover in Western Europe, 1945–2015.” *European Journal of Political Research* 57(3):637–661.
- Herzog, Alexander and Slava J. Mikhaylov. 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 Carroll McKibbin. 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.
- Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin and Luke Sonnet. 2019. *Voteview. Congressional roll-call votes database*.
URL: <https://voteview.com/data>
- Linde, Jona and Barbara Vis. 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, Richard E. and Donley T. Studlar. 2004. “Determinants of legislative Turnover. A cross-national analysis.” *British Journal of Political Science* 34(1):87–108.
- Munzert, Simon. 2019. “Measuring the Importance of Political Elites.” Unpublished Manuscript. <https://osf.io/preprints/socarxiv/t8gs5/download>.

MySociety. 2018. *EveryPolitician*.

URL: <https://everypolitician.org/>

Putnam, Robert D. 1976. *The comparative study of political elites*. Englewood Cliffs, NJ: Prentice-Hall.

Rauh, Christian, Pieter De Wilde and Jan Schwalbach. 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.”.

URL: <https://doi.org/10.7910/DVN/E4RSP9>

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

Sieberer, Ulrich, Thomas Saalfeld, Tamaki Ohmura, Henning Bergmann and Stefanie Bailer. Forthcoming. “Roll-call votes in the German Bundestag. A new dataset, 1949–2013.” *British Journal of Political Science* .

Vote Smart. 2018. *Project Vote Smart. Facts matter*.

URL: <https://votesmart.org/>

Wagner, Claudia, Marcel Reif, Kandhasamy Rajasekaran, Tevriz Iusupova and Evgenii Vasilev. 2017. *Politicians on Wikipedia and DBpedia*. Köln: GESIS.

The Comparative Legislators Database

Online Supplementary Information

Contents

Appendix A Literature Review	2
Appendix B Details on Data Collection	3
Appendix C Data Coverage, Quality, and Verification	5
Appendix D Introduction to the R package	13
Appendix E Software Statement	17

Appendix A Literature Review

The survey was conducted with the support of three research assistants. We conducted a search in the archives of four flagship general-interest journals of political science, the *American Political Science Review*, the *American Journal of Political Science*, the *Journal of Politics*, and the *British Journal of Political Science*, as well as the official journal of the Legislative Studies Section of the American Political Science Association, *Legislative Studies Quarterly*. Using JSTOR’s search mask, we used the query “(MP OR legislator) AND data” for all fields, filtered for articles only as item type, and restricted the search to articles published in a period of ten years, between 2009 and 2018. For periods not covered by JSTOR, we referred to the search function for the respective archive hosted by the journal’s publisher. Overall, this search identified 535 articles. In the next step, we screened the abstract of every article to identify studies that potentially used individual-level data of members of national-level legislative bodies. This resulted in 225 articles. In the third step, three research assistants inspected the body of the articles and collected information on the data sources used by the studies as well as their type (whether the data was provided by governmental sources or non-governmental institutions, journalistic data sources, databases provided by other scholars, or whether the study involved an original data collection effort), countries and years covered, and the types of features that were used. Here, we distinguished between sociodemographics (e.g. age, gender, ethnicity), political behavior (e.g. roll call votes, ideology measures), office-related variables (e.g. tenure, committee memberships, leadership positions), and constituency-specific variables (e.g. election results of legislator’s district, district public opinion, district sociodemographics). 16 studies turned out to be irrelevant in the last coding step.

Figure A1: Usage of data on national political representatives over time.

Note: Based on a survey of articles published in five top political science journals between 2009 and 2018.

Appendix B Details on Data Collection

After identifying parliamentary members' index sites on Wikipedia, we downloaded the respective HTML files. Next, we created legislature-specific scraper functions with the XPath query language. These functions were used on the HTML files to automatically extract basic data (name, party, constituency, period of service, and session) on representatives and URLs to their personal Wikipedia pages. The URLs were then applied to gain access to the individual Wikipedia page and Wikidata IDs via Wikipedia's API. Using several custom-built API bindings, we finally fed the collected IDs to the APIs of Wikipedia, Wikimedia, and Wikidata to automatically extract all remaining data. For the Wikipedia and Wikimedia APIs, we simply set the parameters to include the desired information (traffic, revision histories, and portraits). For the Wikidata API, we tapped content-specific properties pre-specified by Wikidata (IDs, offices, professions, social media, and all remaining sociodemographic data).

