Describing the nature of legislation through roll call voting in the Chilean National Congress, a linked dataset description

Francisco Cifuentes-Silva\textsuperscript{a,b,*}, Felipe Rivera-Polo\textsuperscript{b}, Jose Emilio Labra-Gayo\textsuperscript{a} and Hernán Astudillo\textsuperscript{c}

\textsuperscript{a} Department of Computer Science, University of Oviedo, Oviedo, Asturias, Spain
E-mails: francesco.cifuentes@weso.es, labra@uniovi.es
\textsuperscript{b} Victoria s/n Valparaíso, Biblioteca del Congreso Nacional de Chile, BCN, Chile
E-mails: francesco.cifuentes@weso.es, frivera@bcn.cl
\textsuperscript{c} Department of Informatics, Universidad Técnica Federico Santa María, Santiago Chile
E-mail: hernan@inf.utfsm.cl

Abstract. This article shows the dataset and Semantic Web model that includes ontologies, RDF Shape expressions and URI patterns for describing bills and roll call voting of bills (legislation), processed by the Library of the Chilean National Congress during the last three decades (1990 - 2020). The data includes over 1.2 million RDF triples about voting events belonging to more than 2,700 bills, related to more than 500 distinct Chilean Members of Congress. The dataset is integrated with others, such as those about parliamentary biographies, legal norms, bills, and debates from sessions.

The dataset offers a public SPARQL endpoint which allows to develop different analysis. As an example, we present an exploratory analysis about polarity and political alignment metrics, verifying that it is possible to establish quadrants to classify the nature of bills into four categories defined by: I) Ideological stance (high polarization, high alignment), II) Personal interests (high polarization, low alignment), III) Sectoral interest (low polarization, low alignment) and IV) Technical consensus (low polarization, high alignment). We have identified that in each quadrant there is a co-existence of bills related to underlying topics, evidencing that voting is not totally governed by ideological factors. This type of information is relevant to describe the political scene and to carry out more in-depth studies in political science, establishing mechanisms that allow to optimize the legislative process, promoting more efficient parliamentary work, increase transparency and helping to improve the image of Congress and its members by the citizens.

Keywords: Linked dataset description, Semantic Web, Legal informatics, Polarization, Political alignment, Linked Open Data

1. Introduction

The arrival of the 21st century has come hand in hand with an important questioning of the legitimacy of the institutions of liberal representative democracies, which has been defined as a break between those represented and their representatives [1]. This rupture has led the politics to branch out, rejecting any form of mediation or political representation [2], even ceasing to have the nation-state as the sole interlocutor. It is in this context that participation and transparency have acquired increasing political and social importance, as there is a growing demand for a direct democracy mechanisms without intermediation [3].

At the same time, the integration of information technology in economies, has become today a key element to contribute to the development of nations. Although IT has developed globally in recent decades, the gap in technological development in areas such as AI and computerization of the State, between developed countries and Latin America is evident [4], which translates into comparative advantages: greater deliv-
In this context, this work presents a detailed description of a dataset using semantic web technologies, the automated loading process, and an example of consumption of published data. The latter is done through an analysis of roll-call votes, where a novel method is presented to classify bills in one of four quadrants, where each quadrant defines the group behavior that Members of Congress have (ideological stance, personal interests, sectoral interest and technical consensus), and in turn latently, allows identifying those issues that emerge associated with these types of behavior.

For the first part, metrics of political alignment $A$ and polarization $P$ are defined and calculated, both in a range from 0 to 100% obtaining a coordinate $(A, P)$ which determines the quadrant to which the vote belongs. Finally, all this exercise is performed directly using SPARQL and is available for query using R language scripts in a public repository.

Although the above mention concepts of political alignment and polarization have been widely studied in political science, the use of semantic web technologies, and particularly its use in the field of open data, marks precedents in transparency that offer new analytical possibilities, enhancing the reproducibility of the studies.

2. Semantic Web data model

All the data model has been developed according to the Linked Open Data best practices [16], with the idea of enabling a natural interoperability mechanism that allows to share public information. In this context, the main semantic web pieces of the dataset are: the definition of OWL ontologies, the formalization of shapes (ShEx/SHACL) to describe the data model, the description of the dataset \(^5\) by means of DCAT version 2 vocabulary [17], and the usage of URIs for identifiers in conjunction to a Linked Data frontend to access RDF resources.

