Transforming parliamentary libraries: Enhancing processes delivering new services with Artificial Intelligence

Francisco Cifuentes-Silva, Hernán Astudillo & Jose Emilio Labra-Gayo

Abstract

The integration of Artificial Intelligence (AI) in libraries offers wide impact on the evolution of information access and management. It allows both streamlining internal processes and transforming the way users interact with information resources, thus enhancing effectiveness and operational efficiency while enriching the user experience. This article presents the experience in incorporating several AI techniques in Chile's Library of Congress (BCN), and describes three initiatives: 1) publishing legislation as linked open data with Semantic Web technologies, combining machine-readable comprehension to high standards of interoperability; 2) maintaining the history of legislation, via automatic tagging of legislative documentation with natural language processing; and 3) predicting law approval based on current political context, using machine learning. The use of these technologies has allowed BCN to offer a wide variety of knowledge management services, providing useful and timely information for parliamentary work, and automated human-based repetitive tasks for efficient use of public resources.

Keywords

Semantic Web, Linked Data, Natural Language Processing, Legislation, Artificial intelligence, e-Government

Introduction

Artificial Intelligence (AI) is becoming widely used to facilitate automation and support in various human tasks, and its integration in libraries is having significant impact on the evolution of information access and management. Techniques that are increasingly frequent include Natural Language Processing (NLP), intelligent search engines or multimodal discoverers, and the use of physical or virtual robots. These uses allow to streamline internal processes, and also transform the way users interact with informational resources, thus enhancing operational efficiency and effectiveness while enriching the user experience.

In more legislative-specific use cases, AI tools are being used to process session video records, analyze legislative documents, build citizen engagement tools, and

support legal and parliamentary advice, among others Chalkidis et al. (2017); Fitsilis (2021); Gagnon and Azzi (2022); Elsawy and Shehata (2023); Oksanen et al. (2019); Klapwijk et al. (2021); Ben-Porat and Lehman-Wilzig (2020).

In line with this global trend, the Congress Library of Chile (BCN[1]), a parliamentary library with significant citizen engagement, is committed to maintaining the highest standards of quality and efficiency in generation and use of a wide range of products and services. Technological tools have played a crucial role, improving productivity processes and enabling a continuous expansion of services for both Congress members and staff and the public at large.

This article describes three altogether different AI application in BCN, each using different technologies: (1) publication of linked open data on legislation, using semantic web technologies that combine machine-readable comprehension (vocabularies, ontologies, and Shape expressions) with high standards of computational expressivity and interoperability, delivering over 52 million RDF triplets for public use; (2) creation of "History of the Law" and "Parliamentary Labour", using automatic tagging of legislative documentation using NLP, ML, Linked Data / Linked Open Data (LOD), and the Akoma-Ntoso XML standard (AKN), reducing document processing times by 37.5%; and (3) implementation of an approval predictor for individual laws, using a logistic regression model that takes information from the current political context, achieving an 86% accuracy rate.

Developing these projects with AI-based technology has allowed BCN to offer a wide variety of knowledge management services, providing useful and timely information to the Parliamentary Community, Legal Community, and Citizens at large.

Towards a Semantic Web based Library

The Semantic Web

The Semantic Web emerged around the year 2000, based on the idea of extending the link between two existing pages on the (non-semantic) Web, which lacks a meaning other than a simple connection (in HTML, the <a href...> tag), to one where the links have a specific, machine-readable meaning. This would make it possible to create intelligent agents capable of navigating, understanding, and solving human problems by exploring the Web. To achieve this, the Semantic Web proposes a technological stack based on two fundamental elements derived from the traditional Web: the first is the concept of URI, or Uniform Resource Identifier, which is basically a resource identifier from an abstract point of view (like an ID or passport number for a person) and is slightly different from URLs that locate documents. This means that a single resource can have one URI but several URLs, depending on the representations of the resource. For example, if an article is published at a URI, accessing it can provide a *machine-readable* representation for an application, while a human user might obtain a readable copy in their language. The second element is the Resource Description Framework or RDF, which is a declarative format that allows describing a resource

[1] Biblioteca del Congreso Nacional (BCN): https://www.bcn.cl

defined by a URI through attributes and relationships with other resources and can be implemented in multiple syntaxes, such as JSON, XML, or CSV. The idea behind URIs and RDF is that both data and data models (vocabularies, taxonomies, and ontologies) can be described in the same format and under an interoperability model based on HTTP.

With this aspect resolved, in 2006, the creator of the Web, Tim Berners-Lee, proposed the Web of Data and LOD Berners-Lee (2006), promoting the international adoption of data publication standards using Semantic Web technologies at government level. The LOD Cloud project[2] has grown from 12 open linked datasets in 2007 to 1314 datasets in 2023.

Motivation in the library

The availability of linked open data has multiple justifications in a public sector library:

- Public data facilitate studies and research.
- Open systems facilitate external contributions.
- Public data belongs to citizens, as they are financed through taxes.
- Public data builds trust by promoting information transparency.

The first Semantic Web project in BCN aimed to publish the database of Chilean legislation as Linked Open Data (where *legislation* is the set of rules, laws, decrees, etc that define the legal order). Most nations have mechanisms for publication of legislation: Official Gazette, Official Bulletin, Government Publishing Office, and so on. In Chile, the *legal fiction of knowledge* was sadly true, because the Official Gazette is not freely accessible; approved laws mainly amend previous texts rather that providing texts in force; and accessing older legislation is very challenging.

To address this, BCN launched in 2008 the "Ley Chile" (Chile Law) website, containing full text of legal norms; their versions, amendments, and related information, like related bills and jurisprudence; a search engine; and several tools for use and interoperability. While the primary objective of this website was to address the *legal certainty* issue, it serves current texts to Congress and citizens at large.

In 2024, it averages 80,000 daily visits, a high number given its specialized content and the Chilean population of about 17 million internet users. Well aware of the public value of this database, in 2011 BCN provided access as LOD for Ley Chile, releasing the datos.bcn.cl portal Cifuentes-Silva et al. (2011).

Semantic Web infrastructure

The datos.bcn.cl portal serves linked open data generated by Congress, belonging to several collections and originating from diverse sources. To achieve this, it implements a computing infrastructure with several elements, as follows.

[2] https://lod-cloud.net/#about

RDF Triple Store A key aspect of a technological infrastructure for the Semantic Web is a database engine that not only allows for proper storage and management of data but also enables querying for intra-organization use and public availability. In this scenario, the appropriate component is an RDF Store Hertel et al. (2008), which can be implemented either as a native RDF database or as a relational database with RDF and SPARQL mapping components. This component has three desired characteristics: firstly, it should provide a query interface in the SPARQL language, the de facto standard query language in the Semantic Web; secondly, it should have the ability to store information in structures called graphs (represented by a URI), which allow grouping RDF triples similar to how a table does in a relational database. Strictly speaking, a data tuple or quad is composed of four elements: the three parts of an RDF triple (subject, predicate, object) plus the graph it belongs to. The third desired characteristic of the database is the ability to execute federated queries by accessing other public databases through what is defined in the SPARQL 1.1 Protocol World Wide Web Consortium (2013). For BCN's case, the implementation of this part of the technological stack utilizes the Openlink Virtuoso RDF database.

Linked Data Frontend A Linked Data Frontend (LDF) is an essential tool that enables the HTTP publication of all URIs existing in the RDF Triplestore. The main idea is that the tool receives HTTP requests to URIs defined in a mapping file (commonly a regular expression associated with a SPARQL or SQL query). At that moment, it connects to the RDF database and retrieves all RDF triples corresponding to the resource associated with the URI. Then, through content negotiation (HTTP 303 code), it delivers the data representation that best fits the HTTP request (one of the RDF syntaxes, which can vary between HTML, RDF+XML, RDF+JSON, and many others). For the datos.bcn.cl project, we implemented the WESO-DESH tool, an open-source LDF written in Java that integrates RDFa into its HTML view.

Data models For the publication of open data, it is essential to describe the model that shapes the data in such a way that they can be understood. From the technological stack of the Semantic Web, tools are provided that allow expressiveness at different levels:

Vocabularies and Ontologies
An ontology is a formal specification of a representational vocabulary for a shared domain, encompassing classes, relations, and other objects. In short terms, it is an explicit specification of a conceptualization Gruber (1993). In the field of Semantic Web, ontologies enable the description of the semantic aspects of a data model through technologies such as RDF[3], RDF Schema[4], and OWL[5], facilitating the specification for sharing and reusing data.

For implementing ontologies and vocabularies for our own data, it's ideal to reuse other schemas (vocabularies and ontologies) that define properties or classes

[3] https://www.w3.org/RDF/

[4] https://www.w3.org/TR/rdf-schema/

[5] https://www.w3.org/TR/2004/REC-owl-ref-20040210/

semantically equivalent to our domain model. This is mainly because there are multiple general-purpose and specific-purpose vocabularies that have been around for a long time and are well-known within the community, making their use for describing data models natural and easy to understand. If the elements of the domain model we need to model haven't been modeled before, or if we define a specific use case, it will be necessary to design an ontology that allows us to model the problem domain. To design and be able to reuse the data, there are several ontology matching techniques available that enable reuse Kalfoglou and Shorlemmer (2003); Giunchiglia et al. (2005); Jean-Mary et al. (2009). It's also important to consider the principles of *reuse, don't reinvent* and *mix freely* Bizer et al. (2008) . At a technical level, it's ideal to use OWL World Wide Web Consortium (2012), RDFS Guha and Brickley (2014), or a combination of both. RDFS allows for the expression of classes, properties, hierarchies, domain and range of properties, and other elements such as defining sequences. On the other hand, OWL enables the description of higher levels of expressiveness such as transitive relations, set operations, negations, quantifiers, cardinalities, and other property attributes.

