Enhancing Access to Legal Data through Ontology-based Representation: A Case Study with Brazilian Judicial Appeals

Melissa Zorzanelli Costa[1,*], Thiago Baiense Peçanha Vieira[2], Jean-Rémi Bourguet[2], Giancarlo Guizzardi[3] and João Paulo A. Almeida[1]

[1]*Federal University of Espírito Santo, Av. Fernando Ferrari, 514, Vitória, ES, Brazil*

[2]*Vila Velha University, Av. Comissário José Dantas de Melo, 21, Vila Velha, ES, Brazil*

[3]*University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands*

Abstract

In Brazil, legal requirements for public information access, as mandated by Law n° 12.527/2011, have amplified the role of the open data portals in disseminating data of collective and general interest. Despite legal provisions, there are persistent difficulties in presenting data in first-class semantic formats, which ultimately creates obstacles for digital citizens to fully exercise their newfound rights to information access. These obstacles can be addressed by building semantic data warehouses to enhance the use of open data through computational ontologies. In this paper, we demonstrate the use of a well-founded legal ontology for representing data from legal decisions extracted from a Brazilian judicial organ website. We focused our approach on a specific type of appeal in the Brazilian legal system, the Request for Standardization (RS) of interpretation of federal law, which seeks to standardize the understanding of the Appeals Panels of Federal Special Courts. Employing web scraping techniques, we built a complete ETL (Extract, Transform, Load) process to triplify data on RS appeals and their rulings. We used a gUFO-based OWL renderization of a previously developed OntoUML ontology (called OntoRS) to transform the extracted data into a suitable RDF format and populate a Virtuoso triple store. Thus, the OntoRS ontology allowed us to perform SPARQL queries to obtain new insights, metrics and small RDF graphs.

Keywords

ETL process, Web scraping, Legal ontology, Triplification, SPARQL queries

1. Introduction

In today's data-driven world, citizens are increasingly interested in accessing valuable insights mined from vast amounts of data. In Brazil, Federal Law n° 12.527/2011 regulates access to public information which is now considered a key citizen's right with the corresponding state's duty. This law has made a significant contribution to enhancing transparency by mandating proactive disclosure of data of collective and general interest (referred to as active transparency).

Proceedings of the XVI Seminar on Ontology Research in Brazil (ONTOBRAS 2023) and VII Doctoral and Masters Consortium on Ontologies (WTDO 2023), Brasilia, Brazil, August 28 - September 01, 2023.

[*]Corresponding author.

melissa.inf.ufes@gmail.com (M. Zorzanelli Costa); thiago.vieira@uvvnet.com.br (T. B. P. Vieira); jean-remi.bourguet@uvv.br (J. Bourguet); g.guizzardi@utwente.nl (G. Guizzardi); jpalmeida@ieee.org (J. P. A. Almeida)

0000-0002-4695-5799 (M. Zorzanelli Costa); 0009-0000-8301-8000 (T. B. P. Vieira); 0000-0003-3686-1104 (J. Bourguet); 0000-0002-3452-553X (G. Guizzardi); 0000-0002-9819-3781 (J. P. A. Almeida)

For example, the Open Data Portal[1] is considered as a central point of reference for Brazilian public data on any subject, aiming to standardize data access, reuse and interoperability. To better explore the potential of knowledge encoded in data and in order to be able to use some reasoning mechanisms, a usual practice is to provide semantic data warehouses by enriching open data through computational ontologies [1]. Despite the efforts of the National Open Data Infrastructure to establish standards and formats and stimulate government open data distribution [2], the vast majority of the public agencies are struggling to expose this data through first-class semantic formats [3]. During the last decade, a substantial number of works emerged from different areas to offer solutions for transforming data into open formats (e.g., concerning academic output [4], patient data [5], epidemiology [6] or public budgets [7]).

An important source of public data is the judicial system. This is a particularly sensitive information domain, as access to information by citizens is directly related to their exercising of fundamental rights, such as the right to due process enshrined in the Brazilian constitution. In addition, access to judicial information is relevant to all actors of the judicial system, including lawyers, judges, clerks, etc. The judicial system in Brazil is highly digital, mainly in its federal sphere, with a significant number of digital-born documents and processes. Therefore, there is a lot to gain from openly exposing existing judicial process data in semantic formats. This entails not only the extraction of data from legal (web-based) systems, but also its representation in domain-adequate and easily processable formats. Despite the potential benefits, there are also significant challenges, as legal process data is highly specialized, often expressed with impervious jargon. Therefore, a successful approach requires interdisciplinary teams and the use of ontological analysis methodologies.

