www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

Bulk Ingestion of Congressional Actions and Materials Dataset

Ryan Delano, Aaron Rudkin & In Song Kim ✉

Congressional data are essential for analyzing the U.S. law-making process, political networks, and policy outcomes. However, accessing up-to-date official data for scientific research often requires a framework to ingest large-scale, unstructured data from government sources, along with an automated pipeline to validate, curate, and synthesize these data. We introduce the Bulk Ingestion of Congressional Actions & Materials (BICAM) dataset, which includes eleven components of congressional activities, actors, and materials, covering all electronically available official records from 1789 to the present: *Bills*, *Amendments*, *Members*, *Committees*, *Committee Reports*, *Prints*, *Meetings*, *Nominations*, *Hearings*, *Treaties*, and *Congresses*. To support integration with external datasets and facilitate quantitative and qualitative research on the U.S. Congress, BICAM also provides standardized identifiers for each component. By linking BICAM to filings from the Lobbying Disclosure Act of 1995, we demonstrate its applicability and potential to advance empirical research on legislative processes.

Background & Summary

The United States Congress produces extensive data through its activities, which are critical for scientific research on American politics. Despite significant resources dedicated to studying Congress, accessing congressional data for research remains challenging. Bulk downloads are scarce and often disrupted by discontinued services, while comprehensive analysis to achieve a complete picture of congressional activities requires linking disparate datasets—many of which are sourced separately and lack standardized identifiers for integration. As a result, researchers face a trade-off between using government sources, which are authoritative but difficult to work with, or third-party solutions, which are often unreliable for long-term use, posing challenges for reproducibility and continuity. To address these issues, we introduce BICAM (*Bulk Ingestion of Congressional Actions & Materials*), a novel relational database that consolidates the following eleven respective datasets, covering all electronically available official records from 1789 to the present: (1) *Bills*, (2) *Amendments*, (3) *Members*, (4) *Committees*, (5) *Committee Reports*, (6) *Prints*, (7) *Meetings*, (8) *Nominations*, (9) *Hearings*, (10) *Treaties*, and (11) *Congresses*. We describe each component in detail in below in Data Records.

Figure 1 describes the institutional process through which legislation becomes law and highlights how some of the eleven components described above are instrumental to this process. For instance, a *Member* in one or both of the chambers of Congress sponsors (introduces) a *Bill*. It is assigned a number and referred by legislative leadership to a *Committee*, a subset of all members that focuses on legislation relating to specific topic or issue focus (taxation, science, commerce, trade, defense, etc.). The Committee considers the bill, holds *Committee Meetings*, produces a *Committee Reports* on the bill, and votes whether or not the legislation should be considered by the entire chamber. If the legislation is scheduled for debate in the chamber, members may make statements about or propose *Amendments* to the bill. A vote may then be scheduled; if the vote passes, the bill is said to have passed or cleared the chamber. Both chambers must pass the same bill in the same *Congress* to refer the bill to the U.S. President for signature or veto. Vetoed bills may be reconsidered by Congress for override of the veto. If the President signs the bill or the veto is overridden, the bill becomes law and can be enforced. BICAM encompasses all of the above steps, allowing researchers to conduct systematic analyses of the legislative process.

We overcome a number of specific challenges that limit the usability of existing data offerings. Government sources such as Congress.gov[1] and GovInfo.gov[2] — widely used in Political Science[3–5] — provide high-quality, official raw resources (and indeed serve as our data providers). However, these offerings often require considerable expertise to navigate effectively. While these government sites provide APIs (application programming interfaces) that allow researchers to programmatically query and download data, APIs pose three significant challenges. First, no ready-made software packages are available for querying these APIs, necessitating custom

Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. ✉e-mail: insong@mit.edu

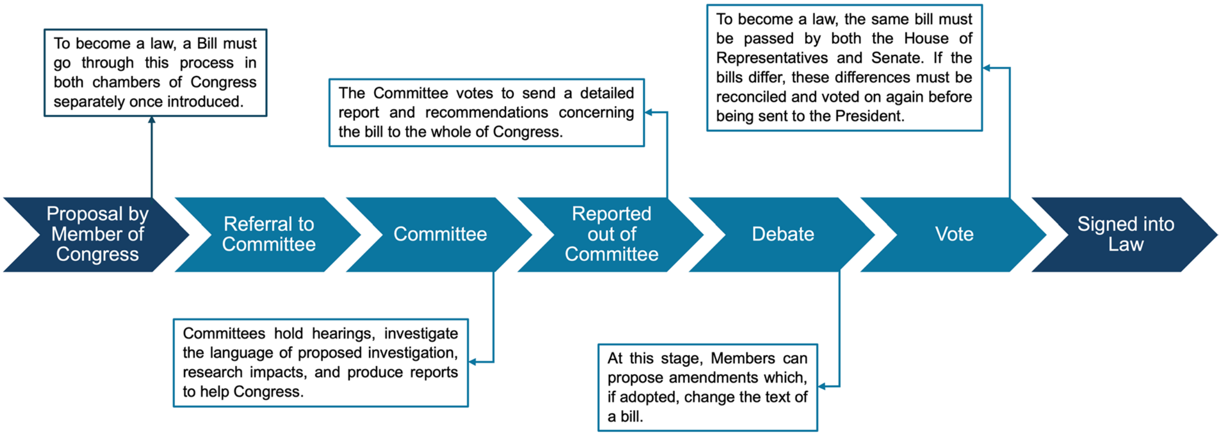
To become a law, a Bill must go through this process in both chambers of Congress separately once introduced.
The Committee votes to send a detailed report and recommendations concerning the bill to the whole of Congress.
To become a law, the same bill must be passed by both the House of Representatives and Senate. If the bills differ, these differences must be reconciled and voted on again before being sent to the President.
Proposal by Member of Congress
Referral to Committee
Committee
Reported out of Committee
Debate
Vote
Signed into Law
Committees hold hearings, investigate the language of proposed investigation, research impacts, and produce reports to help Congress.
At this stage, Members can propose amendments which, if adopted, change the text of a bill.