2.1. Ontologies

The dataset model is expressed by means of two ontologies (the Biographies ontology \(^3\) and the Legislative resources ontology \(^4\)) as well as by reusing concepts from other ontologies and vocabularies like Dublin Core, Time Ontology, SKOS, BIO and Wikidata.

The two base ontologies of the data model are composed by more than one hundred classes and a similar number of properties. In the case of the Biographies ontology, it describes political characters in terms of their public personal information (name, picture, web

\(^{1}\)http://opendata.camara.cl

\(^{2}\)Available on named graph http://datos.bcn.cl/recursocatalogo/votaciones

\(^{3}\)http://datos.bcn.cl/ontologies/bcn-biographies/doc

\(^{4}\)http://datos.bcn.cl/ontologies/bcn-resources/doc

\(^{5}\)bcnbio - benares at prefix.cc
page, URI, birth/death date/place, wikidata URI, etc.), political party affiliation by period, public offices by period, or related persons, among others. The Legislative Resources ontology describes a partial data model of legislative processes, which includes document types and structures, stages in the process of law discussion, voting procedures, semantic annotation of debates, and others. Both ontologies are written in Spanish (using CamelCase convention) but documented also in English by means of RDF language tag annotations.

2.2. RDF Shapes

In order to improve the usage specification of the dataset, an RDF Shapes model has been published. RDF Shapes are an emergent component of the Semantic Web Stack that provides a method to describe and validate RDF data, describing shapes or the topology of a node group in the context of a specific RDF graph, extending the expressivity of data specification, and filling a validation space not covered by ontologies and vocabularies. ShEx [18] and shapes constraint language (SHACL) [19] are the most widely accepted proposals to define and validate RDF graph’s topology, and although SHACL has become a W3C recommendation [20], ShEx is being used in many different scenarios [7, 21–24] due to its concise and human-readable syntax, and an increasing set of open source and community tools that are currently being developed. The diagram in Figure 1 shows the Shape Expressions model⁶ which was created with the ShEx author tool⁷, and is available to validate the described dataset. Consequently, the SHACL specification was developed starting from ShEx, using the shEx2Shacl conversor tool from RDFShape⁸.

2.3. URI patterns

In order to follow the linked data principles, Cool hierarchical URIs⁹ were used, as they allow to partially reflect the conceptual scheme behind the data they represent. This facilitate debugging and quality control, which is important since the entire data model is used in other systems that are in production within BCN.

### Table 1

<table>
<thead>
<tr>
<th>URI pattern</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>/persona/{number} example</td>
</tr>
<tr>
<td>Bill</td>
<td>/cl/proyecto-de-ley/{pattern1} example</td>
</tr>
<tr>
<td>Voting</td>
<td>/cl/proyecto-de-ley/{pattern1}/votacion/{number} example</td>
</tr>
<tr>
<td>Vote</td>
<td>/cl/proyecto-de-ley/{pattern1}/votacion/{number}/voto/{number} example</td>
</tr>
<tr>
<td>Militancy</td>
<td>/persona/{number}/militancia/{number} example</td>
</tr>
<tr>
<td>Political Party</td>
<td>/cl/organismo/partido-politico/{pattern2} example</td>
</tr>
</tbody>
</table>

### Regex patterns and URI prefix

*number*  
*pattern1*  
*pattern2*  
*bcn*  
*bcnp*  

All dataset URIs are defined under the following URL base:

http://datos.bcn.cl/recurso/
2.4. Linked data frontend

We have implemented a linked data front-end following the preceding URI patterns using WESO-DESH\textsuperscript{11}, a Java linked data front-end (LDF) tool. This application offers RDF data as both human and machine readable ways, and was already employed in the first version of the BCN data portal \[5\]. It comprises the following main features:

- A native HTML+RDFa output,
- Content negotiation using HTTP 303 code,
- Independence of SPARQL endpoint technology, which allows to show federate results as single resource,
- URIs defined through regular expressions, allowing to design complex URI patterns (hierarchical or REST queries),
- It allows to execute multiple types of SPARQL queries (CONSTRUCT, ASK and DESCRIBE).

\textsuperscript{11}https://github.com/weso/weso-desh

WESO-DESH enables to use dereferenceable URIs, and brings an optimized view of RDF data, in a similar way to applications such as Pubby\textsuperscript{12}.