RDF data shapes It is another aspect 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. Shape Expressions (or simply ShEx) Prud'hommeaux et al. (2014) and Shapes Constraint Language (SHACL) Knublauch et al. (2017) are the most widely accepted proposals to define and validate RDF graph's topology, and although SHACL has become a W3C recommendation, ShEx is being used in many different scenarios Solbrig et al. (2017); Labra-Gayo et al. (2014); Thuluva et al. (2018); Cifuentes-Silva et al. (2020); García-Gonzalez et al. (2020) due to its concise and human-readable syntax, and an increasing set of open source and community tools that are currently being developed. Shapes, Vocabularies, and Ontologies must be described based on an HTTP URIs, which can later be treated in abbreviated syntax as a namespace or prefix.

Documentation portal

A documentation portal provides indispensable functions for the datasets published by an organization, including:

- Data Description: It allows for detailed descriptions of the ontologies, vocabularies, and Shapes that structure the data, which is essential for correct interpretation. It also enables describing the dataset's content, purpose, origin, data lifecycle elements (such as creation and modification dates), and the source or provenance of the data, helping users understand the value and potential applications of the data. Additionally, it is useful to describe the URI patterns of the exposed RDF graph, providing an even greater level of understanding of the dataset.
- Interoperability Mechanisms: It allows for describing the technical specifications on which the data is published, including data formats, URIs, and

publication standards. This encompasses the existence and access methods of the SPARQL endpoint or APIs, available content negotiation mechanisms, and the presence or absence of data dumps associated with the various published datasets and periods.

- Promote the Use of Open Data: This can be achieved by publishing manuals or user guides that explain how to query and use the data, either through SPARQL examples or direct access via the URIs where the data is exposed.
- Provide a Communication Channel with the Community: It allows for establishing a link between the organization and the data user community, where channels such as contact forms, forums, or mailing lists related to the data can be set up. At the same time, it enables the publication of data usage terms and news about updates to the available data.

Semantic Web in use

The first set of linked open data published by BCN was the datasets of Chilean laws in 2011 Cifuentes-Silva et al. (2011). The main idea behind this implementation was to conduct a proof of concept to validate the technology and explore potential uses. The work involved designing an ontology of laws associated with the Chilean model and then generating RDF triples for all the existing laws in the Ley Chile database. Additionally, an update tool was installed to keep the data up-to-date as new laws were published. This implementation included an RDF graph that allowed linking and navigating laws across their various versions, relating them to the different organizations that generate them, and, in the case of international treaties, linking them to their respective countries (connecting these to their URIs in DBpedia and Wikidata), among other things. All this was published in URIs that were not only unique but also readable, opting for the use of hierarchical URIs. Although only the basic metadata of the laws (such as title, date, creating organization, modification date, type of law, link to the XML text, and others) were published at that time, the experience proved useful for the interoperability and availability of the data to other state agencies and the public, as well as for conducting proof-of-concept tests with SPARQL and considering the possibility of mixing this data with other data. In this first project, 300,000 laws were published in RDF, equivalent to 8 million RDF triples. As of June 2024, there are 748,000 laws in RDF and 13.7 million RDF triples.

Then, we defined an ontology[6] of legal norms and a namespace prefix for the ontology, which is related with the particular context of the national reality. We considered a structure extensible to others domains such as parliament, education, health and others. This ontology has been written using both RDF Schema and OWL, making possible the application of inferences to RDF graph. Another important feature of the ontology, is that it has been composed using previous ontologies and datasets

[6] http://datos.bcn.cl/ontologies/bcn-norms/doc/

such as SKOS Bechhofer and Miles (2009), Dublin Core Baker (2000), FOAF Brickley and Miller (2007), Geonames[7], Organization[8] and DBPedia[9].

By utilizing the latter two, we were able to link data from the legal norm graph to external record sets, specifically international treaties and countries. This task was challenging due to the significant amount of manual labor required. Finally, the ontology was stored in the RDF store to enable inferences, as already published on the web using RDF/XML and Turtle syntax text files. The documentation was published in both Spanish and English.

During the same year 2011, BCN began developing its first projects based on linked open data. These projects aimed to facilitate interoperability within BCN's systems and enable citizens to reuse data and access all products and services based on unique identifiers (URIs). Two of these projects were strategic for BCN's definitive implementation as a Semantic Web Based Library: The *History of the Law* (HL) system, which aimed to collect, process, and publish the collection of all the documents generated during a law's legislative processing; and the *Parliamentary Labor* (PL) system, which aims to compile all legislative activity carried out by a parliamentarian during the exercise of his office, such that it has been registered in printed media belonging to the legislative power, such as a parliamentary motion, a session journal or a commission report. For the implementation of both systems, a diverse set of ontologies was implemented, and several datasets were published, including:

Parliamentary biographies: This dataset is composed by two main elements: an ontology (Biographies ontology[10]) that provides a model of classes and properties in RDFS and OWL, which is defined based on the FOAF, Dublin Core, and Time ontologies, and allows for the description of people, political parties, and other related concepts such as events (birth, death) and political positions (senator, deputy, president), among others; and a second element, the data, published as Linked Open Data in RDF, which provides basic information about each person, their periods of membership to political parties and parliamentary positions held since 1810. By June 2024, the database contains 5,296 people related to the political history of the country. Figure 1 displays, on the left side, a portion of the biographies ontology model associated with the person entity. On the right side, it shows the data of Gabriel Boric, Chile's President, in an RDF representation (HTML + RDFa).

The first data load 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, contained RDFa metadata that was extracted and transformed into RDF. This was accomplished using a model of dereferenceable URIs, allowing seamless navigation through various types of resources. Although a large amount of data was normalized during this process, due to the fact that the Wiki did not have validation mechanism

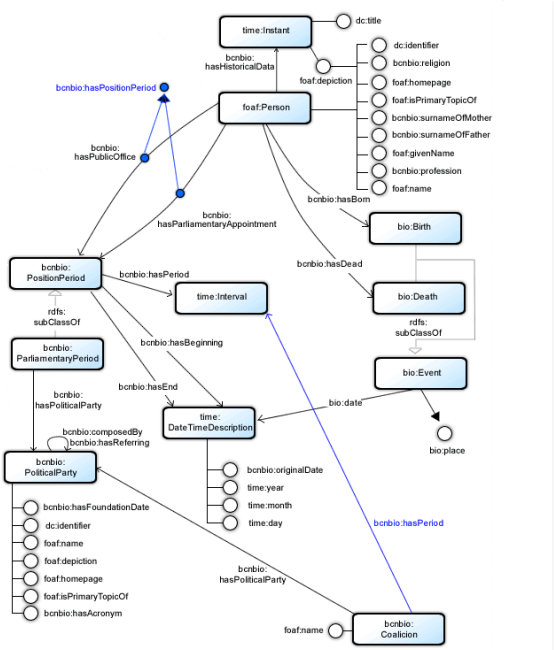
[7] http://www.geonames.org/ontology/documentation.html

[8] http://www.epimorphics.com/public/vocabulary/org.html

[9] http://wiki.dbpedia.org/Ontology

[10] https://datos.bcn.cl/ontologies/bcn-biographies/doc/

of the inputs, there have been minor errors related to formats and some inconsistencies in the information such



- http://datos.bcn.cl/recurso/persona/4536
  - foaf:depiction = https://www.bcn.cl/laborparlamentaria/imagen/4536.jpg
  - foaf:givenName = "Gabriel"^^xsd:string
  - rdfs:label = "Gabriel Boric Font"^^xsd:string
  - bcnbio:surnameOfFather = "Boric"^^xsd:string
  - bcnbio:surnameOfMother = "Font"^^xsd:string
  - bcnbio:nationality = http://datos.bcn.cl/recurso/pais/chile
  - foaf:isPrimaryTopicOf = http://www.gabrielboric.cl/
  - rdf:type = frbr:ResponsibleEntity
  - rdf:type = https://www.wikidata.org/wiki/Q5
  - rdf:type = foaf:Person
  - bcnbio:bcnPage = https://www.bcn.cl/historiapolitica/resenas_parlamentarias/wiki/Gabriel_Boric_Font
  - bcnbio:hasMilitancy = http://datos.bcn.cl/recurso/persona/4536/militancia/11056
  - bcnbio:hasMilitancy = http://datos.bcn.cl/recurso/persona/4536/militancia/11055
  - bcnbio:hasMilitancy = http://datos.bcn.cl/recurso/persona/4536/militancia/11258
  - skos:prefLabel = "Gabriel Boric Font"^^xsd:string
  - foaf:thumbnail = https://www.bcn.cl/laborparlamentaria/imagen/110x110/4536.jpg
  - bcnbio:hasPositionPeriod = http://datos.bcn.cl/recurso/persona/4536/cargo/110583
  - bcnbio:hasPositionPeriod = http://datos.bcn.cl/recurso/persona/4536/cargo/10478
  - bcnbio:hasPositionPeriod = http://datos.bcn.cl/recurso/persona/4536/cargo/10316
  - bcnbio:twitterAccount = "gabrielboric"^^xsd:string
  - dc:identifier = "4536"^^xsd:integer
  - bcnbio:idCamaraDeDiputados = "972"^^xsd:string
  - foaf:gender = "hombre"^^xsd:string
  - foaf:name = "Gabriel Boric Font"^^xsd:string
  - http://www.wikidata.org/entity/P2002 = "gabrielboric"^^xsd:string
  - wikidata-prop:P21 = "hombre"^^xsd:string
  - http://www.wikidata.org/entity/P856 = http://www.gabrielboric.cl/
  - foaf:img = https://www.bcn.cl/laborparlamentaria/imagen/4536.jpg
  - bcnbio:hasParliamentaryAppointment = http://datos.bcn.cl/recurso/persona/4536/cargo/10316
  - bcnbio:hasParliamentaryAppointment = http://datos.bcn.cl/recurso/persona/4536/cargo/10478
  - bcnbio:hasBorn = http://datos.bcn.cl/recurso/persona/4536/nacimiento
  - bcnbio:lastUpdate = "2018-11-29T21:44:31Z"^^xsd:dateTime
  - bcnbio:profession = "Egresado de Derecho"^^xsd:string

Figure 1. Parliamentary biographies dataset - **Top**: part of ontology model, **Bottom**: person in RDF

as duplicate periods of militancy or dates in different formats that have progressively been corrected manually. Currently, the data editing process is carried out directly on the RDF Triplestore, and most of the databases of the various systems in BCN use the parliamentarians' URIs for interoperability.