Fig. 1 Condensed timeline of U.S. legislative process. This figure depicts the process by which legislation becomes law, simplified and linearized for legibility. While there are many more intricacies and special cases than those that are captured in the figure, this general overview helps to contextualize the environment in which this congressional data is produced and disseminated. The darker blue steps only occur once per bill, while the lighter blue steps occur in each congressional chamber (House of Representatives and Senate) per bill.

programming, which introduces substantial and redundant costs. Second, the documentation provided is often limited in scope and inconsistent in quality, making it difficult to learn and use effectively. Third, these APIs impose restrictions on the volume of data that can be retrieved, significantly increasing the time required to prepare data for research—issues we discuss further in Methods. In an era where computational social science depends on large, comprehensive datasets spanning extensive time periods and integrating disparate types of data, these limitations create substantial barriers to research.

Given these challenges, third-party solutions have periodically emerged to fill the gap–most notably efforts by the Sunlight Foundation, which were for a time the canonical source for Political Scientists looking for data from Congress[6]. While third-party resources have been valuable, they often constrain researchers in two key ways: first, by limiting the ability to reproduce datasets based on raw data sources due to proprietary or non-public underlying technologies, and second, by the lack of long-term service availability required for sustained research. For example, researchers seeking congressional data from the Sunlight Foundation today would encounter difficulties, as their tools were discontinued in 2016. Sunlight referred users to ProPublica, which provided similar services until they, too, ceased their congressional data offerings in 2024. Similarly, other discontinued third-party sources include the New York Times Congress API and the GovTrack.us API, both of which are now non-public or closed. Researchers are well aware of the frustrations caused by these emergence and disappearance cycles, to say nothing of the challenges posed for the reproducibility of published research.

Of course no data provider can insulate against all possible futures: if the U.S. Congress stops gathering or making available data, all projects relying on the data will face insurmountable odds. Regardless, we believe that as academic researchers, we are well-situated to do better than the private organizations that preceded us. As researchers, our incentives are well aligned: we need this data for our own research and will for the foreseeable future. The marginal cost of making it available to the public on top of gathering it ourselves is limited, and we have allocated funding sufficient to cover operating expenses over the long-term horizon.

BICAM provides five key advantages. First, in contrast to commercial data providers (such as LegiScan[7]), BICAM is free. Second, we offer cleaned, vetted, and linked data that are available for immediate download without requiring any coding through the webpage https://bicam.net/. Third, we merge data from both upstream congressional APIs such that if one is discontinued in the future or migrated, we can easily adapt and in the meantime continue to provide the best data possible. Fourth, we release all source code and encourage open-source contributions, which allows researchers to guard against any future risks to our infrastructure and if necessary use our work as a baseline for future data offerings. Finally, we integrate additional datasets into BICAM, including record linkages to connect it with other important data sources, such as Voteview/NOMINATE[8] based on standardized identifiers. For researchers studying interest group politics, BICAM additionally includes a comprehensive set of record linkages between congressional bills and Section 16 of lobbying disclosure reports gathered under the Lobbying Disclosure Act of 1995, enabling researchers to conduct studies of interest group and industry lobbying activity with greater precision than any prior source[9]. Taken together, BICAM improves data quality and lowers barriers to conducting new research on the U.S. Congress.

We observe that Political Scientists have periodically made available specific components of this data. The Congressional Bills Project (CBP) makes available a broad swath of Congressional bills from 1947–2008[10]. While the CBP offers some advantages in terms of depth (in that it includes human-augmented coding of bill subject matter), our data covers a wider temporal period (from the 6th to current Congresses), and *Bills* are just one of many components of BICAM. The Comparative Agendas Project, similarly, offers hearing, law, and vote data, augmented with variables for coding policy issues[11]. The Comparative Legislators Database offers legislator (*Members*) data, including additional linkages to social media accounts and demographic data extracted from legislator biographies, and the Measuring American Diplomacy project offers a dataset of U.S. Treaties comparable to our analogous component[12,13]. Thus, researchers are advised that BICAM's primary advantage is its wide breadth (in terms of subject matter and temporal/historical coverage), easy record linkages, and rolling updates through the time of this article's publication. Researchers interested in deeper metadata for any one component, and for whom historical data is sufficient, may find bespoke data offerings more useful.

In the below data descriptor, we outline the construction of this dataset. Methods describes how we assemble the data. Data Records is a researcher focused introduction to the eleven components of BICAM. Technical Validation describes the steps we take to validate data integrity. Usage Notes provides practical guidance for accessing the data. Finally, Code Availability explains how researchers can access BICAM data and code or contribute to its development.