In this paper, we face these challenges in the case of data concerning a specific type of appeal in the Brazilian legal system, which is part of a specialized procedure in Federal Special Courts that is considered particularly verbose and nontransparent. A Request for Standardization (RS) of interpretation of federal law is a petition that seeks to standardize the understanding of the Appeals Panels of Federal Special Courts. As an appeal, its objective is to reform the judgment handed down by the Appeals Panel or by the National Uniformization Panel (TNU) [8]. This type of appeal was the subject of [9], in which the authors presented a multi-viewpoint conceptual model of an RS by combining a structural perspective modeled with an ontology-driven conceptual modeling language (OntoUML) with a dynamic perspective modeled in a business process notation (BPMN). Here, we go one step further and describe an ETL (Extract, Transform, Load) process that uses as target an operational version of the aforementioned OntoUML Request for Standardization (RS) ontology. We perform the 'triplification' of knowledge extracted from judgments available in an unstructured format in the official TNU jurisprudence website[2]. For that, we crawled the TNU web site supported by classical web scraping techniques. Once the data was extracted, we transformed it into a suitable RDF format according to the operational ontology and populated a Virtuoso[3] triplestore. Thus, we were able to perform SPARQL queries to obtain new insights, metrics and small RDF graphs.

This paper is further structured as follows: in Section 2, we discuss the current state of affairs

[1] https://dados.gov.br/
[2] https://www.cjf.jus.br/jurisprudencia/tnu/
[3] https://virtuoso.openlinksw.com/

concerning open legal data in Brazil; in Section 3, we present the conceptual structures that underpin our data transformation; in Section 4, we describe the procedure for data extraction and triplification; in Section 5, we show some SPARQL queries performed on top of our triple store; in Section 6, we discuss works closely related to ours; and finally, in Section 7, we conclude our paper and draw some perspectives.

2. Problem statement

In the scope of the Federal Special Courts in Brazil, the Law n° 10.259/01[4] established the role of a National Uniformization Panel (TNU), in its article 14, with the competence to eliminate differences of interpretation in matters of substantive federal law. Its creation integrating the federal structure took place in 2008. Currently, its functioning is governed further by its own regiment, Resolution n° 586/19[5]. In this work, we deal with the Request for Standardization of a law interpretation, which is the appropriate appeal against the judgment handed down by an Appeals Panel and directed to the TNU. The jurisprudence formed by the decisions handed down by this panel (the TNU) are available on its own website[6], allowing (unstructured) access to citizens and interested jurists, such as law clerks and lawyers. Unlike other judicial bodies such as the Superior Court of Justice (STJ)[7], there is no open API for the TNU. This means users cannot explore data to its full potential and are limited to parameterizing predefined queries at the website and processing query results manually. As a consequence, data users would have to resort to technical workarounds such as web scraping to perform more sophisticated data processing. A number of technical barriers stand in their way, including *captchas*, which hinder automated access. While *captchas* are usually indented to offer resilience to cyberattacks, they also hinder legitimate access to public data as mandated by law.

In addition to these technical barriers, there are also potential legal barriers to data access. This is because web scraping faces the delicate balance between open access to information and privacy [10]. In other countries, for example in the United States, as far as we are aware, there is no statute designed to specifically address web scraping [11], its legality can vary depending on the jurisdiction and some specificities about the data (e.g. copyright laws, data protection regulations, and any applicable intellectual property rights, trademark infringement, false advertising, or dilution by tarnishment [12]). In 2018, the GDPR (General Data Protection Regulation) was approved in the European Union (EU). Inspired by this, the Brazilian Congress, which already had some discussions on data protection laws on the agenda, sanctioned in August 2018 the General Data Protection Law (LGPD) (Federal Law n° 13.709/2018). Fortunately, the LGPD allows the handling and extraction of data for academic and research purposes, which is the case of this work.