Following the automated data extraction on Wikipedia and Wikidata, we identified other sources to fill in some gaps in the data. For several countries, we filled missing religious affiliation by gathering information from governmental databases or other Wikipedia entries and websites. We proceeded, as above, by downloading HTML or XML files and building custom scraper functions to automatically extract the required information. Missing periods of service and social media data were gathered the same way. Period of service data was, in part, also manually extracted by the research assistants from notes on the parliamentary members' index sites. Social media data was, in part, shared with us by colleagues. Missing ethnicity information was partially coded by applying facial recognition software (Face++ API) to collected portraits. If the estimated accuracy was below 95 percent, we recoded ethnicity manually. Government membership and leader positions were manually collected by the research assistants in the form of relevant Wikidata IDs from the index and election-specific Wikipedia sites.

IDs for political science datasets were not collected via Wikipedia or Wikidata. Instead, we and the research assistants wrote scripts to match these datasets to our database. A combination of variables, such as session, name, date of birth, and party affiliation was used to uniquely link politicians. For politicians that could not be matched this way, the research assistants tried to achieve a match via manual research.

The formatting and harmonization of variables within and across legislatures was conducted by applying regular expressions and by grouping information. For instance, religious affiliation was manually grouped into major religious groups and Christian denominations. Messy party affiliations were cleaned and aligned by removing and replacing character sequences using regular expressions.

Appendix C Data Coverage, Quality, and Verification

Figure C1: Percentage of missings for selected variables by country.

Note: Country, pageid, wikidataid, wikipitle, name, session, session start, session end, wikipedia traffic and wikipedia history are complete by definition or due to the way we approach data collection, only very few do not have a pageid, even fewer do not have a wikidataid. Variables on which missingness is not necessarily due to lack of data availability were omitted from the assessment of missings (e.g. death and deathplace, government membership, leadership positions, constituency affiliation, and social media profiles.)

Table C1: Missing data analysis for selected countries.

United States (House)				
	Birthplace	Ethnicity	Portrait	Religion
Female	1.5%	1.2%	1.2%	57.8%
Male	12.4%	28.7%	28.7%	89.8%
Black	1.4%	–	0%	52.1%
Hispanic	3.3%	–	0%	62.5%
Other	2.9%	–	0%	58.8%
White	8.6%	–	0%	86.4%
Recency	0.84	0.74	0.74	0.57
United Kingdom				
	Birthplace	Ethnicity	Portrait	Religion
Female	20%	29.1%	29.3%	93.5%
Male	70.6%	63.7%	63.5%	97.9%
Asian	10.4%	–	12.5%	58.3%
Black	9.1%	–	9.1%	95.5%
White	46.7%	–	0%	96.6%
Recency	0.73	0.89	0.89	0.74
France				
	Birthplace	Ethnicity	Portrait	Religion
Female	1.2%	46%	46%	99.0%
Male	2.9%	71.8%	71.8%	98.4%
Black	4.5%	–	0%	81.8%
Other	0%	–	0%	100%
White	0.3%	–	0%	96.5%
Recency	0.83	0.76	0.76	1.1

Note: For the discrete variables sex and ethnicity the group-specific share of missings with reference to the column variables is reported. The continuous variable recency measures the time difference in years between legislators' birthdate and the current time at the date of computation. Reported is the ratio of recency means for missing and observed data with reference to the column variables. A ratio smaller than 1 indicates recency bias.

Our main data source, Wikipedia, is among the ten most-visited websites worldwide (See <https://www.alexa.com/siteinfo/wikipedia.org>. Last accessed February 2020), which emphasizes its importance as a source of information. The open-collaboration platform is a long-standing online encyclopedia, where information can be created and revised by anyone at any time. Wikipedia asks its editors to maintain a neutral point of view when editing content and employs various bots for picking up and correcting biased content (See https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view. Last accessed February 2020). Indeed, studies confirm the accuracy of information offered on Wikipedia in comparison to traditional encyclopedias (Giles, 2005). While this encourages trust in the CLD, it does not provide explicit verification of the very specific information the CLD offers. Furthermore, several studies indicate that information on Wikipedia can fall prey to political biases (Göbel and Munzert, 2018; Kalla and Aronow, 2015). For this reason, we conducted several verification checks, detailed in the following paragraphs.