Geography: The geography ontology[11] provides a model of classes and properties in RDFS and OWL that describes geographic entities from the different models of territorial division existing in the Chilean political administration. In this way, it defines both the distribution by regions, provinces, and communes (administrative), as well as the division by districts and constituencies (electoral). This model allows, among other things, to solve the problem of the different territorial distributions of a country over time, for example when a territorial division of the region type is divided into two; as well as defining the different electoral divisions, which vary according to how the electors are distributed in the country over time. Currently, the linked open dataset published associated with this domain model covers the various historical geographical distributions of the country and their relationships (country, region, province, commune), as well as the current electoral distributions (electoral district and senatorial constituency).

Bills and Legislative resources: This dataset comprises the data of bills as well as the ontology models of *Congress* and *Legislative resources*, which have been presented in detail in a previous work Cifuentes-Silva et al. (2023, 2024). A bill is a document submitted to the National Congress to propose a legal text, to be debated by Congress with the aim of creating a new law. In Chile, a bill can be introduced either by the executive branch (referred to as a Presidential Message) or by a member of Congress (known as a Parliamentary Motion). The dataset of bills contains information such as basic metadata (title, bill number, submission date, status, URL to the text), access to its processing in the Senate's tracking system, details of each stage it has undergone, and the votes associated with the bill. For bills that have been enacted into law, it also includes a link to the law in the Ley Chile system.

The legislative resources ontology[12] and Congress model[13], both implemented in RDFS and OWL models a wide variety of concepts and relationships associated with the work of the National Congress, including the definition of bills, votes, document structures, elements of the legislative process, and many other abstractions necessary for automating tasks associated with the law-making process. These ontologies are heavily used in several internal applications due to their breadth, and indeed, both serve as the foundation for the metadata of XML documents processed in the open

[11] https://datos.bcn.cl/ontologies/bcn-geographics/doc/
[12] https://datos.bcn.cl/ontologies/bcn-resources/doc/
[13] https://datos.bcn.cl/ontologies/bcn-congress/doc/

standard for legal documents, AKN. Finally, the legislative resources ontology is also modeled using a Shape Expressions[14] schema developed in ShEx.

National budget: As presented in Cifuentes-Silva et al. (2020), the National budget dataset is based on an ontology[15] that represents the Chilean budget law published annually, the relationships between different state agencies that receive and allocate budget amounts to dependent agencies, and the budget execution of each of them. Among the main vocabularies and ontologies used for building this model are SKOS, RDFS, Dublin Core, OWL and the core vocabulary. Also, for this data model Shape Expressions are implemented. In structural terms, the budget law has a hierarchical composition in seven main levels, what is reflected in the ontology. The first level explains in an aggregated form the national budget, the next three coincide largely with state agencies that use public funds, and the next three describe the internal composition of budget within an organization. Each of the elements presented in each level has an associated number defined in the Law, for instance *Chapter 16* for the Ministry of Health.

In this way, the data model is composed by two main types of RDF classes: *metamodel classes*, such as Budget Execution, Budgetary Entity and Budget Entity, which model common behavior for domain business classes and are useful for describing domains and ranges in RDF properties; and *domain business classes*, such as National Budget, Batch, Chapter, Program, Subtitle, Item and Allocation, which can give meaning to data, and to establish constraints such as the usage of *owl:oneOf* declaration.

Regarding the data, the content of the dataset is monthly updated according to budget execution, and can be consumed via the SPARQL endpoint, by content negotiation in other formats or through the monthly dumps available on the website API[16].

Other datasets and models: Since the publication of the first dataset available as linked open data in BCN, several models have been implemented and used transiently or partially in different projects. Some of them are:

- *Communal Reports Ontology*: classes and properties (in Turtle and RDF/XML) that describe indicators from official sources on demographic, social, educational, economic, municipal, and citizen security data of each commune in Chile; this ontology is described in both .
- *Transparency Ontology*: a classes and properties (in RDFS and OWL) to depict hierarchies and labor relations among functional units and officials within BCN; it was used for RDFa markup in an earlier transparency website.

[14] https://datos.bcn.cl/ontologies/bcn-resources/bcn-resources.shex

[15] https://www.bcn.cl/presupuesto/linked-open-data/vocabulario.rdf

[16] https://www.bcn.cl/presupuesto/api

Results

BCN has established itself in Latin America as a library built on an open technology stack, where Semantic Web technologies play a crucial role both in publication and availability of open data for free use, as well as in creation of products based on these technologies. As of June 2024, more than 52 million RDF triples have been published with a sustained growth rate, along with 9 ontologies and domain-specific vocabularies, production systems, and numerous technological tools that utilize the data and models.

Furthermore, over time this philosophy of providing data openly and freely for thirdparty use has become ingrained in the organizational culture, and many new projects use and produce data available as linked open data.

Limitations

Without prejudice to the foregoing, it is important to highlight some of the most significant limitations of this work:

- Although the complete set of legal norms of the state was published in RDF, the dataset contained only the basic metadata of the norms and a link to the XML version, but not the triples associated with their content or normative structure. These could be highly useful for implementing applications and performing analysis with graph-based tools. This omission was due to the fact that a single norm can generate thousands of triples by itself, not to mention updated versions of norms that are automatically composed from the text of amending norms or other linkages.
- Some initiatives associated with RDFa mentioned in the study are no longer available as web pages. This is due to the technological replacement of applications, coupled with a shift in focus from interoperability to usability.
- In the case of the biographies wiki, RDFa was initially used for populating and structuring person-related data. However, RDFa is no longer used to mark biographical entries. Instead, the RDF representation of person entities is now directly derived from the internal parliamentary database.
- Another significant limitation is that linking to external datasets, such as Wikidata, is performed manually on a resource-by-resource basis, without automated mechanisms to streamline the process.
- Finally, it is worth mentioning that a wide variety of datasets and web services not based on Semantic Web technologies were not included in this work.

Natural Language Processing of legislative documents

To discuss natural language processing, it is necessary to provide context about the projects that drove this technology at BCN.

History of the Law and Parliamentary Labor

At the beginning of 2012, the development of the projects History of the Law (HL) and Parliamentary Labour (LP) began, presented in detail in Cifuentes-Silva and Labra-Gayo (2019), which would expand the catalog of products and services that BCN offers to the public and the National Congress.

A *History of the Law* (HL) is the collection of all the documents generated during a law's legislative processing; since the initiative that gives life to the bill, until its discussion in the Congress, the reports of the parliamentary committees that studied it and the transcripts of the debates in the session's rooms, gathering their traceability within the legislative process. The HL allows someone to collect the so-called *spirit of the Law*, allowing its interpretation in a precise way in relation to the scope and sense that was given to the norm when it was legislated. This legal instrument is particularly useful both for judges when preparing judgements and for lawyers when they use certain rules to support their arguments.

Similarly, the *Parliamentary Labour* (LP) is a compilation of all the legislative activity carried out by a parliamentarian during the exercise of his office, such that it has been registered in printed media belonging to the legislative power, such as a parliamentary motion, a session journal or a commission report.

In Chile until 2011, both products were made by legal analysts, only for specific requests, and by processing manually each document related to a Law. The objective of the HL and PL project was the establishment of a technological infrastructure support and the necessary processes for the semi-automated elaboration of the History of the Law, covering all the laws of the republic, and its subsequent availability to the public, both through open data and via a public access web portal.

For the electronic and semi-automated elaboration of both documentary collections, it is required to have a granular database, which registers all the documents of the legislative process where any reference to bills or parliamentarians is made, allowing later to extract and recover selectively, what was discussed around a bill that will become law, as well as what a certain legislator has said in any context. For this reason, the OASIS standard for legal documents - AKN Palmirani and Vitali (2011) has been used for the construction of these project, since it provides electronic representations of parliamentary, normative and judicial documents in XML using semantic markup and annotation of textual documents through Linked Open Data (LOD), allows the addition of semantic marks on the text, which in turn allow the precise identification of the location of parliamentary interventions, the presence of debates around a bill, the processing phases, and many other types of metadata.

Datasets

To create HL and LP products, we used several datasets, as follows.

Legislative Process Documents: This dataset includes basic metadata schema (URI, title, date, originating chamber, document type, and more) and documents content in text (TXT) and XML (AKN version). There are several document types, coming from different formats. First, the legislative process has several types of documents, chiefly

the Session Log; it includes attendance records, a transcription of the floor discussion (parliamentarians' speeches), and often several documents presented during the session, like official letters, bill proposals (motions or messages), communications, and more. Another important document type is the Committee Report, which provides an extended summary of what occurred in the session of a committee; it has a structure similar to the Session Log, but focused on issues within remit of each committee.

Each document type, depending on its period and originating chamber, has different drafting rules and structures, which have evolved over time; as an extreme, older documents use an outdated Spanish orthography (Andres Bello's orthography)´ [17].

The documents format also varies according to their period. Documents before September 1973 (when Congress was dissolved) were only available in physical format, i.e. on paper; therefore, processing them requires digitization. Documents from 1990 onward are native digital, in formats like *.doc*, *.pdf*, or even *.xml*, thus allowing processing of the file itself.

This dataset includes all Session Logs, from both Congress chambers, from 1965 to 1973, and from 1990 on; and all Committe Reports, messages, motions, and official letters, as preserved for History of the Law (HL), for all laws from 1990 and many decrees issued between 1973 and 1990.