[4] https://www.planalto.gov.br/ccivil_03/leis/leis_2001/l10259.htm

[5] https://www.cjf.jus.br/cjf/corregedoria-da-justica-federal/turma-nacional-de-uniformizacao

[6] https://www.cjf.jus.br/jurisprudencia/tnu/

[7] https://dadosabertos.web.stj.jus.br/dataset/

3. Reference Ontology and its Implementation

Our approach is based on a reference ontology for this domain developed in OntoUML, an ontologically well-founded UML profile whose primitives reflect ontological distinctions of the Unified Foundational Ontology (UFO [13]). We extend here the legal reference ontology presented in [9] (dubbed OntoRS) to represent the analysis of the `Granted and Admitted RS`s by the TNU. We also obtain a gUFO-based [14] operational implementation of OntoRS in OWL that is used further for data triplification.

Figure 1 shows the fragment of the reference ontology that is relevant for the purpose of this paper. The classes in green are types of relators, i.e., reified relationships according to the UFO that can change qualitatively in time while retaining their identity. This includes an overall `Judicial Process` including the `Petitions` that are part of it, such as the subkind of `Appeal` which is the focus of our attention here, the `Request for Standardizing the Interpretation of a Federal Law (RS)`. The ontology covers the decisions that are handed for an `RS` (and are modeled as types of events). The analysis of the admissibility of an `RS` is made in accordance to the aforementioned internal regulations of the TNU (Res n° 586/19) [9]. When an RS that was filed against an `Appellate Decision on SFCA` (Special Federal Court Appeal) is considered a `Granted and Admitted RS`, it is received by the `Minister-President`

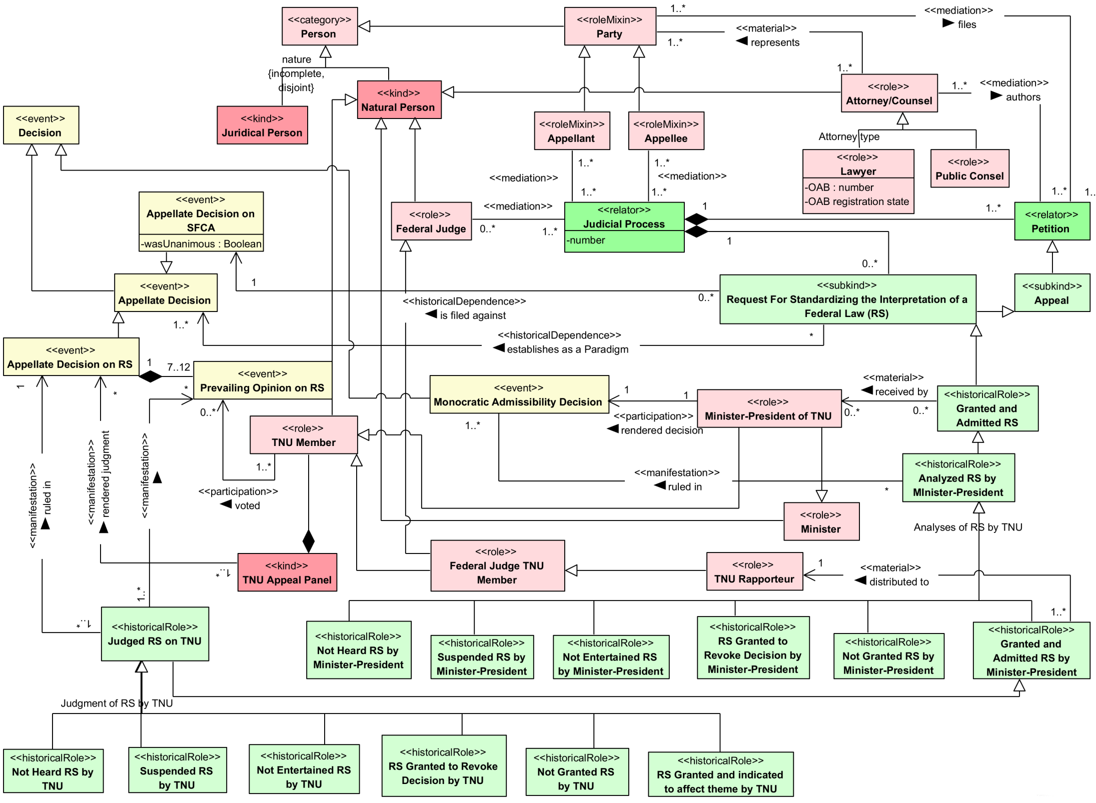


Figure 1: Request for Standardizing on TNU Legal Ontology (OntoRS).