Recent research suggests that both the number of editors and the amount of editorial experience are related to the degree of biased information on the encyclopedia (Greenstein, Gu and Zhu, 2017). Contributors were found to target content that was especially slanted way from their viewpoint and enter into dialogue with each other during editing. This indicates heterogeneous conversations that drive a bias-reducing environment. Naturally, the more participants involved in editing an article, the more this collective intelligence can thrive, and the more we can trust the information it generates. This lends skepticism towards articles that attract only a few editors. However, the study found that editorial experience plays an important role as well. It shows that contributors grow more moderate in their editing behavior, i.e. the information they produce is less slanted and biased, the more articles they edit.

Figure C2 shows the amount of editorial experience and the number of editors for each legislator’s Wikipedia page in the database. We measure domain-specific editorial experience as the amount of legislators’ Wikipedia pages an editor contributed to within the respective legislature. We further weigh this amount by a contributor’s

Figure C2: Editor experience and number of editors on legislators' Wikipedia pages.

mean number of edits. Plotting the measures against each other, we find that legislators’ Wikipedia pages with few editors usually do attract more experienced contributors. As the number of editors on an article increases, the average amount of editorial experience decreases. Hence, editorial experience steps in when collective intelligence does not and vice versa.

To further verify the quality and accuracy of the specific information we offer, we compared part of the database against information available in other data projects, which relied on human coders. We picked two projects integrated with the CLD: (1) the Bundestag Roll Call Vote Data (BTVote) (Sieberer et al., Forthcoming), which includes roll call votes taken in the German Bundestag from 1949 to 2013, together with several personal characteristics of the respective legislators, and (2) Voteview (Lewis et al., 2019), which provides roll call votes and basic biographical information of legislators for all chambers of the United States Congress. We chose these datasets for comparison because of their longitudinal scope, and because data on roll call votes are among the most prominent and frequently used in political science. In addition, their coder-based mode of data collection is the gold standard in terms of data accuracy. A student assistant joined both datasets with the CLD using the identifiers offered in the CLD’s *IDs* table, harmonized variables available in both the CLD and the datasets to facilitate comparison, and filtered information where the CLD and the datasets disagree. In a case of disagreement, the student assistant searched for official information and noted whether the value in the CLD or the value in the other datasets is correct. Figure C3 presents results from the verification exercise. On the compared variables, the CLD shows very high agreement with information offered in BTVote and Voteview. Disagreement is either resolved in favor of the CLD or tied between the CLD and the other datasets. This shows that our crowd-sourced approach mirrors the human gold standard in terms of accuracy of information.

Finally, a student assistant conducted a Google search of official sources for a random selection of 500 legislators from the CLD to assess both the availability and accuracy of information. The assistant was instructed to search for “[legislator’s

Figure C3: Comparison of CLD to human gold standards on selected variables.

name] member of [country] parliament” on Google and screen the first landing page for official information. We considered pages such as personal, parliamentary, governmental, or campaign websites as official. Other sources such as Wikipedia and Wikidata were omitted. In case no official page was available for a legislator, the

assistant resorted to the Google info box to verify information. However, this happened only for a handful of legislators. Usually, there was either an official source or none at all. Once official sources were located on the first landing page, the student assistant searched those pages for a set of individual information, including sex, date of birth and death, ethnicity, religion, birthplace, place of death, legislative session, party, constituency, period of service. If the respective information was traceable and included in the CLD, the student assistant noted whether it confirmed (C) or disconfirmed (D) information in the CLD. If the information was traceable but not included in the CLD, it was similarly categorized as disconfirmed. If information was not traceable through official sources on the first landing page but available in the CLD, it was categorized as not found (NF). Figure C4 shows that for most variables, information in the CLD is confirmed by official sources. Interestingly, the share of information in the CLD that was disconfirmed by official sources is negligibly small, except for places of birth and death, and periods of service. However, the bulk of disconfirmed information for these variables arises because the information was traceable via official sources but not available in the CLD. For the other variables, if information could not be verified it was primarily because it was not traceable via official sources. This supports both the accuracy and coverage of the CLD.