3. Data processing

In order to curate the political and legislative Knowledge Graph, the data has been obtained from multiple sources, and consequently has been particularly processed and transformed for each case. The main sources of data are the two Chilean National Congress chambers, which provide an open data portal with XML Web Services and data about legislative process, as well as their own web pages.

Other important source of data is the BCN Archive, and in particular, the Political History portal and repository\textsuperscript{13}, where among several resource types, the parliamentary biographies are published and maintained. It should be mentioned that, although these three sources of data are common bodies of the National

\textsuperscript{12}http://wifo5-03.informatik.uni-mannheim.de/pubby/

\textsuperscript{13}https://www.bcn.cl/historiapolitica
Congress, they do not implement de facto a common standard or web service schema, hindering a clear and consistent integration of data published by the Chamber of Deputies and the Senante separately. Indeed, in each of the chambers, the published Web services are described by different XML schema and details. For example, the lists of active senators and deputies have distinct and disjoint identifiers (even name descriptors or dates are described under different standards or formats).

This aspect also occurs with other kinds of resource types, such as party militancies and information about bills and voting, all of which are not integrated either (with the exception of bill number which is a functional code), and there are even restrictions on the limit of data allowed to harvest. This scenario has hampered data processing and curation. Nevertheless, due to an early strategic decision to use Semantic Web technologies in the BCN, this labour has been carried on in an incremental and progressive manner during years, keeping to date several processes that automate the data integration.

With regards to the mechanisms of data acquisition, it has been mixed, a part harvested from various XML Web services from the legislative congress open data page, as well as a web scraping processing from the Congress chambers web pages. Once captured, the data has been curated, integrated and modelled in RDF using the Legislative Resources ontology (which includes bill voting), finally being published as Linked Open Data.

LOD publishing is a policy established at BCN since 2011, where the first legal norms ontology were published[5]. Thence, a variety of datasets and vocabularies have been published in RDF at the LOD portal through its public SPARQL endpoint14, among which are the bill voting and the biographies dataset, which are the data sources of this work.

To date, the dataset of bills is composed of 1.2 million of RDF triples, while the dataset of Members of Congress is close to 433,000 RDF triples. Regarding the stability of dataset, a relevant part of the data contained are used in strategic products of the BCN[6], and all data are part of the institutional core business in terms of archival of political history of Chile. In these terms, the published KG is mature and relatively stable, although it has a sustained growth, and some minor errors need to be repaired once detected.

In the following sections we describe more details about the data that has been used and their sources.

3.1. Members of Congress and Political Parties dataset

This dataset is composed of information from all Members of Congress and political parties that have been part of the National Congress (which is bicameral), having information from the 1990 period onwards. The data, published as Linked Open Data in RDF, provides basic information about each person, their periods of membership to political parties and parliamentary positions held from the aforementioned date, which is open to the public under the principle of transparency.

The data was collected from an institutional wiki (based on MediaWiki) where biographical reviews of the main political actors in the history of the country are stored, archived and maintained. This institutional wiki, developed in 2010, contains RDFa marks that have been extracted and transformed into RDF triples in accordance with the URI convention described in Table 2.3. Although a large amount of data was normalized during this process, due to the fact that the Wiki did not have validation mechanisms of the inputs, there have been minor errors related to formats and some inconsistencies in the information such as duplicate periods of militancy or dates in different formats (such as descriptive text or other formats other than ISO-8601) that have progressively been corrected manually.

Although the database contains over 4,500 people related with the political history of the country, the total number of different Members of Congress who have participated in project voting during the period analyzed is 555.

This can be explained because many Members of Congress have been reelected in the same chamber or have changed chambers between elections (usually from the Chamber of Deputies to the Senate), as well as the incompleteness of voting in the entire period. Although this in part demonstrates the low turnover of Members of Congress in the previous 30 years, during 2020 a law was passed and published that limits the re-election of authorities [25] with retroactive effect, being able to remain in the Senate a maximum of 2 terms of 8 years (16 years in total) and in the Chamber of Deputies a maximum of 3 terms of 4 years (12 years in total) if they are reelected.