Named Entities: The second dataset has data on several named and interrelated entities. It is used for tagging and referencing operations within XML documents, configuring processes in the processing workflow, and supporting delivery and query functions. The most relevant entity sets are:

- People: enables the identification of parliamentarians and the creation of the LP. It includes metadata according to the biographies ontology, which covers basic information (full name, last names), known name, nationality, public positions held in Congress and the executive branch, birth and death events (including place and date), links to other resources such as social media, Wikidata URI, gender, thumbnail image URL, and political affiliations (including political parties and periods of membership).
- Organizations: includes state agencies involved in the legislative process, such as the chambers of the National Congress, ministries, and other entities like the governing junta (the ruling body from 1973-1989) or the Constitutional Convention of 2020.
- Political Parties: contains information on political parties, such as their description, logo, founding date, founders, and other relevant details.
- Other Datasets: datasets for sessions and legislatures, to allow time-aware identification and differentiation of session documents; legislative phases, defining constitutional and regulatory procedures for each law; and bills, as previously defined .

[17] https://en.wikipedia.org/wiki/Bello_orthography

Document Structure Elements: Each document type is defined based on a hierarchical structure of sections, composed of various instances of the "SeccionEstructural" class. These different data collections enable the dynamic configuration of an XML editor based on the type of document being edited, while also allowing other applications to navigate the document's structure according to the defined subparts.

Automatic markup of XML documents

Possibly one of the most tedious tasks, requiring most time from a human, is identifying and marking in a text both *structural sections* (such as chapters or subchapters) and *named entities*, and their association with specific identifiers. A more efficient option is implementing a component that automates this task, with humans only reviewing the marking done by the algorithm. We will call this tool the *automatic marker*, which transforms a plain text unformatted document into an XML document with AKN schema.

This problem has already been addressed using machine learning (ML) Akhtar et al. (2004); textual visual properties Burget (2007); content-associated patterns Bolioli et al. (2002); and combinations of rules Abolhassani et al. (2003). We used three ways:

1. Knowledge engineering: manual implementation of rules or ad-hoc algorithms; we used regular expressions and exact matches to detect structural elements and identify entities in the text; and quality is verified with a documents sample and a complete list of entity descriptors. The model precision and effectiveness depend on the quantity and quality of implemented rules.
2. Machine learning: pre-labeled documents divided into training and test sets; training documents are fed to classification algorithms that use pattern recognition (e.g. Hidden Markov Models, Naive Bayes, Conditional Random Fields, or Neural Networks) to yield a classification model; and quality is verified with metrics (like Precision, Accuracy, Recall, and F-Measure Baeza-Yates and Ribeiro-Neto (2011)) upon the test sets. The most commonly used marking components are entity recognizers (e.g. *Stanford NER*, *spaCy*, and *OpenNLP*).
3. Hybrid approach: combines both approaches: complex structural marking is executed with knowledge engineering, taking as input ML-recognized entities (restricted to key sections to improve marking efficiency and precision).

In History of the Law projects, the complexity and detail of marking legal documents with multiple tasks (i.e. detection and disambiguation of named entities, structural recognition of document parts, and specific formatting) required to implement a four-component pipeline:

Named-Entity Recognizer (NER): Finds mentions of nouns or "entities" in the text (e.g. dates or figures written in narrative prose); and identifies the entity type of each. We use a customized version of the Stanford NER software[18], implemented by a CRF classifier, which enriches the input text with recognized entities, possible entity types (see Table 1), and confidence scores. As a training corpus, we used session documents

[18] Stanford NER: https://nlp.stanford.edu/software/CRF-NER.shtml

manually labeled with 64,727 words, each with a type. To evaluate the deployed model, we used ten-fold cross-validation (90% training - 10% testing); the NER detected on average 97% of entities in the text, and correctly assigned their type in 89% of cases.

Mediator: Assigns the URI associated with an entity mentioned in the text, performing *Entity Linking* or *Disambiguation*, similarly to DBpedia Spotlight Mendes et al. (2011), AGDISTIS Usbeck et al. (2014), or WikiME Tsai and Roth (2016) .

The Mediator is based on a RDF Triplestore, accessible through a SPARQL endpoint that stores information ranging from basic descriptions (like names or text descriptors, dates, etc.) to more complex structures (like membership periods or event occurrences), all associated with entities of the types described in Table 1.

Table 1. Types recognized by NER and number of entities in the Knowledge Base

Named–Entity Type	Example	Total in LKB
Person	Gabriel Boric, Salvador Allende Gossens	5.296
Organization	Ministerio de Educación, SERNATUR	8.169
Location	Valparaíso, Santiago de Chile	1.251
Document	Ley 20.000, Diario de Sesión Nº12	861.518
Role	Senador, Diputado, Alcalde	434
Events	Nacimiento de Eduardo Frei Montalba, Sesión Nº 23	17.963
Bill	Boletín 11536–04, Prohíbe fumar en espacios cerrados	17.697

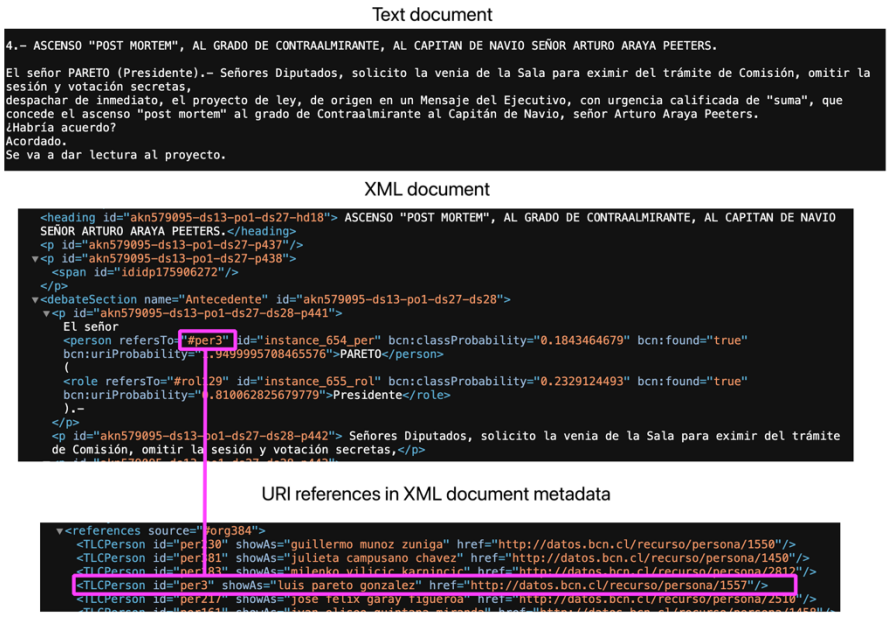


Figure 2. Text to XML markup with NLP

Prepared using sagej.cls

The simplest use case involves sending a label via REST as a parameter, with the mediator returning a list of suggestions in JSON format, structured as (uri, label, score) and ordered by decreasing score. To improve accuracy, parameters such as the entity type or a session ID can be added. Additionally, the mediator can operate by receiving an XML file with recognized entities as input. It will return an XML file where it adds URI and label attributes to each entity, assigning the URI with the highest calculated score for each case, provided this score meets or exceeds a configurable threshold.

To enhance the tool's precision, context data can be assigned to narrow down the set of possible alternatives when selecting the URI for each label to be identified. Since this tool is used to disambiguate entities in documents of the National Congress, useful context data may include the session date, the chamber of the document, the session number, or the legislative period. To evaluate and refine the implementation's accuracy, the first set of session journals from the 1965–1973 period was processed using the NER system, which assigned URIs to the entities identified in the text. These assignments were manually reviewed both individually and through an aggregated view that highlighted anomalies (e.g., persons identified who did not belong to the parliamentary period). This process enabled fine-tuning of the filtering algorithms associated with contextual information. As a result, a tool was developed that currently operates with an error rate of less than 1% in its productive use.

Structural marker: We define the structural detection of text as the task of identifying groups of consecutive strings that, to a human reader, correspond to elements such as titles, subtitles, paragraphs, sections (groups of paragraphs under the same title or subtitle), chapters, annexed documents integrated into the text, enumerations, and lists. Additionally, given the context of application in documents containing parliamentary debate, we define a special type of structural element called an Intervention (speech), which describes what is spoken by a person and can consist of one or more consecutive paragraphs.

The structural marker is the tool that performs structural detection by adding marks to the text to indicate the beginning and end of each structural element. In our case, the marks added to the processed text output are in XML format.

The primary strategy for detecting structural and hierarchical sections in text documents combines the use of regular expressions with the application of rules that encapsulate programming logic, triggered when specific regular expressions are detected. This approach is particularly practical for documents generated by the National Congress, as they generally adhere to standardized drafting rules. The tool can identify first, second, and third-level structural sections, sequences of elements based on numbered and unnumbered lists (including nested ones), and parliamentary speeches.

Given that the documents to be processed mainly consist of political debates, the primary element to be recognized is a block called *participation*. This block is composed of one or more speeches involving an actor who moderates the session (usually the President of the Chamber), an actor who owns the participation (the primary speaker), and potentially other actors who interrupt. In this composite block, it is essential to automatically identify the participation author, necessitating an

analysis of the speeches to detect the underlying structures of participation and implement a rule with automaton characteristics. This particular implementation is needed because participation, in structural terms, is unique as it lacks a conventional title and body structure and is embedded within other recognized structural sections of the debate.

Technically, the structural marker receives plain text with or without entities as input. From this text, a DocumentPart object is generated, containing a property with the full text and an empty list of DocumentPart objects (sub-parts). The process involves running specific rules depending on the document type and the depth level of the DocumentPart object. Each rule processes the text of a DocumentPart object and returns all identifiable DocumentPart objects, adding them to the sub-parts list. The final task is to serialize the object in XML format.

XML converter to AKN Schema Takes XML output of structural marking, entity recognition, and entity linking—referred, and yields an XML—into AKN format, editable by tools like LIME[19], Legis Pro[20], AT4AM[21], Bungeni[22], xmLegesEditor[23], or LEOS[24].