Figure C4: Google search for information on 500 randomly drawn legislators in CLD.

Note: C = Confirmed, D = Disconfirmed, NF = Not found. Information not available in CLD but found through a Google Search was categorized as disconfirmed before computing shares. Otherwise, information not available in CLD is excluded here.

Appendix D Introduction to the R package

Access to the database is managed via the `legislatoR` R package. The package can be installed from GitHub as follows:

```
# install from GitHub
install.packages("devtools")
library(devtools)
install_github("legislatoR")
```

After having installed the package, it must be loaded via:

```
# load library
library(legislatoR)
```

The package provides dataset-specific function calls named after the datasets and preceded by “get_”. Each of these functions take only one single argument, the ISO 3166-1 alpha-3 code of the country for which a dataset shall be fetched from the server. Below are lists of datasets with corresponding functions as well as countries with corresponding codes.

Dataset	Function	Country	Function Argument
Core	<code>get_core()</code>	Austria	“aut”
IDs	<code>get_ids()</code>	Canada	“can”
Offices	<code>get_office()</code>	Czech Republic	“cze”
Political	<code>get_political()</code>	France	“fra”
Portraits	<code>get_portrait()</code>	Germany	“deu”
Professions	<code>get_profession()</code>	Ireland	“irl”
Social	<code>get_social()</code>	Scotland	“sco”
Wikipedia History	<code>get_history()</code>	Spain	“esp”
Wikipedia Traffic	<code>get_traffic()</code>	United Kingdom	“gbr”
		United States Congress	“usa_house/_senate”

Datasets can be accessed for specific countries the following way:

```
# access Core dataset for Canada
can_core <- get_core("can")

# access Traffic dataset for Germany
deu_traffic <- get_traffic("deu")
```

```
# access Political dataset for United States Senate
usa_senate_political <- get_political("usa_senate")
```

These commands download a whole dataset for a country and assign it to the R environment under the user-specified name provided to the left of the assignment arrow. Datasets are stored in the R environment as `data.frame` objects.

Usually, we wish for somewhat more targeted access, i.e., sociodemographic data about politicians from a specific legislative session or social media handles for politicians of a specific party. The relational database was designed to achieve this via filtering joins. To allow for such joins, the `dplyr` package is required.

```
# install the dplyr package
install.packages("dplyr")

# load the dplyr package
library(dplyr)
```

Extracting sociodemographic data for politicians from a specific session is fairly easy. Since the *Core* and *Political* datasets are directly linked via the Wikipedia page ID, this requires only one semi-join. A semi-join returns all rows from a dataset in X where there are matching values of a key from a dataset in Y, keeping only columns from X.

```
# extract sociodemographic data about United Kingdom politicians
# from the 42nd legislative session
uk_core_sub <- semi_join(x = get_core("gbr"),
                        y = filter(get_political("gbr"),
                                   session == 42),
                        by = "pageid")
```

This chunk of code downloads the *Core* and *Political* datasets for the UK. It then filters the *Political* dataset to only include members from the 42nd legislative session. Lastly, it extracts *Core* data of members appearing in the 42nd session *Political* dataset and assigns it to the environment. This process takes less than a second. It

assigns only a data frame of size 239.9 KB with 12 variables for 691 politicians into the environment, not the full *Core* and *Political* datasets called in the process and amounting to 6.9 MB.

Extracting social media handles for politicians who are members of a specific party is slightly more complicated. Since the *Social* and *Political* datasets are not directly linked via a common key, a detour through the *Core* dataset is required. In essence, this simply means that we need two instead of one semi joins.

```
# extract social media handles for German politicians who are members
# of the SPD party
deu_spd <- semi_join(x = get_core("deu"),
                     y = filter(get_political("deu"),
                                party == "SPD"),
                     by = "pageid")
deu_social_spd <- semi_join(x = get_social("deu"),
                           y = deu_spd,
                           by = "wikidataid")
```