Methods

We begin by describing the method we use to build BICAM, which consists of three main steps: (1) collecting congressional data on a comprehensive array of activities from ground-truth (official) sources; (2) building a relational database which connects, validates, and synthesizes the data; and (3) augmenting the data with record linkages connecting BICAM to external data sets. The result is a comprehensive data source of congressional activity.

First, we gather ground-truth data from two sources maintained by the U.S. government: the Congress.gov API[1] and the GovInfo.gov API[2]. These are free government services that provide congressional data, subject to the limitations described earlier. While the two sources offer overlapping and complementary data, each has unique strengths and drawbacks. Why build BICAM rather than retrieving the data on demand? Two primary reasons: the process is technically complex, and it is slow. For example, retrieving the full dataset used to build BICAM requires interacting with over 100 API endpoints (specific sources for querying data) on Congress.gov[1]. Even with programming expertise to automate this task, researchers face a second challenge: the API imposes a limit of 5,000 requests per hour. Note that constructing our *Bills* component alone requires at least 10 requests to gather all the data for a *single* bill. Thus, it is clear that retrieving information for all 408,000 bills included in BICAM would be prohibitively time-consuming. Furthermore, filtering, sorting, and searching within this API are limited, creating additional hurdles. While GovInfo.gov[2] is comparatively easier to use and imposes fewer restrictions on data retrieval, it also covers a narrower scope of congressional data, lacking useful endpoints for matters like amendments and committee meetings. BICAM addresses these challenges by capturing all congressional data from both providers. This example illustrates a broader issue: bill data is just one of eleven components integrated into BICAM.

We solve both challenges: for researchers who wish to retrieve data from Congress themselves, BICAM is software that can be used to do so. We include functionality to pause and resume, to recover from errors, and to gather data more quickly in parallel. The code can be run to augment existing data with newly available data, allowing researchers to "catch up" periodically. But for the majority of researchers who are uninterested in gathering the data themselves, we make the full data available for bulk download. This first stage of data collection is powered by a custom Python package that retrieves, cleans, and stores virtually all of the data provided by both of these sources.

Second, having gathered the data, we import it into a database (technically, a PostgreSQL database), where they are assembled and connected. Relational databases connect imported data, automatically providing data validation benefits. We describe this validation further in Technical Validation. Records that contain errors are corrected using a series of rules. We merge records from our two upstream APIs to create a composite record, containing the most thorough information from both APIs if available.

For researchers who are comfortable using the SQL programming language and who want a more powerful way of querying data, we also make available the means to set up a relational database version of BICAM on researchers' local machines using open-source and publicly available code. This code is lightweight, with very few external dependencies, can flexibly use any SQL-based database backend, and can be run modularly: in short, the code allows researchers to future-proof their access to our data pipeline.

The resulting database is described in Fig. 2, which shows a mapping of relational connections between our main components. We use this relational database to build the static exports which we make available as CSV files to end users.

We describe each component in turn in Data Records below and describe how the relational database aids in data validation in Technical Validation. At this stage, we build additional helper tables to ease connection of components and to document data better for researchers.

Finally, we augment our existing data with record linkages to external datasets. Most modern congressional research involves connecting data about official congressional activities with external measures or behaviors: for example, connecting legislation (a congressional activity) to roll-call voting (a widely studied source of data, often provided by an external provider like GovTrack.us[14] or Voteview[8]), and from there to donation data (provided by the Federal Election Commission).

The components in BICAM include standardized versions of congressional record identifiers, but we also build two custom record linkages to key external sources: roll-call voting data and lobbying activity. The first record linkage connects the *Bills* component of BICAM, described below, to roll-call voting data and congressional ideology data provided by VoteView[8]. Mapping bills to votes requires a custom identifier (since bill numbers are used across multiple congresses). Second, we connect our bill data to lobbying reports filed under the Lobbying Disclosure Act of 1995 (LDA). LDA data is available publicly from the U.S. House and Senate, and is compiled and structured for researchers elsewhere by LobbyView[15] and OpenSecrets[16]. We make it easier to leverage those sources of data by creating a significantly improved record linkage between *Bills* and lobbying activity which detects significantly more cases of bill lobbying than existing sources[17], and for the first time also captures the full breadth of ex-post lobbying (lobbying of laws and implementation details after bills