---

14http://datos.bcn.cl/sparql
3.2. Bills dataset

A bill is a document presented in the National Congress, whose function is to propose a legal text to be discussed by the Congress and to create a new law. The presentation of a bill in Chile can be carried out at the initiative of the executive branch (called Presidential Message), or by a Member of Congress (Parliamentary Motion). Generally speaking, each bill is entered into legislative proceedings, and enters a workflow where both chambers participate, where the proposed legal text is evaluated in full (in general) and at the level of its basic normative units (in particular) by the Members of Congress.

During this evaluation, the votes are carried out to reach a consensus on the views of the legislators and define the final version of the law that will be published. The process of processing the law involves great complexity according to its regulations, which will not be exposed in this article. However, the ontology of legislative resources brings an overview of the process in its main stages (Constitutional and Regulatory Procedures defined by bcnres:TramiteConstitucional and bcnres: TramiteReglamentario respectively), as well as various aspects that are currently processed, recorded and published as open data, within which there are various types of entities, documents and link properties among others.

Figure 2 shows a graph with the distribution by type and year of the bills published in RDF on the open data portal, differentiating the initiatives of the Executive Power from those carried out by Members of Congress.

Within the data there are 21 bills prior to year 1990, which have been created in the database to digitize relevant historical norms or that remain in force, such as political constitutions and other norms created during the 1973–1989 dictatorship period (Other).

The graph shows data from 1978 onwards, although there is a bill that was created to build the history of the 1925 constitution. These data have been obtained mainly from three different sources: 1) the BCN project processing database, 2) a database created in 1990 that was replaced in 2010 by the Web services that provide the open data portal of the Congress (with which there is currently an automatic update service), and 3) by manual creation from the History of Law system [6].

The query in Figure 5 can be executed in the SPARQL endpoint to obtain the RDF representation of roll call voting data of bill 9404–12, which returns a set of datatype values and dereferenceable URIs.

4. Proof of concept

The main idea of the analysis is to characterize bills by two metrics: political alignment and polarization. In this way, through SPARQL the voting events and votes of each bill are obtained, as well as the voting Members of Congress and their political party. Based on these data, the coefficient of each vote is calculated using two algorithms:

1. Political alignment coefficient, which indicates the degree of cohesion in the vote that Members of Congress have with respect to their party (only in the context of voting).

2. Polarization coefficient, which indicates the degree to which the vote divides the group of voters into opposite poles.

Subsequently, the average values of each index are calculated for each bill, allowing the project to be characterized by a single value for each metric. With these values at project level, a scatterplot is constructed with political alignment on the X axis and polarization on the Y axis. Finally, on the diagram, quadrants associated with ranges are defined in the values of the indices (polarization $\geq 50\%$ high, $<50\%$ low, alignment $\geq 70\%$ high, $<70\%$ low), allowing four quadrants to be set. A category has been assigned to these quadrants, which has been built inductively, taking as a reference the types of projects voted associated with each quadrant. In this way, four categories are established, which are defined below:

1. **Ideological stance**: bills with high polarization and high alignment in voting. This category establishes a differentiation in the political axis between left and right, so that the projects voted are ideologically ordered.

2. **Personal interest**: bills with high polarization and low alignment in voting. This category establishes a differentiation between the parliamentarian and his political coalition, which indicate personal interests over party principles.

3. **Sectoral interest**: bills with low polarization and low alignment in voting. This category contains projects of sectoral or territorial interest, so the parliamentarian is a representative of these interests, and the antagonism is against the disinterest of other Members of Congress.
4. Technical consensus: bills with low polarization and high alignment in voting. This category contains those projects where technical consensus was established, and with no political antagonisms in voting.

4.1. Metrics

Below we describe the algorithms to use in calculating the polarization and political alignment indexes.

4.1.1. Political alignment

Political alignment will be defined as a characteristic that describes the degree of convergence or coincidence that occurs within a group of individuals with respect to a certain opinion. Other variants of the political alignment (or just alignment) concept that are considered synonymous for the purposes of this article are party cohesion and party discipline [26]. This metric can be used both at the group level (political party or coalition), personal (Member of Congress depending on the group), by bill or by voting event. In particular, in the case of Member of Congress votes on bills, the political alignment describes the degree of similarity in the votes of a group of parliamentarians from the same political party.