The first chunk works similar as before. It first downloads the *Core* and *Political* datasets for Germany, filters the latter by SPD party membership, and assigns only *Core* data of respective politicians to the environment. The second call now downloads the *Social* dataset for Germany but keeps only politicians included in the previously extracted *Core* data for SPD members. To achieve the same with even less code and having to assign only the desired result into the environment, we can use the pipe operator from the *magrittr* package.

```
# install the magrittr package
install.packages("magrittr")

# load the magrittr package
library(magrittr)

# extract social media handles for German politicians who are members
# of the SPD party
deu_social_spd <- semi_join(x = get_core("deu"),
```

```

      y = filter(get_political("deu"),
                  party == "SPD"),
      by = "pageid") %>%
semi_join(x = get_social("deu"),
          y = .,
          by = "wikidataid")

```

The pipe operator `%>%` takes the output of a function call and passes it on to the next function call. The output of the previous call is referred to via a dot.

In addition to conditional subsets of datasets, we might also want to combine information from different datasets. This can be achieved via mutating joins, which are also implemented using the `dplyr` package. Lets say we wish to combine sociodemographic and political data.

```

# extract combined sociodemographic and political data for the Czech
# parliament
cze_soc_pol <- left_join(x = get_core("cze"),
                        y = get_political("cze"),
                        by = "pageid")

```

The left join that is used here appends the information stored in the *Political* dataset to politicians and information in the *Core* dataset. This results in some kind of panel data, where politicians are repeated the number of sessions they are observed in. A format like this can be used immediately to study individual party change between legislative sessions.

For those who wish to conduct their analyses using other software, such as EXCEL, SAS, STATA, or SPSS, the following code can be used to export the extracted data in the desired format.

```

# save data frame as .csv for use with Excel
write.csv(cze_soc_pol, "cze_soc_pol.csv")

# install haven package for export into other formats
install.packages("haven")

```

```
# load haven package
library(haven)

# save data frame as .sas for use with SAS
write_sas(deu_social_spd, "deu_social_spd.sas")

# save data frame as .dta for use with STATA
write_dta(can_core, "can_core.dta")

# save data frame as .sav for use with SPSS
write_sav(deu_traffic, "deu_traffic.sav")
```

We offer further specialized tutorials online at <https://cran.r-project.org/web/packages/legislatoR/vignettes/legislatoR.html>. There, we illustrate how to extract and combine data with external datasets, construct cross-country panels, loop over countries in the database, or combine filtering and mutating joins for even more targeted data access.

Appendix E Software Statement

All scripts were run under Windows 10 x86-64 using R version 3.4.2. The following R packages were used for assembling the database and for analyses reported in this paper.

Table E1: R packages.

coefplot (Lander, 2018)	pageviews (Keyes and Lewis, 2019)
cowplot (Wilke, 2019)	plyr (Wickham, 2011)
crayon (Csárdi and Gaslam, 2017)	purrr (Henry, Wickham and RStudio, 2019)
data.table (Dowle et al., 2017)	R.utils (Bengtsson, 2019)
dplyr (Wickham et al., 2020)	readxl (Wickham et al., 2019)
eeptools (Becker and Knowles, 2019)	reshape2 (Wickham, 2017a)
extrafont (Chang, 2014)	rvest (Wickham, 2016)
finalfit (Harrison, Drake and Ots, 2019)	stargazer (Hlavac, 2018)
ggplot2 (Wickham, 2009)	stringr (Wickham, 2017b)
gtools (Warnes, Bolker and Lumley, 2018)	tibble (Müller and Wickham, 2019)
haven (Wickham and Miller, 2019)	tidyr (Wickham, Henry and RStudio, 2019)
httr (Wickham, 2019)	tidyselect (Henry and Wickham, 2018)
jsonlite Ooms, Lang and Hilaiel (2018)	toOrdinal (Betebenner, Martin and Erickson, 2019)
lubridate (Grolemund and Wickham, 2011)	vroom (Hester et al., 2020)
magrittr (Bache and Wickham, 2014)	WikidataR (Keyes et al., 2017s)
mpoly (Kahle, 2019)	Wikipediatrend (Meissner, 2017s)
pacman (Rinker et al., 2017)	xtable (Dahl et al., 2019)
padr (Thoen, 2019)	zoo (Zeileis and Grothendieck, 2005)