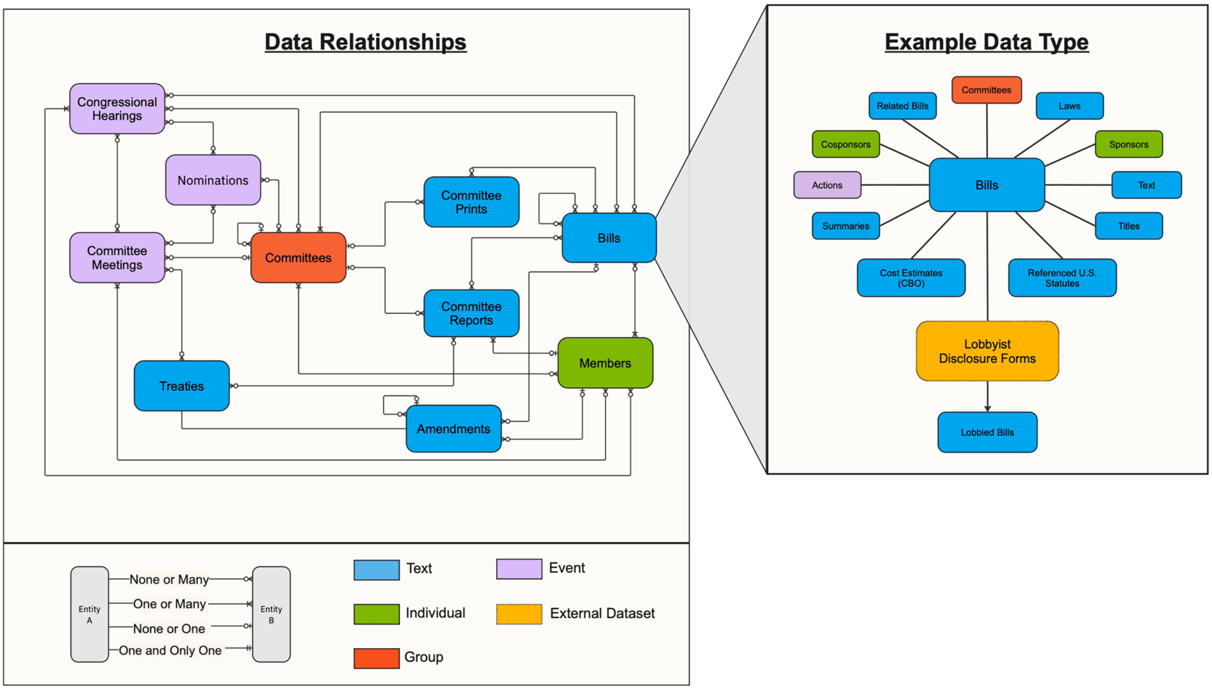

Fig. 2 Entity-relationship diagram of BICAM's architecture. Left panel: Components and their interconnections, with colors indicating categorical classifications: text-based content (blue), events (purple), individual actors (green), and group entities (red). Relationship lines denote cardinality (none/one/many) between connected entities. Right panel: Detailed view of the *Bills* component, showing 11 associated components, and demonstrating how BICAM's congressional data can be easily connected to external data sources. In this example, we connect BICAM to LD-2 congressional lobbying disclosure reports, making use of Section 16, which is used to disclose specific bills and issues lobbied. The schema demonstrates the dataset's comprehensive coverage of legislative activities and its interoperability with external data sources.

have already been adopted)[18]. We discuss our external record linkages in the "Bill-LDA Linkages" section of our Supplementary Information file.

Having summarized our process for constructing the data, we now provide an introduction to each of the 11 components of BICAM to assist researchers with using the data.

Data Records

BICAM is made available primarily as a series of `.CSV` files designed for easy viewing, downloading, and connection. Each component and sub-component is distributed as its own file, named the same as the component itself (the *Amendments* component is found in `amendments.csv`). Submission versions of this data are deposited with the Harvard Dataverse[19] and the most up-to-date versions of all data are available on our website at https://bicam.net. Users can directly load our data as a relational database (see the "Database Reconstruction by User" section of our Supplementary Information file for details) or use our BICAM-Collection package to gather data from the congressional APIs themselves. As previously mentioned, there are 11 main components of the dataset: *Bills, Amendments, Members, Committees, Committee Reports, Prints, Meetings, Nominations, Hearings, Treaties,* and *Congresses*. Each of these is subject to different temporal horizons, but we make available all the data that has been digitized and assembled upstream for each component.

BICAM includes virtually all of the data which the U.S. Congress makes available. We document a few exceptions here: we do not make available the full Congressional Record (transcript). Nor do we make available Congressional Communications (written statements presented to Congress by outside groups), reports of the Congressional Research Service, or press releases or reports written by individual members of Congress. These are excluded primarily due to project scope, and we anticipate incorporating these elements in due time. We also do not make available data produced by executive branch agencies, including political donations or election results (available from the Federal Election Commission). Finally, we do not provide rollcall vote data or lobbying data, both of which are well-covered by existing Political Science data sources, though we provide record linkages to help facilitate the use of those sources.

In this section, we present each of them in sequence. A comprehensive list of sub-components for each component is provided in the "Component and Sub-Component File Directory" section of our Supplementary Information file, and additional details can be found in the codebook documentation included with the data downloads. Users can download the complete BICAM dataset or individual components as zipped `.CSV` files.

Bills. Bills contain valuable information summarizing diverse legislative activities of Congress. A vast body of theoretical and empirical research, therefore, has been dedicated to understanding the structure and dynamics of congressional bills[20–23]. Bill text corpera, in particular, have been subject to extensive study[24–26]. The *Bills* component of BICAM includes a comprehensive dataset of 421,764 bills spanning the 6th to 119th Congress. Each bill is linked to other key components and related sub-components through a unique `bill_id`. These sub-components encompass congressional actions associated with bills (e.g., reporting, debating, voting, and scheduling the bill on the calendar), cost estimates, recorded votes, sponsors and co-sponsors, bill text, and more. Additionally, we provide a crosswalk that connects our bill identifiers to those used by Voteview[8], enabling seamless integration with detailed legislator ideology and voting behavior data.