To convert raw XML into AKN, we first tried XSLT, but the required style sheets to identify entity references and group them headers (essencial to AKN) were too complex. Given the diversity of document structures, we took a programmatic approach that takes the input XML file, generates a DOM-like representation, and traverses it to convert each raw node to AKN XML.

Other NLP-based tools There is a product derived from the History of Law called ”History of Law by Article” (HLA), which exclusively documents everything that has occurred to a specific article or other normative unit during the law’s legislative process. To create this product, various tools have been implemented to assist human analysts in their work. These tools operate together on two or more documents, allowing for crossreferencing between their sections, with the aim of subsequently compiling related information based on different criteria. Some of these tools include:

- *Link Generator:* Operates on two consecutive versions of a bill’s XML (consecutive in terms of modifications introduced during the legislative process) that have previously undergone automatic tagging. In the XML of each version, the basic normative units (such as articles, numerals, paragraphs, or enumerated letters) are identified. The tool’s function is to compare the normative units of both versions of the bill and establish text traceability for

[19] http://lime.cirsfid.unibo.it

[20] https://xcential.com/legispro-xml-tech/

[21] https://at4am.eu

[22] https://github.com/bungeni-org

[23] http://www.ittig.cnr.it/lab/xmlegeseditor

[24] https://ec.europa.eu/isa2/solutions/leos

each unit. It determines whether a unit has been affected by text modifications, added, removed, or moved.

- *Reference Generator*: Operates on an XML version of a bill and a document containing discussions, such as a session journal or committee report. Its function is to automatically identify all sections of the discussion document where references are made to a specific normative unit, with the aim of compiling the associated debate for that unit across all documents where the law is discussed.
- *Update Generator*: Operates on an XML version of a bill and an amendment document. An amendment is an instructive document that specifies changes to the normative units of a bill version, resulting in a new version of the bill once these changes are applied to the text. An amendment can include instructions such as "replace word X with Y," "remove point Z," or "add article X," among others.

To implement these tools, several techniques have been used, including XML tree traversal, text comparison using cosine distance, stopword removal, lemmatization, different types of tokenization, regular expressions, and rules through programming classes (see Gacitúa et al. (2016) for details). The evaluation mechanism for all these tools followed the same approach, based on a comparison between manually labeled documents and the versions produced by the automated tools.

Constitutional History

This use case started with the History of the Law project. In late 2019, Chile suffered a *social outbreak*[25], characterized by large-scale civil protests across the country, arising from dissatisfaction with the prevailing economic model. In an effort to quell the social unrest and address citizen demands through a structural legislative approach, most political parties in Congress settled on an "Agreement for Social Peace and the New Constitution"[26][27]. A referendum[28] decided to draft a new constitution (78.27% approval), initiating the so-called 2020 Constitutional Process. BCN decided to document this process, similarly to History of the Law and Parliamentary Labour, all the debate surrounding it, to capture the spirit and foundations of each part of the draft New Constitution and the work of each drafter. Since all the technological infrastructure was already in place, a new website was open with the history of Chile's

[25] https://en.wikipedia.org/wiki/2019-20_Chilean_protests

[26] https://obtienearchivo.bcn.cl/obtienearchivo?id=documentos/10221.1/

[27] /1/Acuerdo_por_la_Paz.pdf

[28] https://www.bcn.cl/procesoconstituyente/plebiscito2020

constitutions [29], as well as the history of constitutional processes that had not produced an approved draft.

History of the Law

Once the AKN files have been annotated both automatically and manually and have successfully passed the quality assurance phase, the publication process extracts the knowledge expressed in the document as RDF triples and tuples for a query data base. All data extracted from the document during its publication in RDF (including structural sections, entities, parliamentary speeches, references to bills, and more) are transformed into new RDF triples. These triples then complement the previously defined basic information in the RDF triplestore, making them available for querying at the SPARQL endpoint.

The publication process, packaged as a web service, implements a parser that traverses the XML tree of the AKN document, looking for predefined structures within the document sections. Each type of document uses small data extractors per section, encapsulated in specific classes, allowing for reuse and specific implementations as needed. Subsequently, with the data extracted from all XML documents in which a law is discussed (HL) or a parliamentarian speaks (LP), a publication process generates a physical file stored in a digital repository (DSpace[30]), indexed by library specialists.

BCN launched in 2014 the portals for the History of the Law[31] and Parliamentary Labor[32], and in 2020 the Constitution History portal. These platforms publish open linked open data.

Results

By June 2024, the HL system has processed 4,143 laws, associated with 49,030 documents, generating 520,219 participations from 1,578 parliamentarians from 1965. Additionally, it has documented the history of all Chilean constitutions and constitutional proposals, as well as the work of the two recent constitutional conventions (2020 and 2022).

The automated tagging has reduced work by 37,5% Cifuentes-Silva and Labra-Gayo (2019), without consider the impact on the constant generation, updating, and indexing of LP document dossiers for each parliamentarian.

Limitations

Some of the most relevant limitations of this work are as follows:

- Although session documents are processed in full text, relevant sections such as voting records are neither retrieved nor utilized. This is mainly because these records are directly consumed from web services provided by the chambers of Congress for integration into the database. However, future

[29] https://www.bcn.cl/historia-de-la-constitucion/

[30] https://www.dspace.org

[31] https://www.bcn.cl/historiadelaley

[32] https://www.bcn.cl/laborparlamentaria

plans include incorporating this information, particularly for widely agreed-upon votes that do not record nominal votes but are of interest in the legislative process.

- Initially, the project included a wide range of features associated with detailed markup tasks that were mostly intended to be performed by human analysts, with the idea of enabling the development of other products and services in the future. However, due to limitations in access to both legal analysis and IT development personnel, the system was consolidated around its core objective: building the legislative history and parliamentary activities.
- To date, the system has not integrated tools based on Large Language Models (LLMs), which could improve the accuracy of processes such as the Update Generator or the Reference Generator.
- Similarly, while the initial project envisioned linking metadata between intervention texts and session videos (using video subtitles based on the text of interventions), this task has not yet been addressed.
- Although XML tagging processes could automatically identify additional information in the texts of legislative debate documents (e.g., identification of subjects or free terms for indexing), this remains a pending task that could enable a wide range of new products and services.

Law approval prediction

The crafting of legislation is a complex process influenced by various factors. While legislative initiatives often arise from citizen needs, many of them lack the necessary support within the legislative process to advance through their processing and ultimately become law. This is partly due to the large number of projects in various subject areas, the time required for parliamentary debate in each case to gather diverse public perspectives, and the sensitivity of establishing legal norms that govern society. In any case, it is of great interest for both the executive branch and members of Congress to know the probability of whether a proposed bill will ultimately become law or not. This knowledge enables them to undertake the necessary political efforts to promote its processing and eventual enactment as law.

The third experience using AI corresponds to a law approval predictor for bills processed in the National Congress of Chile, based on a logistic regression model. A logistic regression model is a statistical model used to predict the probability of a binary event, meaning an event that has one of two possible outcomes. It mathematically models the relationship between a dependent variable (the variable to be predicted) and one or more independent or predictor variables. A logistic regression model, in the context of AI, is a supervised learning technique used to solve binary classification problems by predicting the probability that an instance belongs to one of the two possible classes using the logistic function.

For this use case, the specific utilization of this type of model is based on several factors, including its high interpretability, which is essential in an advisory environment for members of Congress across all political sectors. In such a setting, there must be no political preference or biases, making it crucial for predictive models

to be explainable and ideally simple. Additionally, the training data size is relatively small (fewer than 10,000 examples), making other models like deep learning less suitable.

Therefore, the following will present the design and results of a law approval predictor implemented for the Chilean legislative process.

Datasets

To achieve this, data from 14,738 bills introduced to Congress since 1990, along with their processing details, political parties, and the actors involved in the legislative process, were used. This data was obtained from the open data portals of the National Congress, supplemented with ad-hoc calculated exogenous variables. In detail, the development of the predictor utilized three sources of data:

1. Linked Open Data from BCN: as mentioned earlier, the open database of BCN that provides information about parliamentarians from 1990 to the present, including their terms in office, which includes the person's ID, start and end dates, and the position they have held. It also includes information about bills introduced from 1990 to the ends 2022, their authors, and their political parties.
2. Opendata Congress XML Database: An XML service database that provides records of the processing of bills, which is updated daily. This data is published by the Senate from the bill processing system.

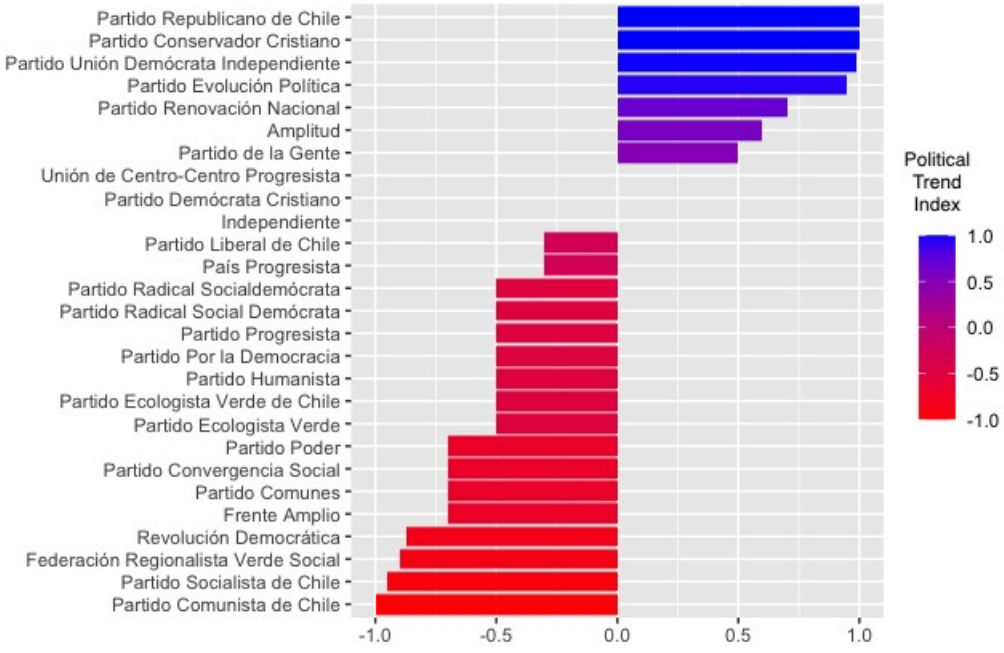


Figure 3. Political trend index of parties based on author's perception

3. Exogenous Database: A database created specifically for this occasion, containing election years (a list) and a table of political parties with a *Political Trend Index - PTI*, created based on the author's perception.