References

- Bache, Stefan M. and Hadley Wickham. 2014. *magrittr: A Forward-Pipe Operator for R*. R package version 1.5.
URL: <https://CRAN.R-project.org/package=magrittr>
- Becker, Jason P. and Jared E. Knowles. 2019. *eeptools: Convenience functions for education data*. R package version 1.2.2.
URL: <https://CRAN.R-project.org/package=stringr>
- Bengtsson, Henrik. 2019. *R.utils: Various programming utilities*. R package version 2.9.0.
URL: <https://CRAN.R-project.org/package=R.utils>
- Betebenner, Damian W., Andrew Martin and Jeff Erickson. 2019. *toOrdinal: Cardinal to ordinal number and date conversion*. R package version 1.1.
URL: <https://CRAN.R-project.org/package=toOrdinal>
- Chang, Winston. 2014. *extrafont: Tools for using fonts*. R package version 0.17.
URL: <https://CRAN.R-project.org/package=extrafont>
- Csárdi, Gábor and Brodie Gaslam. 2017. *crayon. Colored terminal output*. R package version 1.3.4.
URL: <https://CRAN.R-project.org/package=crayon>
- Dahl, David B., David Scott, Charles Roosen, Arni Magnusson, Jonathan Swinton, Ajay Shah, Arne Henningsen, Benno Puetz, Bernhard Pfaff, Claudio Agostinelli, Claudius Loehnert, David Mitchell, David Whiting, Fernando da Rosa, Guido Gay, Guido Schulz, Ian Fellows, Jeff Laake, John Walker, Jun Yan, Liviu Andronic, Markus Loecher, Martin Gubri, Matthieu Stigler, Robert Castelo, Seth Falcon, Stefan Edwards, Sven Garbade and Uwe Ligges. 2019. *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4.
URL: <https://CRAN.R-project.org/package=xtable>

- Dowle, Matt, Arun Srinivasan, Jan Gorecki, Tom Short, Steve Lianoglou and Eduard Antonyan. 2017. *data.table: Extension of 'data.frame'*. R package version 1.10.4.
URL: <https://CRAN.R-project.org/package=data.table>
- Giles, Jim. 2005. "Internet encyclopaedias go head to head." *Nature* 438:900–901.
- Göbel, Sascha and Simon Munzert. 2018. "Political Advertising on the Wikipedia Marketplace of Information." *Social Science Computer Review* 36(2):157–175.
- Greenstein, Shane, Yuan Gu and Feng Zhu. 2017. "Ideological segregation among online collaborators. Evidence from Wikipedians." *NBER Working Papers* .
- Grolemund, Garrett and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40(3).
URL: <https://www.jstatsoft.org/article/view/v040i03>
- Harrison, Ewen, Tom Drake and Riinu Ots. 2019. *finalfit: Quickly create elegant regression results tables and plots when modelling*. R package version 0.9.4.
URL: <https://CRAN.R-project.org/package=finalfit>
- Henry, Lionel and Hadley Wickham. 2018. *tidyselect: Select from a set of strings*. R package version 0.2.5.
URL: <https://CRAN.R-project.org/package=tidyselect>
- Henry, Lionel, Hadley Wickham and RStudio. 2019. *purrr: Functional programming tools*. R package version 0.3.2.
URL: <https://CRAN.R-project.org/package=purrr>
- Hester, Jim, Hadley Wickham, Jukka Jylänki, Mikkel Jørgensen and RStudio. 2020. *vroom: Read and Write Rectangular Text Data Quickly*. R package version 1.2.0.
URL: <https://CRAN.R-project.org/package=vroom>
- Hlavac, Marek. 2018. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. R package version 5.2.2.
URL: <https://CRAN.R-project.org/package=stargazer>