A typical bill record is the *Build Back Better Act* (`bill_id = hr5376-117`). The primary *Bills* component has fields summarizing key attributes of the bill's congressional debate, for example that 283 amendments to this bill were proposed. Data supporting these summaries can be found by connecting to sub-components via `bill_id`. Connecting *Bills* to its *Bills-Texts* and *Bills-Sponsors* sub-components, we can read six versions of the bill's legislative text; connecting it to *Members*, we learn the bill was sponsored by John Yarmuth, and can connect the bill to other legislation he introduced, allowing for researchers to investigate co-sponsorship relationships[27–29].

Amendments. Amendments often reflect legislative bargaining and have therefore been closely studied in social science, particularly in the context of social choice theory and legislative agenda control[30]. BICAM includes 123,061 amendments which were introduced from the 97th to the 119th Congress. *Amendments* data is linked to other main components and related sub-components via `amendment_id` (comprised of amendment type, amendment number, and Congress number). Sub-components include congressional actions on amendments, recorded votes for those actions, sponsors, and texts. Each amendment can be linked to the bill, treaty, or amendment it amends (for example, by connecting *Bills* to *Amendments* via `bill_id`). The amendment with the `amendment_id = samdt5499-117` amends the *National Defense Authorization Act*; in turn, 736 further amendments to this amendment were proposed! Most amendments include both full-length and summary text changes, enabling researchers to conduct substantive analyses of how initial proposals develop into final laws and, in some cases, infer the strategies lawmakers use to advance their agendas.

Members. All 2,597 members of Congress that served from the 81st to 119th Congresses are included in the *Members* component of BICAM (including non-voting delegates and resident commissioners serving non-state U.S. territories). Member data is linked to other main components and related sub-components via `bioguide_id`, the same identifier used by Congress for their internal identification system. Sub-components include member terms, party history, and leadership roles. We additionally provide ICPSR (Inter-university Consortium for Political and Social Research) IDs, the de facto standard means for academic researchers to uniquely identify legislators, in the `icpsr` field. These two keys, together, make it possible to link BICAM's *Member* data to most other commonly used data sources on federal elected officials. One special case includes members who switch parties. *Joe Manchin III* (`bioguide_id = M001183`) left the Democratic Party in 2024 to serve as an Independent: if researchers need to treat party-switchers as multiple distinct legislators, they can use `icpsr`. Our data also includes contact information (office address, phone number) and information on the member's legislative district. The *Members* component can be easily linked to *Bills* (to attach members to sponsored legislation), *Committees*, and much of our other data.

Committees. Committees have long been considered perhaps the central institution of U.S. congressional politics[31–35] — famously described by Woodrow Wilson as "little legislatures" within Congress[36]. BICAM has a total of 226 committees and 587 sub-committees, with committee history spanning from the 1st to the 119th Congress. *Committee* data is linked to other main components and related sub-components via `committee_code` (the internal Library of Congress THOMAS ID for the committee). Sub-components for both *Committees* and their subcommittees include related bills and history, while subcommittees are linked to their parent committee and parent committees are also linked to *Committee Reports*.

Consider the *House Committee on Science, Space, and Technology* (`committee_code = hssy00`). All committees have codes ending in 00, while subcommittees have the same alphabetic representation as their parent committee with additional numbers (e.g., *House Sub-committee on Space and Aeronautics* [`committee_code = hssy16`]. HSSY has considered 8,880 bills in our data and released 330 reports. It has multiple subcommittees, which can be connected to committees either via the text schema described above or via the *Committees-Subcommittees* sub-component. This committee has also undergone several name changes, which are recorded in the *Committees-History* sub-component, beginning as the Committee on Science and Astronautics in 1958. By including name changes, we make it easier to connect our data to external, unstructured text data which may record historical names.

Committee Reports. The major outputs of committees are described in the next four components: *Committee Reports*, *Prints*, *Meetings*, and *Congressional Hearings*. We include 106,818 committee reports from the 16th to the 119th Congress. Committee reports are the summary of the investigation, findings, and disposition of the relevant committee towards legislation, and are believed to shape congressional voting[37]. This data is linked to other main components and related sub-components via `report_id` (comprised of report type, report number, and Congress number). Sub-components include associated bills, associated treaties, and texts.

Committee reports, associated with both *Bills* and *Treaties*, often include relevant text describing the potential impacts of legislation. These texts provide information on budget implications, expected testimony, and explanations of legal consequences. For example, the *Build Back Better Act* (`report_id = hrpt430-1-117`) includes over 1,500 pages in the Congressional Record.

Committee Prints. *Committee Prints* are internal research documents created by committees during their deliberations and represent one of the primary ways professional staff contribute to the policymaking process[38]. BICAM indexes 4,346 committee prints from the 67th to the 119th Congress. Committee print data is linked to other key components and related sub-components via `print_id`. Sub-components include associated bills, associated committees, and texts. Committee prints are less common than reports but provide an equally rich source of text for relevant legislation. For example, three prints were published about the *Build Back Better Act* (`bill_id = hr5376-117`) with respective `print_id`s of `hprt46233-117`, `hprt46234-117`, and `hprt46235-117`. Each was printed by the *House Rules Committee* (`committee_code = hsru00`) and describes proposed amendments and changes to the bill for the committee's internal reference.