The *PTI* differentiates the political tendencies of parties on a scale from -1 to 1, where -1 represents a more left-wing ideology, while +1 corresponds to a more right-wing ideology, as shown in Figure 3. According to the authors' perception, the Communist Party of Chile would have an index of -1, while the Republican Party of Chile would have a value of +1.

The main idea behind this index is to convey the political trend value to parliamentarians affiliated with the party in various circumstances, whether as authors of a bill, as the President of the Republic (who holds executive power), or as the President of the Chamber of Deputies or the Senate, who have direct influence on the prioritization of bill discussions. Figure 4 shows the Political Trend Index (PTI) applied to the positions of President of the Republic, President of the Senate, and President of the Chamber of Deputies from 1990 to 2024. The idea behind this index is to have an indicator that differentiates and systematizes political support during the bill's processing period, which will be translated into a limited set of features for regression analysis.

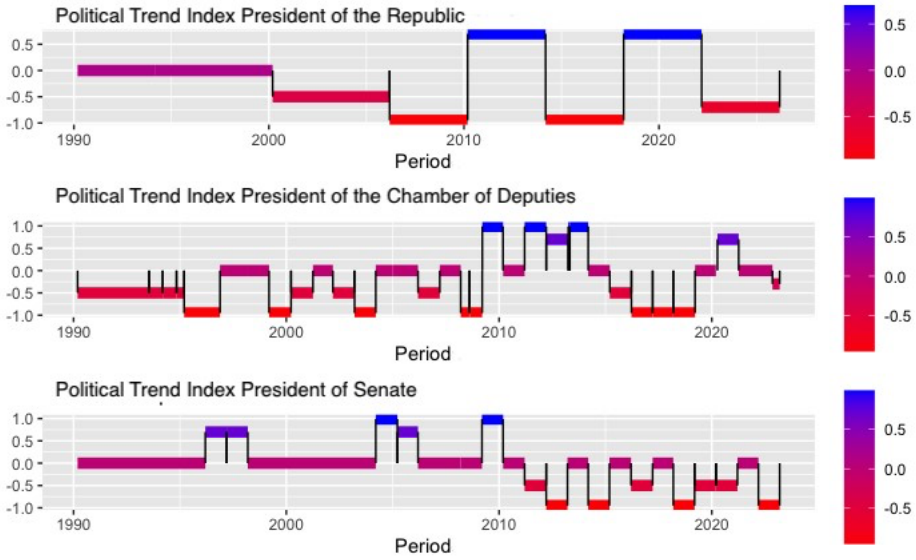


Figure 4. Political trend index of parties based on author's perception

Tool design

With this data, an initial database of bills was created with 23 features, including the current status of the bill (enacted as law, rejected, in process, archived), date-related characteristics (such as separating year and month for each date into different features), identification of events that could affect legislative processing, such as determining if a year is a presidential election year or not (assuming this could affect the processing of certain types of bills), the number of days since entry into processing or since the last movement in Congress, and 13 characteristics calculated based on political trend measured from various perspectives (sums, averages, etc.), among others.

The primary idea behind using these variables was to characterize the political landscape to determine if a particular bill could be approved, assuming that the actors involved in the process coincided in some way due to their political tendencies.

Thus, following the aforementioned principle, political trend values were applied to characteristics such as authors, the President of the Republic, the Presidency of the Senate, and the Presidency of the Chamber of Deputies at the time of the bill's entry into processing, as well as at the time of the last processing recorded in the Senate's processing system.

Having defined the first set of characteristics, three groups of projects were generated:

1. Bills that became law (Enacted): all those that were transformed into law and have critical information in the database (dates and status).
2. Bills that could still become law (In process): all those with critical information in the database and currently in the "In process" status.
3. Bills that did not become law (Rejected): all those not excluded in the first stage and not belonging to the Enacted or In process categories. This category includes Archived bills, Unconstitutional ones, and others.

Additionally, a fourth group was excluded from the exercise due to having too many missing or inconsistent data. The purpose of identifying these groups is to implement a set for training the logistic regression model (Enacted + Rejected), and another set to predict whether a bill will become law (In process) later. Table 2 provides totals by project type.

Bill status	Total
Enacted	3.282
In process	6.800
Rejected	4.953
Excluded	487

Table 2. Bills by status for training

Results

With the prepared data, a logistic regression model was trained in R language[33]. Initially, 23 preliminary features were considered for the model. During the regression analysis, the Akaike Information Criterion (AIC) technique was used for feature selection, transitioning from the initial model with 23 features to a simplified and optimized final model with 9 features, all of which were statistically significant (p-value < 0.001). Of the discarded features, more than 10 exhibited high correlation (multicollinearity), and were therefore removed primarily to avoid issues of accuracy and sensitivity in predictions with new data. Other discarded variables showed low statistical significance, and were therefore also removed. Table 3 provides a detailed description of the features for the final logistic regression model.

The final model shows a classification accuracy of 86.23%, with a 95% confidence interval of (0.8546, 0.8697), a sensitivity value of 0.89, and a specificity value of 0.82. Additionally, it is worth mentioning that the cutoff point for the regression is set at 0.38, which is the threshold at which the prediction changes from approval to rejection.

As a testing mechanism for the model's functionality and accuracy, simulations were conducted to mimic shifts in political trends among various actors associated with PTI-related characteristics. These changes directly influenced the calculated probability of a bill's approval, aligning with expected outcomes.

The Figure 5.B displays the confusion matrix of the predictor using the training data, and Figure 5.A presents a comparison between the actual cases and the values provided by the regression model for the training data.

Limitations

While the model demonstrates acceptable performance, it is not without limitations, which include:

- The absence of some key features, such as identifying whether a bill is designated as urgent legislation—a prerogative of the executive branch—or whether the bill pertains to a topic of current political debate.
- The use of an arbitrarily defined PTI by the authors, which may affect the accurate representation of the data and the results obtained. Future work in this area could include a dynamic calculation of real PTI values based on voting distances (greater differences among voters imply larger distances), wich is observed between different political parties, potentially segmenting the PTI by topic.

[33] https://www.r-project.org/

Law's Predictor prototype

Figure 6 shows the user interface of the first functional prototype of the predictor. It features a line graph where two zones are differentiated on the vertical axis: one for approval (green) and one for rejection (red). The horizontal axis shows the evolution of the prediction over time according to changes in the political context as well as information related to the bill. Additionally, a thermometer-type graph indicates the current probability of approval.

Feature	Description
Regression intercept	Float number, base for regression
Type of bill	1 if Parliamentary motion, 2 if Presidential Message
Total authors with PTI zero	Float number, total number with value Zero in PTI
Average PTI of authors	Float number, the average of the PTI of all authors
Average PTI of left-wing authors	Float number, the average of the PTI of all authors with PTI <0
Month of last processing in the Senate system	Number of month from 1 - 12
Are there elections in the year of the last processing in the Senate system?	1 if is election year, 0 if not
Total days processing	Integer number
PTI of President at submission	Float number, PTI of the President of the Republic at the time of bill submission
Current PTI of the President of the Senate	Float number

Table 3. Features for optimized final model

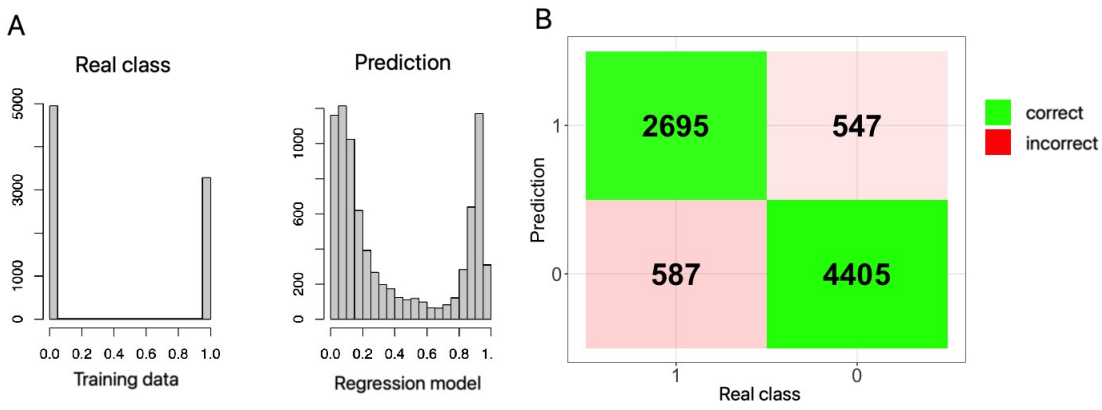


Figure 5. A: Comparison between real data and prediction - **B:**Confusion matrix for Law predictor

This tool still has a restricted access as we are still improving the validation and user interface.

Figure 6. Law's predictor user interface

Related work

Although this work explores various lines of development in artificial intelligence (AI) implemented in BCN, there are also similar experiences worldwide that can offer different approaches to enrich the discussion. Below, international cases analogous to each of the described experiences are presented.