of TNU, that analyses the RS and renders a Monocratic Admissibility Decision for it. According to the Monocratic Admissibility Decision, the RS is considered either: (i) Not Heard RS by Minister-President, (ii) Suspended RS by Minister-President, (iii) Not Entertained RS by Minister-President, (iv) RS Granted to Revoke Decision by Minister-President, (v) Not Granted RS by Minister-President or (vi) Granted and Admitted RS by Minister-President. If the analysis of the RS culminates in a decision that grants and admits the appeal within the scope of TNU (vi), the appeal is distributed to a TNU Rapporteur (that is a Federal Judge TNU Member, that is a member of the TNU Appeal Panel). The TNU Appeal Panel comprises 12 federal judges from the Appeals Panels of the Special Federal Courts, with 2 federal judges from each Tribunal Regional Federal (TRFs) Region in Brazil[8]. A judgment session takes place in which the TNU Rapporteur and other TNU Members present their votes. A Prevailing Opinion on RS then arises as part of the Appellate Decision on RS. According to the judgment of the TNU Appeal Panel, the RS may have been judged as: (i) Not Heard RS by TNU, (ii) Suspended RS by TNU, (iii) Not Entertained RS by TNU, (iv) RS Granted to Revoke Decision by TNU, (v) Not Granted RS by TNU, or (vi) RS Granted and Indicated to affect theme by TNU, in which case, the TNU Appeal Panel chooses this RS as a representative of the overall controversy and declare it a binding precedent.

In order to validate the presented Legal Ontology, it is possible to correlate the classes with the norms in force in Brazil that gave rise to them [9]. LexML norm fragment identifiers [15] were used for this purpose. Furthermore, the most specific norms, in this case, the TNU bylaws, were reviewed and studied. Interviews were conducted with law clerks who work directly with the admissibility of appeals at the TNU presidency as well as those who work with admissibility at the TRF courts of the 2nd and 4th regions, in the end of May 2023.

Along the modeling process, the OntoUML plugin for the Visual Paradigm tool was employed to check for model wellformedness according to the UFO taxonomy rules. Further, the tool provided for a transformation of the reference ontology into an operational OWL Ontology, reusing gUFO [14] (a lightweight implementation of the Unified Foundational Ontology) by specializing and instantiating its elements. The ontology was found consistent using the HermiT reasoner (version 1.4.3.456) in Protégé. Figure 2 shows the resulting taxonomy as anchored in the gUFO notion of Relator.

4. Extraction and Triplification

This section presents the ETL process which starts with data extraction from the TNU website[2] as summarized in the Figure 3. The website presents a search form with some mandatory parameters (e.g. dates, identifiers, keywords). To prototype a first extraction, we chose a given range of dates (August 20th, 2022 to January 1st, 2023) which retrieved all judgments (Appellate Decision on RS) in this specific period. Simple HTTP requests could in principle retrieve the search results. However, the TNU website prevents such requests, possibly as a means of preventing denial-of-service attacks on the server.

Therefore, to overcome this obstacle, we opted to simulate user interaction through the

[8] Each Tribunal Regional Federal court in Brazil is assigned to one of six “regions”.