Committee Meetings. We include records of 12,238 committee meetings (which include some, but not all, congressional hearings; see below) from the 114th to the 119th Congress. *Committee Meeting* data is linked to other key components and related sub-components via `meeting_id`, which utilizes an internal congressional event identification number. Sub-components include associated bills, treaties, nominations, meeting documents, witness documents, and witnesses. Beyond hearings, meetings cover broader committee work, such as bill markup (evaluating and recommending changes to bills). Committee meetings can be directly mapped to the *Bills*, *Treaties*, and *Nominations* components. One example is `meeting_id = 108090`, a hearing titled "Facebook: Transparency and Use of Consumer Data." The sole witness for this testimony was the CEO of Meta Platforms, Inc. (née Facebook), Mark Zuckerberg, who testified about the Cambridge Analytica scandal. Witness statements and other supporting documents (including letters of complaint issued by the Federal Trade Commission against Meta) are available in PDF format; in future releases of BICAM, we plan to extract and directly provide the text of these documents for easier searching.

Congressional Hearings. Testimony at hearings has been the focus of significant research by social scientists[39,40]. BICAM has a total of 74,918 hearings from the 79th to the 119th Congress. Congressional hearing data is linked to other main components and related sub-components via `hearing_id`. Sub-components include dates, transcripts, and associated bills. BICAM's hearing data for hearings on nominations can be easily mapped to presidential nominations, allowing users to read the exact comments on specific nominees prior to confirmation.

Treaties. We index 758 treaties considered from the 81st to the 119th Congress. *Treaty* data is linked to other main components and related sub-components via `treaty_id`. Sub-components include actions on treaties, countries involved in treaties, index terms, and titles. Treaties are the main means by which the United States enters into binding commitments to harmonize actions with other countries, and they include a variety of topics including extradition, marine pollution, and arms control[41]. Treaties can be easily linked to *Committee Reports*, *Amendments*, *Bills*, and *Committee Meetings*, so scholars can easily investigate recent developments, including the Kigali Amendment to the Montreal Protocol (`treaty_id = td117-1`), which expanded the list of substances banned due to their impacts on Earth's ozone layer.

Nominations. 43,858 Presidential nominations (including judicial, cabinet, and sub-cabinet offices) from the 97th to the 119th Congress are included in BICAM. The *Nomination* component is linked to other main components and related sub-components via `nomination_id`. Sub-components include actions on nominations, committee actions on nominations, associated hearings, positions up for nominations, and nominees for those positions.

Nominations can be filtered by type (civilian / military), by date, and whether the nomination was successful. Although Supreme Court (Merrick Garland's failed Supreme Court nomination: `nomination_id = PN1258-00-114`) and Cabinet nominations (Merrick Garland's successful Attorney General nomination: `nomination_id = PN78-07-117`) attract the bulk of scholarly attention[42], the broader presidential nomination power has also been the subject of research[43,44]. Over 2,000 nominations occur during each session of Congress, highlighting a rich source of data on federal bureaucratic leadership for researchers to investigate.

Congresses. BICAM has basic details of all 119 sessions of Congress, including their start year, end year, the type of session, and references to available published congressional directories associated with the Congress.

Summary. Every component in BICAM is clearly labeled, documented, and ready for connection with other components through a variety of key columns, many of which are described above. The work put into synthesizing, cleaning, and preparing this data serves a key purpose: making research on the activities of the U.S. Congress simpler, removing the need for individual researchers to make the connections by hand.

Technical Validation

This section provides an overview of the measures we take to ensure the integrity of the data contained in BICAM. The primary goal of our project is to synthesize and make original ground-truth data from the U.S. Congress accessible, addressing the complexities described above. Since Congress itself serves as a ground-truth arbiter of this data, our primary validation efforts focus on ensuring, through rigorous integrity checks, that the data we compile is complete, accurate to the sources, and correctly synthesized. We outline five validation measures employed after importing the data: download checks, data type checks, identifier uniqueness, record linkage, and synthesis of our two main data sources.

First, we validate that our data gathering from the original sources is complete. We ensure that all sequential records have no unexpected gaps, that all pages of all records have been downloaded, and that both APIs have finished gathering all available data.

Second, we gather the data from our source APIs and store it separately and unconnected. At this stage, our goal is to verify the integrity of the data type. Types refer to a set of rules that describe the values that a given field of the data can take. For example, if we expect that we are ingesting a Congress number that contains only digits like "118", type checks help to alert when our data sources return an incorrect value, like "MISSING" or "One-Hundred-Eighteenth." Because we use a relational database to connect the data, we can enforce type requirements on fields, which immediately surface errors when they occur and allow us to build edit and repair processes (see Fig. 2 for examples of various data types).

Next, we ensure that identifiers are present and unique. Having multiple instances of the same record raises issues of authenticity: how would we know if two bills have the same `bill_id`, and which is correct? We designate identifier columns as unique and primary keys for this purpose. These identifiers guarantee uniqueness, alerting us if this criterion is violated. For sub-components that do not have natural identifiers, we take a combination of columns to construct identifiers.

Fourth, we verify consistent linkages between all components. Since components and sub-components are connected by identifiers previously verified as unique, we can guarantee that the expected connections are present. For example, if an amendment references a bill, that bill must exist in our *Bills* component. This process automatically flags missing or potentially incorrect data, which can then be reviewed, edited, or removed.