In the field of parliamentary LOD and libraries, we can find the ParliamentSampo system Hyvönen et al (2022), a LOD service and semantic portal designed to investigate political culture and language in the Finnish Parliament. Based on a knowledge graph containing nearly one million parliamentary speeches (from 1907 to 2021) and data about Members of Parliament, it utilizes the Parla-CLARIN format for analysis and application development, aiming to enhance transparency in political decision-making for diverse user groups. The article provides an interesting analysis of the speeches but does not focus on the details of open data models, and while it mentions them, it does not delve into the process of generating open data.

The study by Koryzis et al. (2021) analyzes the digital transformation of parliaments today, highlighting trends toward open data, standardized processes, and participatory inclusion. Based on surveys conducted with parliamentarians and parliamentary professionals from 25 countries across five continents, it proposes a transformational framework that identifies essential tools for sharing information and knowledge. One of the most relevant findings related to our case study is the recognition of technological aspects with high utility and maturity associated with the use of linked open data, which are even deemed more significant than the use of machine learning.

The article by Niu (2020) examines the diffusion and adoption of LOD in libraries, revealing it as a decentralized, continuous process with significant reinvention. Through extensive analysis, it identifies three diffusion paths (inter-library, intra-library, and inter-librarian) and highlights the roles of leading institutions, professional organizations, vendors, and funders in facilitating adoption. The study underscores the importance of standards, tool reuse, and commercialization in overcoming barriers such as resource limitations and steep learning curves. By mapping the stages of LOD adoption, this research offers a comprehensive framework to guide libraries in embracing LOD for enhanced interoperability and knowledge sharing.

Also in the field of Semantic Web and LOD technologies, several advancements are reported in the literature, which report improvements both in access to the information provided to users, as well as improvements in the impact on the internal functioning of the institutions that implement it. Among these are the use of Shape Expressions (ShEx) for ensuring the quality of LOD (Candela et al., 2023), the replacement of MARC standards with BIBFRAME, which is LOD-based (Samples and Bigelow, 2020; Park et al., 2020), ontology-based systems for processing scientific information in digital libraries (Malakhov et al., 2023), the use of RDA and Wikibase (the technological backbone of Wikidata) for managing named entities (Zapounidou et al., 2024), and the preservation and access to cultural heritage using Semantic Web technologies (Silva and Terra, 2023).

Gagnon and Azzi (2022) introduce the concept of "Legislative Intelligence" (LegisIntel), which leverages AI and semantic analytics in parliamentary contexts to enhance citizen engagement through semantic annotation of debates. Using NLP technologies,

ontologies, and knowledge graphs, these tools enable customized indexing, entity recognition, and tailored recommendations for legislative critiques. Their study highlights an international open-source initiative aimed at developing LegisIntel solutions, detailing its core functionalities and proposing a robust architectural framework.

In the context of NLP applied to legislative tasks, it is worth highlighting several underexplored areas. Notable advancements include the use of language models for tasks such as automatic summarization of regulatory documents Klaus et al., (2022) and multilingual legislative frameworks Zmiycharov et al., (2024); Gesnounin et al., (2024). Additionally, Retrieval-Augmented Generation (RAG) models Lewis et al. (2020) have been employed in tasks such as legal question answering Louis et al. (2024), Wiratunga et al. (2024) and the development of virtual assistants tailored to the legislative domain Rafat (2024).

In the realm of legislative prediction tools or similar applications, several notable experiences from different approaches deserve mention:

- Nay (2017) developed a legislative predictor for the United States Congress using word vectors and an ensemble model, achieving approximately 68% accuracy using only textual data, without incorporating additional contextual features.
- Bari et al. (2021) introduced a vote predictor for congressional members during legislative discussions, serving as a basis for determining law approval, with an 80% accuracy rate.
- Katz et al. (2017) applied a model to predict votes in the United States Supreme Court, achieving 71.9% accuracy in predicting individual justices' votes and 70.2% accuracy in case outcome predictions.

Conclusions

The use cases presented in this article demonstrate the application of AI tools in a parliamentary library. We believe it is important to showcase real cases, not only from the perspective of recent applications like Large Language Models (LLMs) and RAG but also projects that highlight other technological facets that have contributed over time. These contributions span from AI to automation, process efficiency improvements, and the development of new products.

Both from the perspective of Semantic Web technologies and Open Data, which lay the groundwork for building systems based on Knowledge Graphs and enriched data, to the use of natural language processing for enhancing and streamlining computer-assisted human processes, as well as providing tools that encapsulate human reasoning to generate legislative predictions, AI has become fundamental in BCN as a parliamentary library. This integration of AI-based technologies has been crucial in offering a broader and better catalog of products and services to users.

The practice of publishing open legislative data can be highly beneficial in other cultural contexts, as it helps ensure legal certainty through open access to data and enables citizens to develop applications that make practical use of national legislation and legislative data. At the same time, open data also serves as a mechanism for promoting active transparency, facilitating direct access for the population and fostering trust in the government.

The creation of products such as the History of Law and Parliamentary Labour through the use of interoperability standards like AKN, artificial intelligence tools such as NLP, and automated archival systems within repositories constitutes a highly replicable framework for organizations responsible for providing information services related to archives and historical collections. The adoption of standards like AKN enables the use of widely available workflow management tools, accelerating the development of projects in similar environments.

Similarly, the legislative approval predictor presented in this work is a highly replicable tool in other contexts, such as legislative bodies in different countries. It also provides general guidelines for designing similar tools in diverse scenarios where the dependent variable is the approval or rejection of a measure, based on a set of factors that may vary over time.

Authors

Francisco Cifuentes Silva is the Head of Research Projects in the Information Technology Department at the Library of Congress of Chile and a PhD candidate in Computer Science at the University of Oviedo. He holds a degree in Computer Engineering from Universidad de La Frontera and a Master's in Web Engineering from the University of Oviedo. His expertise spans the Semantic Web, open data, machine learning, text processing, data science, and data visualization. He led the development of the *History of Law and Parliamentary Labor* system, and pioneered the first legislative open data portal in Latin America.

Hernán Astudillo is Professor, Universidad Andrés Bello (Chile), PI at Institute for Health and Wellness Technologies (ITiSB-UNAB). Informatics Engineer (UTFSM, 1988), Ph.D. Information and Computer Science (Georgia Tech, 1995). Formerly, has been at MCI.Systemhouse, Financial Systems Architects, Solunegocios, and UTFSM. His main interest is identification, evaluation and reuse of technological decisions with trade-offs using imperfect information. Conducts research, teaching and technology-transfer in Software-Architecture and Process-Improvement, and their applications in e-Health, Digital-Government, and Cultural-Informatics. Has been Executive-Secretary of CLEI (Latin-American Informatics Association); published over 180 articles in peer-reviewed conferences and journals; organized several national and international conferences; and lead R&D projects and international collaborations.

Jose Emilio Labra Gayo is full Professor at University of Oviedo, Spain since 2021. Founder and main researcher of WESO (Web Semantics Oviedo) research group. He is coauthor of the books "Validating RDF data" and "Knowledge graphs", from Springer

Nature as well as more than 100 research papers on semantic web and related technologies. He maintains several open source libraries to work with RDF like rudof and rdfshape which are implemented in Rust and Scala. He was coordinator of the Master in Web Engineering, and Dean of the School of Computer Science Engineering at University of Oviedo.

References

Abolhassani, M., Fuhr, N. & Govert, N. Information Extraction and Automatic Markup for XML¨ Documents. *Intelligent Search On XML Data*. (2003)

Akhtar, S., Reilly, R. & Dunnion, J. Automating XML markup using machine learning techniques. *Journal Of Systemics, Cybernetics And Informatics, Vol 2, Number 5*. pp. 1216 (2004)

Baeza-Yates, R. & Ribeiro-Neto, B. Modern Information Retrieval: The Concepts and Technology behind Search. *ACM Press New York*. 82 pp. 944 (2011)

Baker, T. A grammar of Dublin Core. *D-lib Magazine*. 6, 3 (2000)

Bari, A., Brower, W., & Davidson, C. Using artificial intelligence to predict legislative votes in the united states congress. In 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA) (pp. 56-60). IEEE. (2021).

Bechhofer, S. & Miles, A. SKOS Simple Knowledge Organization System Reference. (2009) Available at: http://www.w3.org/TR/2009/REC-skos-reference-20090818/ (Accessed 20 December 2024)

Ben-Porat, C. & Lehman-Wilzig, S. Political discourse through artificial intelligence: Parliamentary practices and public perceptions of chatbot communication in social media. *The Rhetoric Of Political Leadership*. pp. 230-245 (2020)

T. Berners-Lee. Linked data - design issues. (2006) Available at: http://www.w3.org/DesignIssues/LinkedData.html, (Accessed 20 December 2024)

Bizer, C., Cyganiak, R. & Heath, T. How to publish Linked Data on the Web. (2008) Available at: http://tomheath.com/slides/2008-10-karlsruhe-how-to-publish-linked-data-on-the-web.pdf (Accessed 20 December 2024)

Bolioli, A., Dini, L., Mercatali, P. & Romano, F. For the automated mark-up of italian legislative texts in xml. *Legal Knowledge And Information Systems (Jurix 2002*. pp. 21-30 (2002)

Brickley, D. & Miller, L. FOAF vocabulary specification 0.91. (2007)

Burget, R. Automatic Document Structure Detection for Data Integration. *Business Information Systems*. pp. 391-397 (2007)

Candela, G., Escobar, P., Sáez, M. D., & Marco-Such, M. A Shape Expression approach for assessing the quality of Linked Open Data in libraries. Semantic Web, 14(2), 159-179. (2023)

Chalkidis, I., Nikolaou, C., Soursos, P. & Koubarakis, M. Modeling and Querying Greek Legislation Using Semantic Web Technologies. *The Semantic Web*. pp. 591-606 (2017)

Cifuentes-Silva, F., Sifaqui, C. & Labra-Gayo, J. Towards an architecture and adoption process for linked data technologies in open government contexts. *Proceedings Of The 7th International Conference On Semantic Systems - I-Semantics '11*. pp. 79-86 (2011)