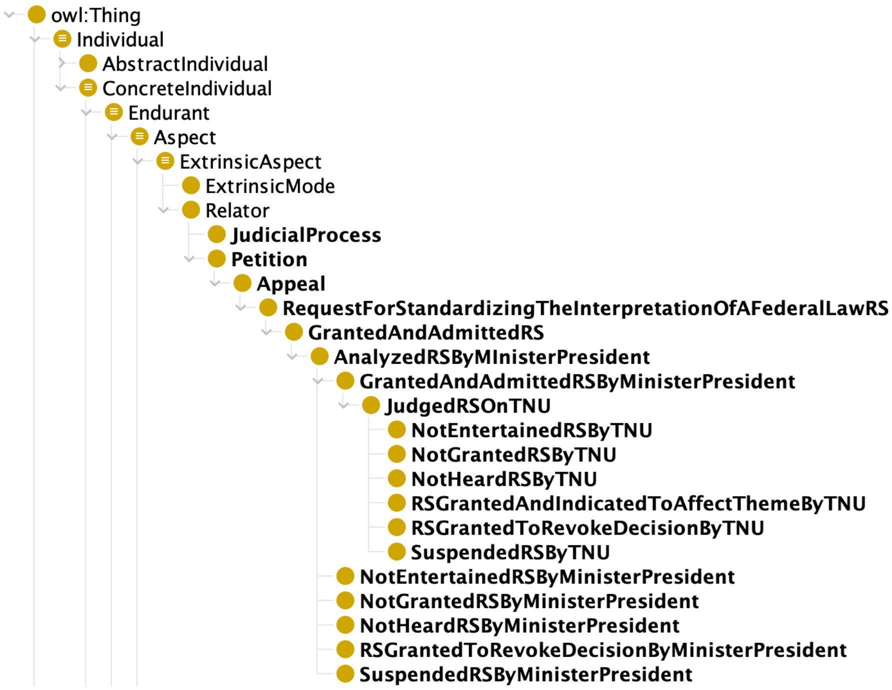

owl:Thing
Individual
AbstractIndividual
ConcreteIndividual
Endurant
Aspect
ExtrinsicAspect
ExtrinsicMode
Relator
JudicialProcess
Petition
Appeal
RequestForStandardizingTheInterpretationOfAFederalLawRS
GrantedAndAdmittedRS
AnalyzedRSByMInisterPresident
GrantedAndAdmittedRSByMinisterPresident
JudgedRSOnTNU
NotEntertainedRSByTNU
NotGrantedRSByTNU
NotHeardRSByTNU
RSGrantedAndIndicatedToAffectThemeByTNU
RSGrantedToRevokeDecisionByTNU
SuspendedRSByTNU
NotEntertainedRSByMinisterPresident
NotGrantedRSByMinisterPresident
NotHeardRSByMinisterPresident
RSGrantedToRevokeDecisionByMinisterPresident
SuspendedRSByMinisterPresident

Figure 2: Taxonomy of Appeals in Protégé anchored in the gUFO notion of Relator.

Selenium library[9]. With this tool, a so-called headless browser navigates the content of Web documents, selecting elements of the document's tree structure by using criteria (e.g. ID, name, class, link, etc.) as specified by an `xpath` expression. In Selenium, as shown in Listing 1, the methods `find_element` and `send_key` respectively find the element using `xpath` and send the data to simulate form filling.

```
navegador.find_element('xpath', '//*[@id="formulario:j_idt25_input"]').send_keys("20/08/2022")
```

Listing 1: Methods to find and send data.

Once we managed to automatically fill out the form, we had to properly extract the full contents of the decisions ("*inteiro teor do acórdão*"). For that, we picked up from the syllabus (i.e. a brief summary of the decision) an access link to the full decision. We used the `beautifulsoup` library[10] which allows browsing through an HTML page and searching for specific elements. Then, a list is built with the respective links of the full decisions, enabling the automation of the entire process as described in Listing 2.

[9] https://pypi.org/project/selenium/
[10] https://pypi.org/project/beautifulsoup4/