Finally, we need to synthesize the data from our two ground-truth sources. Any records that overlap between both sources should be the same. In cases where they are not, we must intervene to correct the discrepancy. While one would expect the number of discrepancies to be small, in reality a considerable amount of manual effort must go into emending incorrectly formatted or missing data. A common, significant discrepancy involves GovInfo's API[2] failing to list the `bioguide_id` for members that use nicknames, such as for Charles "Chuck" Fleischmann with `bioguide_id = F000459`. The Congress.gov API[1] does handle this case, allowing our ingestion process to overcome this problem. Another category of data quality issues consists of incorrect representations of the `committee_code` identifier, such as using the letter "o" instead of the number "0" when referencing the House Natural Resources Committee's `committee_code` of `hsii00`". Our pipeline thus incorporates solutions to these and many other data errors through a rules-based approach in the merging step.

In the "Validating Bill-LDA Linkages" section of our Supplementary Information file, we describe additional measures taken to validate the record linkages between our *Bills* component and reports filed under the Lobbying Disclosure Act of 1995 (and in the "Appropriations Bills" section of our Supplementary Information file, appropriations bills specifically). Although these reports serve as a form of ground truth, they are unstructured, making the process of validating data extraction challenging. To further our goal of providing comprehensive congressional data, we match extracted bill numbers and titles from Section 16 of individual lobbying reports to our *Bills* component using distance measures, which provide a quantitative estimate of confidence. Matches with lower confidence are subjected to additional scrutiny. We have cataloged common sources of error and developed rules to mitigate them. The resulting record linkages are both high-quality and easy to use.

Usage Notes

In this section, we briefly describe how readers can use BICAM. The BICAM website offers CSV downloads of each of the component, along with associated documentation. Researchers interested in directly integrating BICAM data into their research workflow can use the BICAM Python package[45]. Researchers interested in using BICAM's data extracting and cleaning infrastructure to build their own relational database can use the BICAM-Collection Python package[46]. Once gathered, each component of BICAM can be linked to other components through linking identifiers: For example, the *Bills* component of BICAM can be linked to its sub-components (*Bills-Actions*, *Bills-Actions-Committees*, *Bills-Sponsors*, and *Bills-Lobbied*) via the `bill_id` identifier. In turn, *Bills* can be linked to the *Members* component via the `bioguide_id` identifier in the *Bills-Sponsored* sub-component. We offer an example of connecting BICAM to external data in the "Using BICAM Record Linkages" section of our Supplementary Information file.

Code availability

The newest version of all BICAM data, including updates made since submission of this article, is available for download at https://bicam.net. Package code for the BICAM-Collection package is available on GitHub at https://github.com/bicam-data/bicam. Package code for the BICAM data downloader package is available at https://github.com/bicam-data/bicam. Submission versions of BICAM's code and data are stored on the Harvard Dataverse at https://doi.org/10.7910/DVN/TALQZL. We recommend researchers interested in using BICAM data to seek the most current version from our website. This data and code is made available under the MIT license, with no further restrictions on access.

Received: 29 January 2025; Accepted: 31 July 2025;
Published online: 26 September 2025

References

1. Library of Congress. Congress.Gov API https://gpo.congress.gov/.
2. United States Government Publishing Office. GovInfo.Gov API https://api.govinfo.gov/.
3. Eatough, M. & Preece, J. Crediting Invisible Work: Congress and the Lawmaking Productivity Metric (LawProM). *American Political Science Review* **199**(2), 566–584, https://doi.org/10.1017/S0003055424000224 (2025).
4. Lorenz, G. Prioritized interests: Diverse lobbying coalitions and congressional committee agenda setting. *The Journal of Politics* **82**(1), 225–240 (2020).
5. Grier, K., Grier, R. & Mkrtchian, G. Campaign contributions and roll-call voting in the US House of Representatives: The case of the sugar industry. *American Political Science Review* **117**(1), 340–346 (2023).
6. Wilkerson, J. & Casas, A. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science* **20**(1), 529–544 (2017).