Stated in formal terms, we will describe the group alignment as follows:

\[ A_g = \frac{\sum_{i=1}^{n} A_i N_i}{N} \]

(1)

Where:
- \( A_g \) corresponds to group alignment.
- \( A_i \) corresponds to the alignment of the subgroup of individuals who voted for the option \( i \).
- \( N_i \) corresponds to the total number of individuals who voted for the option \( i \).
- \( N \) corresponds to the total number of individuals in the group.

where \( A_i \) is defined as follows:

\[ A_i = \frac{N_i}{N} \]

(2)

where:
- \( A_i \) corresponds to the alignment within the group of those who voted for option \( i \).
- \( N_i \) corresponds to the total number of individuals who voted for option \( i \).
- \( N \) corresponds to the total number of individuals in the group.

In this way and simplifying with an example, if within the same group, in a specific vote the total number of individuals vote against, the alignment of the group is 100%, since they all vote the same. In another hypothetical scenario, if half of the individuals from the same group (for example the same party) vote in favour, and the other half against, the group alignment is 50%, given that the group globally had an opinion divided, although internally there was alignment.

The published social science literature constantly refers to the Rice Index[27] (and variations[28]), to
calculate the cohesion or degree of agreement within a voting event. However, this indicator allows only having a single metric for a complete group under analysis (such as a political party for example), penalizing the entire group for the differences within it. In our version of the political alignment coefficient, it is possible to associate an independent value for each person and vote, as well as for the entire project, obtaining more representative values from that perspective. This, in turn, enables Members of Congress to be characterized through a metric associated with their alignment and the value of their vote. This offers a wider application range than the Rice-Index, without performing complex calculations. For the cases described in the previous example but using the Rice-Index, the maximum alignment would correspond to 100%, but if the vote were divided exactly 50% within the group, the alignment value would be equal to 0%. The image 3 describes the behaviour of Rice-Index, Cos-Rice-Index (variant) and Alignment metrics seen as functions.

4.1.2. Polarization

In the context of legislative votes, polarization will be defined as the lack of agreement on an issue, which leads to a universe of voters grouping into two politically opposed positions. The level of polarization is maximum when there are two groups with an equivalent number of voters facing each other, while it is minimum when the voting universe votes for the same option. The graph in Figure 4 shows the behaviour of the polarization function when testing with different percentages of yes/no votes.

It is important to consider that for polarization only the extreme values (yes/no) are considered, therefore other types of votes are omitted for the calculation or normalized to one of the two options. The interpretation of voting options other than – yes/no – is all-ways relative to the political context, since both abstention and other voting options may represent different grounds. However, in practice, the approval of the vote is achieved by obtaining a certain quorum, which translates into having enough votes in favour.

Considering the above, the formula to calculate the polarization index is as follows:

\[
C_f = P_g = 1 - \sigma_p \times \sqrt{2}
\]

Where:

- \(C_f\) corresponds to the polarization coefficient for votes in favor
- \(C_c\) corresponds to the polarization coefficient for the votes against
- \(N_f\) corresponds to the total votes in favour
- \(N_c\) corresponds to the total votes against

4.2. Data

We have done an experiment using 15,874 voting events, that belong to 2,707 bills. Table 4.2 shows the descriptive statistics about the composition of data corpus\(^\text{15}\). Additionally, for the analysis, the Members of

\(^{15}\)Data available in December 23th of 2020
Congress and political parties dataset available in the data portal was used. Regarding the table data, we note that:

- Some voting events present a number of votes smaller than the total members of the chamber. This is produced mainly by the incomplete register of old bill votes (near to year 1990).
- Votings related to the max number of votes are related mainly to budget law discussion, when a high number of voting events are realized.
- The variant number of Members of Congress through the period also affects the register of votes. Indeed, in 1990 the lower chamber was formed by 120 deputies, while the Senate by 38 members. In 2020, the Chamber of deputies has 155 members and the Senate 43.

The SPARQL query in Figure 5 shows how to get votes for a specific bill from the RDF triplestore. At this point, it is relevant to unveil some design decisions about the experiment:

- Only the types of votes Yes (+) and No (-) have been analyzed. Although there are other rarely used types, these are considered irrelevant in this experiment.
- It is possible to carry out this analysis considering general and particular votes separately, however, to simplify the experiment, both are used interchangeably.

4.3. Data analysis

The first thing that is possible to do is a characterization of the data under analysis. In this sense, the graphs in Figure 6 show in an aggregate way how the polarity and political alignment values are distributed for each camera according to the analyzed data.