Cifuentes-Silva, F. & Labra Gayo, J. Legislative Document Content Extraction Based on Semantic Web Technologies. *The Semantic Web*. pp. 558-573 (2019)

Cifuentes-Silva, F., Fernandez-´ Alvarez, D. & Labra-Gayo, J. National Budget as Linked Open´ Data: New Tools for Supporting the Sustainability of Public Finances. *Sustainability*. 12, 4551 (2020)

Cifuentes-Silva, F., Labra-Gayo, J., Astudillo, H. & Rivera-Polo, F. Using Polarization and Alignment to Identify Quick-Approval Law Propositions: An Open Linked Data Application. *Applied Informatics*. pp. 122-137 (2023)

Cifuentes-Silva, F., Astudillo, H., Labra-Gayo, J. & Rivera-Polo, F. Toward Efficient Legislative Processes: Analysis of Chilean Congressional Bill Votes Using Semantic Web Technologies. *SN Computer Science*. 5, 604 (2024)

Elsawy, E. & Shehata, A. Open government data initiatives in the Maghreb countries: An empirical analysis. *IFLA Journal*. 49, 61-73 (2023)

Fitsilis, F. Artificial Intelligence (AI) in parliaments - preliminary analysis of the Eduskunta experiment. *The Journal Of Legislative Studies*. 27, 621-633 (2021),

Gacitua, R., Aravena-Diaz, V., Cares, C. & Cifuentes-Silva, F. Conceptual distinctions for´ traceability of history of law. *2016 11th Iberian Conference On Information Systems And Technologies (CISTI)*. pp. 1-6 (2016)

Gagnon, S. & Azzi, S. Semantic Annotation of Parliamentary Debates and Legislative Intelligence Enhancing Citizen Experience. *Electronic Government And The Information Systems Perspective*. pp. 63-76 (2022)

García-Gonzalez, H., Boneva, I., Staworko, S., Labra-Gayo, J.E. and Cueva Lovelle, J.M., ShExML:´ improving the usability of heterogeneous data mapping languages for first-time users, PeerJ Computer Science 6 (2020)

Gesnouin, J., Tannier, Y., Da Silva, C. G., Tapory, H., Brier, C., Simon, H., ... & Yang, S. Llamandement: Large language models for summarization of french legislative proposals. (2024).

Giunchiglia, F., Shvaiko, P. & Yatskevich, M. Semantic Schema Matching. *On The Move To Meaningful Internet Systems 2005: CoopIS, DOA, And ODBASE*. 3760 pp. 347-365 (2005)

Gruber, T. A translation approach to portable ontology specifications. *Knowledge Acquisition*. 5, 199-220 (1993)

Guha, R. & Brickley, D. RDF Schema 1.1. (2014). Available at: https://www.w3.org/TR/rdf-schema/ (Accessed 20 December 2024)

Hertel, A., Broekstra J., and Stuckenschmidt, H. *RDF Storage and Retrieval Systems*. (2008).

Hyvönen, E., Sinikallio, L., Leskinen, P., La Mela, M., Tuominen, J., Elo, K., ... & Kesäniemi, J. Finnish parliament on the semantic web: Using ParliamentSampo data service and semantic portal for studying political culture and language. In Digital Parliamentary Data in Action. CEUR-WS. org. (2022)

Jean-Mary, Y., Shironoshita, E. & Kabuka, M. Ontology matching with semantic verification. *Web Semantics: Science, Services And Agents On The World Wide Web*. The Web of Data, (2009)

Katz DM, Bommarito MJ II, Blackman J. A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE 12. (2017)

Klaus, S., Van Hecke, R., Djafari Naini, K., Altingovde, I. S., Bernabé-Moreno, J., & Herrera-Viedma, E. Summarizing legal regulatory documents using transformers. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2426-2430). (2022)

Koryzis, D., Dalas, A., Spiliotopoulos, D., & Fitsilis, F.. ParlTech: Transformation Framework for the Digital Parliament. Big Data and Cognitive Computing, 5(1), Article 1. (2021)

KALFOGLOU, Y. & SCHORLEMMER, M. Ontology mapping: the state of the art. *The Knowledge Engineering Review*. 18, 1-31 (2003)

Klapwijk, W., Wingate Gray, S., Uzwyshyn, R., Wen Sze, T., Chee Kiam, L., Ying Yi, C., Fritz, S., Mason, I., Leclaire, C., Balnaves, E. & Others Trends and Issues in Library Technology (TILT) Newsletter. International Federation of Library Associations. (2021)

Knublauch, H., TopQuadrant, Inc., Kontokostas, D. & University of Leipzig Shapes constraint language (SHACL). *W3C Recommendation*. 11 pp. 8 (2017)

Labra-Gayo, J., Prud'hommeaux, E., Solbrig, H. & Rodríguez, J. Validating and Describing Linked Data Portals using RDF Shape Expressions.. *LDQ SEMANTICS*. (2014)

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih,¨ W., Rocktaschel, T. & Others Retrieval-augmented generation for knowledge-intensive nlp¨ tasks. *Advances In Neural Information Processing Systems*. 33 pp. 9459-9474 (2020)

Louis, A., van Dijck, G., & Spanakis, G.. Interpretable long-form legal question answering with retrieval-augmented large language models. In Proceedings of the AAAI Conference on Artificial Intelligence (2024)

Malakhov, K., Petrenko, M., & Cohn, E.. Developing an ontology-based system for semantic processing of scientific digital libraries. South African Computer Journal, 35(1), 19-36. (2023)

Mendes, P., Jakob, M., Garc´ıa-Silva, A. & Bizer, C. DBpedia Spotlight: Shedding Light on the Web of Documents. *Proceedings Of The 7th International Conference On Semantic Systems*. pp. 1-8 (2011)

Nay, J. J.. Predicting and understanding law-making with word vectors and an ensemble model. PloS one, 12(5). (2017)

Niu, J.. Diffusion and adoption of linked data among libraries. Proceedings of the Association for Information Science and Technology, 57(1). (2020)

Ley 21.200 - Reforma constitucional que habilita proceso constituyente, 24DIC-2019 MINISTERIO SECRETARIA GENERAL DE LA PRESIDENCIA. Available at: https://www.bcn.cl/leychile/navegar?idNorma=1140340. (Accessed 20 December 2024)

D.F.L. Nº 1 - Texto refundido, coordinado y sistematizado del Código Civil chileno, 20-MAY-2023 MINISTERIO DE JUSTICIA. Available at https://www.bcn.cl/leychile/navegar?idNorma=172986. (Accessed 20 December 2024)

Oksanen, A., Tamper, M., Tuominen, J., Makel¨ a, E., Hietanen, A. & Hyvönen, E. Semantic¨ Finlex: Transforming, publishing, and using Finnish legislation and case law as linked open data on the web. *Knowledge Of The Law In The Big Data Age*. pp. 212-228 (2019)

Palmirani, M. & Vitali, F. Akoma-Ntoso for legal documents. *Legislative XML For The Semantic Web: Principles, Models, Standards For Document Management*. pp. 75-100 (2011)

Park, J. R., Brenza, A., & Richards, L.. BIBFRAME linked data: A conceptual study on the prevailing content standards and data model. In Linked open data-applications, trends and future developments. (2020)

Prud'hommeaux, E., Labra Gayo, J. & Solbrig, H. Shape expressions: an RDF validation and transformation language. *Proceedings Of The 10th International Conference On Semantic Systems*. pp. 32-40 (2014)

Rafat, M. AI-powered Legal Virtual Assistant: Utilizing RAG-optimized LLM for Housing Dispute Resolution in Finland. (2024)

Samples, J., & Bigelow, I. MARC to BIBFRAME: Converting the PCC to Linked Data. Cataloging & classification quarterly, 58(3-4), 403-417. (2020)

Silva, A. L., & Terra, A. L. Cultural heritage on the Semantic Web: The Europeana Data Model. IFLA Journal, 50(1), 93-107. (2024).

Solbrig, H., Prud'hommeaux, E., Grieve, G., McKenzie, L., Mandel, J., Sharma, D. & Jiang, G. Modeling and validating HL7 FHIR profiles using semantic web Shape Expressions (ShEx). *Journal Of Biomedical Informatics*. 67 pp. 90-100 (2017)

Thuluva, A., Anicic, D. & Rudolph, S. Shaping Device Descriptions to Achieve IoT Semantic Interoperability. *ESWC 2018*. (2018)

Tsai, C. & Roth, D. Cross-lingual wikification using multilingual embeddings. *Proceedings Of The 2016 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies*. pp. 589-598 (2016)

Usbeck, R., Ngomo, A., Roder, M., Gerber, D., Coelho, S., Auer, S. & Both, A. AGDISTIS-¨ graph-based disambiguation of named entities using linked data. *International Semantic Web Conference*. pp. 457-471 (2014)

W3C, OWL Working Group, OWL 2 Web Ontology Language Document Overview (Second Edition) (2012). Available at: http://www.w3.org/TR/2012/REC-owl2-overview-20121211/ (Accessed 20 December 2024)

W3C, SPARQL 1.1 overview. (2013), Available at: https://www.w3.org/TR/sparql11-protocol/ (Accessed 20 December 2024)

Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., ... & Fleisch, B. CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In International Conference on Case-Based Reasoning (pp. 445-460). Cham: Springer Nature Switzerland. (2024)

Zapounidou, S., Ioannidis, L., Gerolimos, M., Koufakou, E., & Bratsas, C.. Entity Management Using RDA and Wikibase: A Case Study at the National Library of Greece. Journal of Library Metadata, 24(2), 111-131. (2024)

Zmiycharov, V., Tsonkov, T., & Koychev, I. EurLexSummarization–A New Text Summarization Dataset on EU Legislation in 24 Languages with GPT Evaluation. In Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024) (pp. 206-213). (2024)