7. LegiScan, LLP. LegiScan https://legiscan.com.
8. Lewis, J. B. *et al*. Voteview: Congressional roll-call votes database. Technical Report (University of California - Los Angeles (2024).
9. Kim, I. S. Political cleavages within industry: Firm level lobbying for trade liberalization. *American Political Science Review* **111**(1), 1–20 (2016).
10. E. Scott Adler and John Wilkerson Congressional Bills Project http://www.congressionalbills.org/.
11. Jones, B. D. *et al*. Policy Agendas Project: Codebook https://www.comparativeagendas.net/ (2023).
12. Göbel, S. & Munzert, S. The Comparative Legislators Database. *British Journal of Political Science* **52**(3), 1398–1408 (2022).
13. Malis, M. & Thrall, C. U.S. Treaties from the UN Treaty Series: 1945–2022 https://doi.org/10.7910/DVN/AFRCZH.
14. Tauberer, J. GovTrack.us https://govtrack.us.
15. Kim, I. S. Lobbyview: Firm-level lobbying & congressional bills database. Technical report (MIT, 2024).
16. OpenSecrets. OpenSecrets.org Data https://opensecrets.org.
17. Kim, I. S. & Kunisky, D. Mapping political communities: A statistical analysis of lobbying networks in legislative politics. *Political Analysis* **29**(3), 317–336 (2021).
18. You, H. Y. Ex post lobbying. *The Journal of Politics* **79**(4), 1162–1176 (2017).
19. Delano, R., Rudkin, A. F., Kim, I. S. BICAM: Bulk Ingestion of Congressional Actions & Materials https://doi.org/10.7910/DVN/TALQZL.
20. Casas, A., Denny, M. J. & Wilkerson, J. More effective than we thought: Accounting for legislative hitchhikers reveals a more inclusive and productive lawmaking process. *American Journal of Political Science* **64**(1), 5–18 (2020).
21. Ballard, A. O. & Curry, J. M. Minority Party Capacity in Congress. *American Political Science Review* **115**(4), 1388–1405 (2021).
22. Shipan, C. R. & Volden, C. The Mechanisms of Policy Diffsion. *American Journal of Political Science* **52**(4), 840–857 (2008).
23. Kirkland, J. H. The relational determinants of legislative outcomes: Strong and weak ties between legislators. *The Journal of Politics* **73**(3), 887–898 (2011).
24. Wilkerson, J., Smith, D. & Stramp, N. Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science* **59**(4), 943–956 (2015).
25. Gerrish, S. M. & Blei, D. M. Predicting legislative roll calls from text. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011* (2011).
26. Lauderdale, B. E. & Clark, T. S. Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science* **58**(3), 754–771 (2014).
27. Fowler, J. H. Connecting the Congress: A study of cosponsorship networks. *Political analysis* **14**(4), 456–487 (2006).
28. Ringe, N., Victor, J. N. & Gross, J. H. Keeping your friends close and your enemies closer? Information networks in legislative politics. *British journal of political science* **43**(3), 601–628 (2013).
29. Andris, C. *et al*. The rise of partisanship and super-cooperators in the US House of Representatives. *PloS one* **10**(4), e0123507 (2015).
30. Ordeshook, P. C. & Schwartz, T. Agendas and the control of political outcomes. *American Political Science Review* **81**(1), 179–199 (1987).
31. Cox, G. W. & McCubbins, M. D. *Legislative leviathan: Party government in the House*. (Cambridge University Press, 2007).
32. Fouirnaies, A. & Hall, A. B. How do interest groups seek access to committees? *American Journal of Political Science* **62**(1), 132–147 (2018).
33. Powell, E. N. & Grimmer, J. Money in exile: Campaign contributions and committee access. *The Journal of Politics* **78**(4), 974–988 (2016).
34. Berry, C. R. & Fowler, A. Congressional committees, legislative influence, and the hegemony of chairs. *Journal of Public Economics* **158**, 1–11 (2018).
35. Groseclose, T. & Stewart, C. III. The value of committee seats in the house, 1947-91. *American Journal of Political Science* **42**(2), 453–474 (1998).
36. Wilson, W. *Congressional Government*. (Transaction Publishers, 1885).
37. Shepsle, K. A. & Weingast, B. R. The institutional foundations of committee power. *American Political Science Review* **81**(1), 85–104 (1987).
38. DeGregorio, C. Professional committee staff as policymaking partners in the us congress. *Congress & the Presidency* **21**(1), 49–65 (1994).
39. Ban, P., Park, J. Y. & You, H. Y. How are politicians informed? witnesses and information provision in congress. *American Political Science Review* **117**(1), 122–139 (2023).
40. Leyden, K. M. Interest group resources and testimony at congressional hearings. *Legislative Studies Quarterly* **20**(3), 431–439 (1995).
41. Abbott, K. W., Keohane, R. O., Moravcsik, A., Slaughter, A. & Snidal, D. The concept of legalization. *International Organization* **54**(3), 401–419 (2000).
42. Krutz, G. S., Fleisher, R. & Bond, J. R. From Abe Fortas to Zoe Baird: Why Some Presidential Nominations Fail in the Senate. *American Political Science Review* **92**(4), 871–881 (1998).
43. Lewis, D. E. *The politics of presidential appointments: Political control and bureaucratic performance*. (Princeton University Press, 2010).
44. Hollibaugh, G. E. Jr. & Rothenberg, L. S. The when and why of nominations: Determinants of presidential appointments. *American Politics Research* **45**(2), 280–303 (2017).
45. Delano, R., Rudkin, A. F., & Kim, I. S. BICAM Python Package https://github.com/bicam-data/bicam/ (2025).
46. Delano, R., Rudkin, A. F., & Kim, I. S. BICAM-Collection Python Package https://github.com/bicam-data/bicam-collection (2025).

Acknowledgements

The authors thank Becca Sealy for excellent research assistance for this project and others.

Author contributions

R.D. built the BICAM and BICAM-Collection Python packages, website, and made improvements to the record-linkages of congressional bills to lobbying disclosure reports as described in the “Bill-LDA Linkages” section of our Supplementary Information file. A.R. and I.S.K. (corresponding author) wrote initial versions of the congressional data scraping packages and record-linkage algorithms. All authors contributed equally to writing this data descriptor. A.R. revised the data descriptor in response to feedback from reviewers and packaged final data outputs.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05737-8.

Correspondence and requests for materials should be addressed to I.S.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.