When viewing the alignment and polarity distribution graphs in Figure 6, in each of the chambers for the entire period, it is possible to affirm that in terms of political alignment, the senators have a behaviour much more aligned in their way of voting than the members of the Chamber of Deputies. Conversely, in the case of polarity, members of the Senate have a less polarizing behaviour than in the lower house.

Figure 7 shows a scatter diagram where each point represents a bill positioned in one of the four defined quadrants (similar to a Cartesian plane), according to its average polarization and alignment value. It can be seen that the quadrant with the highest number of projects corresponds to the one with low polarization and high alignment, that is, the quadrant previously defined as Technical Consensus.

The way in which Members of Congress are grouped in these projects is better visualized in Figure 8, which represents force graphs calculated with a distance
function between Members of Congress given their voting form. If the Members of Congress vote the same, their distance is 0, and if they vote differently, the distance is 1. This calculation is performed for each voting event of the bill and for all Members of Congress, obtaining the average distance values in a bill for all pairs. At the same time, the red and blue colours have been used to identify the Members of Congress associated with parties of the right or left.

In this way, it can be seen that in quadrant I, called Ideological stance (high polarization, high alignment), graphs are presented (one for each camera) where nodes of similar colour (same political tendency) are closely grouped and polarized with respect to the other group. Discussions like the bill to decriminalize abortion\footnote{http://datos.bcn.cl/recurso/cl/proyecto-de-ley/9895-11} belong to this quadrant.

In quadrant II of Personal interests (high polarization, low alignment), nodes are not grouped by color, but proportionally polarized groups are displayed. An example of bill in this quadrant is titled “Prohibit and penalize driving while smoking”\footnote{http://datos.bcn.cl/recurso/cl/proyecto-de-ley/3836-15}.

In quadrant III of Sectorial interests (low polarization and alignment), voting has a diffuse ordering, and in fact some of them have missing votes due to absences, which may explain their lower number. An example of this quadrant is the bill titled “Facilitate the call for municipal plebiscites”\footnote{http://datos.bcn.cl/recurso/cl/proyecto-de-ley/4228-06}.

Finally, quadrant IV about Technical consensus (low polarization, high alignment), shows that the force graphs are gathered in only one group per chamber, and there is no equivalent distance difference in votes between Members of Congress. An example of a bill in this quadrant is the one titled "Establish benefits for Health Sector personnel"\footnote{http://datos.bcn.cl/recurso/cl/proyecto-de-ley/4545-11}.

It should be noted that for the analysis exercise, some data that did not fit with the designed tools were excluded. Examples of this are abstention-type voting, match (abstentions by pairs), non-voters due to absence, and others. However, it should be mentioned that these data do not represent a data volume greater than 2% of the total, therefore its weight is considered diluted for the experiment.

The implementation of the algorithms for the calculation of indices is available in R language, at the following repository https://github.com/fcifuentes-silva/voting-paper-r-sources.

5. Discussion

Based on our method, the alignment graph in figure 6 shows that the Chamber of Deputies has a less disciplined behavior in voting compared to the Senate, since the trend in the distribution of the latter chamber shows a much larger bias towards 1 (fully aligned). This could be explained by various variables, such as the average age of the Members of Congress, political experience, etc. Regarding polarization, the data distribution graph shows that although the behaviour is similar in both houses, the Senate has a slightly less polarized behaviour than the House of Representatives, since although in the analysed group the Senate has less voting, shows a higher bias towards zero polarity than the Chamber of Deputies.
Regarding the analysis of bills in the context of the quadrants, the tool parsimoniously fulfils the function of characterizing each bill according to how it has been voted. Although a similar number of projects were randomly and manually analysed (without the use of automatic text analysis) to identify a profile and conceptualize each of the four categories, it should be mentioned that in this aspect the analysis is qualitative based on inductive reasoning. However, it is considered valid to indicate that the tool can be useful for political actors, trying to predict the possible scenario that certain bills will face, with the idea of seeking strategies in advance to obtain the approval of quorums.

In the same vein, it can also be useful for the development of artificial intelligence systems associated with making political decisions, where it is necessary to incorporate weighting factors for decision-making based on historical data or associated with specific issues, or be applied to make optimizations to the legislative process, where those initiatives that will be approved more easily are identified to conduct their processing in a simplified way, and giving priority in discussion to those projects that generate greater polarization.