- Kahle, David. 2019. *mpoly: Symbolic computation and more with multivariate polynomials*. R package version 1.1.0.
URL: <https://CRAN.R-project.org/package=mpoly>
- Kalla, Joshua L. and Peter M. Aronow. 2015. “Editorial bias in crowd-sourced political information.” *Plos One* 10.
- Keyes, Oliver and Jeremiah Lewis. 2019. *pageviews: An API client for Wikimedia traffic data*. R package version 0.5.0.
URL: <https://CRAN.R-project.org/package=padr>
- Keyes, Oliver, Serena Signorelli, Christian Graul and Mikhail Popov. 2017s. *WikidataR: API client library for 'Wikidata'*. R package version 1.4.0.
URL: <https://CRAN.R-project.org/package=WikidataR>
- Lander, Jared P. 2018. *coefplot: Plots Coefficients from Fitted Models*. R package version 1.2.6.
URL: <https://CRAN.R-project.org/package=coefplot>
- Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin and Luke Sonnet. 2019. *Voteview. Congressional roll-call votes database*.
URL: <https://voteview.com/data>
- Meissner, Peter. 2017s. *wikipediatrend: Public subject attention via Wikipedia page view statistics*. R package version 1.1.14.
URL: <https://CRAN.R-project.org/package=wikipediatrend>
- Müller, Kirill and Hadley Wickham. 2019. *tibble: Simple Data Frames*. R package version 2.1.3.
URL: <https://CRAN.R-project.org/package=tibble>
- Ooms, Jeroen, Duncan T. Lang and Lloyd Hilaiel. 2018. *jsonlite: A robust, high performance JSON parser and generator for R*. R package version 1.6.
URL: <https://CRAN.R-project.org/package=jsonlite>

- Rinker, Tyler, Dason Kurkiewicz, Keith Hughitt, Albert Wang and Jim Hester. 2017. *pacman: Package Management Tool*. R package version 0.4.6.
URL: <https://CRAN.R-project.org/package=pacman>
- Sieberer, Ulrich, Thomas Saalfeld, Tamaki Ohmura, Henning Bergmann and Stefanie Bailer. Forthcoming. “Roll-call votes in the German Bundestag. A new dataset, 1949–2013.” *British Journal of Political Science* .
- Thoen, Edwin. 2019. *padr: Quickly get datetime data ready for analysis*. R package version 0.5.0.
URL: <https://CRAN.R-project.org/package=padr>
- Warnes, Gregory R., Ben Bolker and Thomas Lumley. 2018. *gtools: Various R programming tools*. R package version 3.8.1.
URL: <https://CRAN.R-project.org/package=gtools>
- Wickham, Hadley. 2009. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
URL: <http://ggplot2.org/>
- Wickham, Hadley. 2011. “The split-apply-combine strategy for data analysis.” *Journal of Statistical Software* 40(1):1–29.
URL: <https://www.jstatsoft.org/article/view/v040i01>
- Wickham, Hadley. 2016. *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.2.
URL: <https://CRAN.R-project.org/package=rvest>
- Wickham, Hadley. 2017a. *reshape2: Flexibly reshape data. A reboot of the reshape package*. R package version 1.4.3.
URL: <https://CRAN.R-project.org/package=reshape2>
- Wickham, Hadley. 2017b. *stringr: Simple, consistent wrappers for common string operations*. R package version 1.2.0.
URL: <https://CRAN.R-project.org/package=stringr>

Wickham, Hadley. 2019. *httr: Tools for working with URLs and HTTP*. R package version 1.4.1.

URL: <https://CRAN.R-project.org/package=httr>

Wickham, Hadley and Evan Miller. 2019. *haven: Import and export 'SPSS', 'Stata' and 'SAS' files*. R package version 2.1.1.

URL: <https://CRAN.R-project.org/package=haven>

Wickham, Hadley, Jennifer Bryan, Marcin Kalicinski, Komarov Valery, Christophe Leitienne, Bob Colbert, David Hoerl and Evan Miller. 2019. *readxl: Read Excel files*. R package version 1.3.1.

URL: <https://CRAN.R-project.org/package=readxl>

Wickham, Hadley, Lionel Henry and RStudio. 2019. *tidyr: Tidy messy data*. R package version 1.0.0.

URL: <https://CRAN.R-project.org/package=tidyr>

Wickham, Hadley, Romain François, Lionel Henry and Kirill Müller. 2020. *dplyr: A Grammar of Data Manipulation*. R package version 0.8.3.

URL: <https://CRAN.R-project.org/package=dplyr>

Wilke, Claus O. 2019. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.0.0.

URL: <https://CRAN.R-project.org/package=cowplot>

Zeileis, Achim and Gabor Grothendieck. 2005. “zoo: S3 infrastructure for regular and irregular time series.” *Journal of Statistical Software* 14(6).