Notwithstanding this, by way of triangulation, the analysis agrees with other studies carried out, where the way in which legislators vote on bills has been analyzed:

- For example, in the US, when legislators vote on issues on which they do not have information [29], their decision is affected by the opinion of their voters. However, in other cases, the opinion may be influenced by interest groups, party leaders and their own preferences. This is similar to the categorization described above.
- Another study [30] suggests that congressmen can vote according to one of three motivational axes, within which are self-interest, exchange of favors and ideology. However, it is mentioned that a vote eventually indicates a direction or preference but not a vote intensity.
- An alternative perspective to this scenario is shown in another analysis [26], where the problem that arises when analysing votes is presented when the data used is lacking in context. A scenario is presented where characteristics of the legislative work are erroneously inferred, as a result of the fact that only the roll call votes are rescued, but not those transmitted orally or that are partial, for which evidence associated with selection biases. Cases are presented about parliaments where all votes are registered, such as the US Congress, or in others where registration is on request, such as the European Parliament. A similar view is presented at [31] and at [32].

In any case, transparency in legislative votes affects the behaviour of the votes, allowing a greater citizen audit, and at the same time that the parties suffer fewer deviations compared to the case of not having public data [33].

Now, although it is possible to make other analyses in detail, such as by type of vote, using gender, age, for each vote in a particular way or others, it was left out of this article by extension, since the idea is to present the dataset and a tool as a plausible case study.

Other analysis, such as identifying the specific parts of a norm that show greater differences based on their votes (in a project there may be few polarizing or aligned votes associated with specific articles), can be difficult in the current scenario, due to the absence of detailed descriptors in the data associated with each vote in open data format. While this information is available for download in PDF documents on the cameras’ websites, obtaining, processing and publishing that part of the data is future work. However, it is considered valid to carry out the analysis at bill level, where there is both a descriptive title and the initiative.

We consider that the potential for analysis provided by this dataset is high, considering that it maintains a relatively constant growth. In addition the sets that coexist and interrelate are varied (and expanding) and they belong to a reliable and persistent source over time.

6. Conclusions

As has been shown, the use of Semantic Web Technologies to publish open legislative data provides high standards for the use of public data in political science and research, which can improve the development and impact of studies that are integrative and multidisciplinary. The importance of having high-quality data, which is persistent over time, from a reliable and available source, which in turn allows replicating or repeating experiments, is today one of the cornerstones in science and the accountability of government authorities. This article aims to contribute along these lines,
where a set of legislative voting data is available with everything necessary to be reused and combined. It can also be seen as an example to develop new data sets, being exposed to external quality checks of data.

At the same time, this type of practice provides citizens with more transparent and reliable public services. Considering that the legislative branch constantly carries a deteriorated image in the eyes of citizens [34][2], this type of initiative contributes to improve the perception of trust in activities that are so relevant to society, but not so well valued, such as politics. Indeed, publishing open data in this field can reduce corruption by increasing accountability and strengthening democracy, allowing voters to make better informed decisions [35].

From another angle, the design and implementation of the voting data set, creates a precedent in the digital offer of information that a public body, such as a parliamentary library, can make available to the community. This dataset, together with the National budget [7], Legal norms [5], Session debates [6] and others, in addition to the corresponding ontologies and data models, can be used and reused in other contexts, even in countries with a similar legislative system. In that sense, it should be mentioned that although the Chilean National Library of Congress provides a reliable and stable source of open data, there is awareness regarding the low availability and persistence of some datasets and ontologies.

Regarding the use of roll-call votes for studies, it is noted that although it is an attractive and growing field of studies, it must always be considered taking into account the restrictions and local circumstances of each political context, legislative and cultural environment. This is because since the institutional provisions restrict roll-call votes in various ways or make them public only under particular circumstances, we need to worry about possible selection biases [36].

In this work, we present the most relevant parts related to a Linked Open Data implementation [16] such as the use of dereferenceable URIs for identifiers, the usage of a Linked Data frontend to access RDF resources, the build focusing on reuse of existing vocabularies and ontologies, the usage of multilingualism as native RDF feature, linking with other sources such as Wikidata, and the usage of Shape expressions for validation and modeling.

References


20https://ourworldindata.org/corruption# which-institutions-do-people-perceive-as-most-corrupt

https://ourworldindata.org/corruption