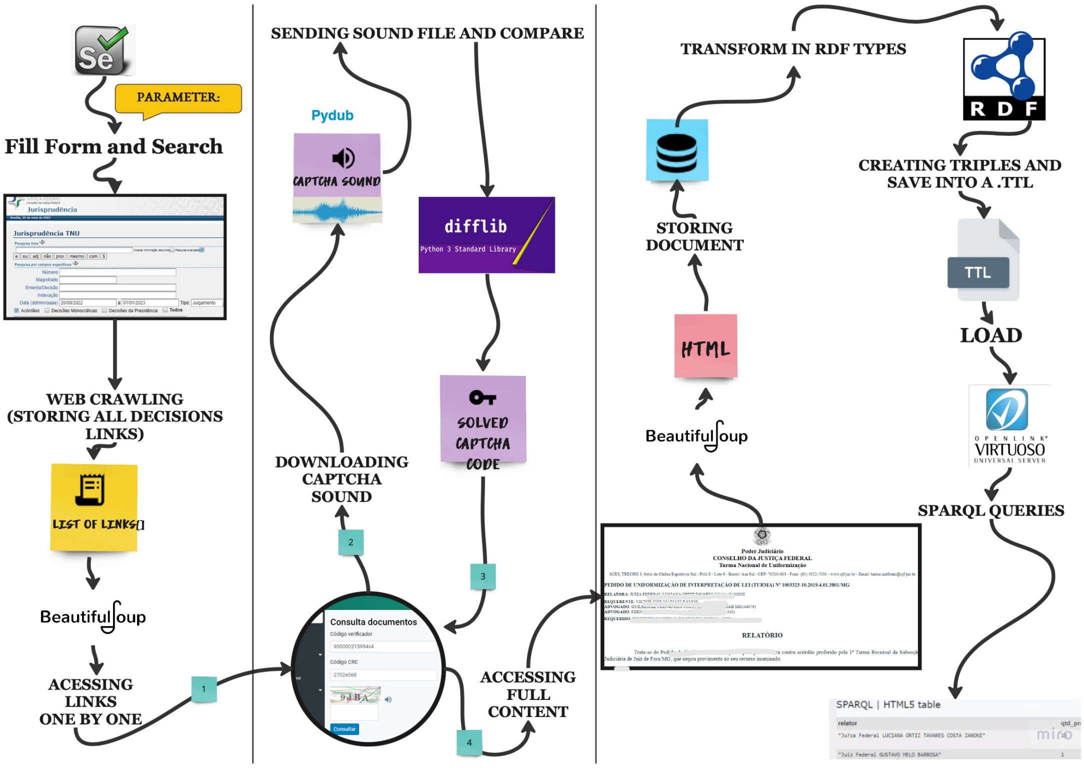


Figure 3: Overview of the ETL process: from web scraping to SPARQL queries in the triplestore.

```
soup = BeautifulSoup(navegador.page_source,'html.parser')
linksInteiroTeor = []
links = soup.findAll( 'a', attrs = { 'target' : '_blank'} )
```

Listing 2: Gathering full decision links.

However, when trying to access these URIs, we were faced with a *captcha*, a little test confirming the human nature of the user generally used to avoid an overload of requests. In our case, the *captcha* necessarily contained a sequence of four digits (letters or numbers) presented to the user via an image or a sound. We opted to interpret the sounds by building a method distinguish the various numbers and letters. We employed the `pydub`[11] and the `difflib` libraries[12] in the process. We omit here the details of the procedure we followed in order to prevent malicious exploits. For our purposes here, it suffices to say that significant expertise and programming effort was required in the process. Our extraction was designed not to cause denial of service because we performed the treatments sequentially, one at a time, outside business hours and with low speed and frequency. Breaking the *captcha* is carried out only

[11] https://github.com/jiaaro/pydub
[12] https://docs.python.org/3/library/difflib.html

for academic purposes and, with the interest of the organizations involved, access would be possible without this type of artificial barrier.

After retrieving the correct sequence of digits, we used again some Selenium methods to insert the *captcha* code and access the inner contents of the judgments as exhibited in Listing 3. This serves as the basis for the next step of the ETL process.

```
navegador.find_element('xpath', '//*[@id="txtInfraCaptcha"]').send_keys(CaptchaCode)
navegador.find_element('xpath', '//*[@id="sbmConsultar"]').click()
soup = BeautifulSoup(navegador.page_source,'html.parser')
```

Listing 3: Methods to access “inteiro teor”.

We carried out the extraction in two ways: (i) picking up metadata from the syllabus, and (ii) downloading the inner contents of the judgment in HTML format. Finally, we performed a sequence of RDF serializations necessary to our transformation phase as illustrated in Listing 4.

```
g = Graph()
ors = Namespace('http://purl.org/nemo/ontors#')
tnu = Namespace('http://purl.org/nemo/tnu/')
numProcesso = soup.find('span', {'data-sin_numero_processo': 'true'}).text.strip()
processo_URI = URIRef(tnu[numProcesso])
g.add((processo_URI, RDF.type, ors.JudicialProcess))
```

Listing 4: Creation of an RDF triple instantiating a process.

We used the RDFLib library[13] to transform the extracted legal data identified through their specific markups and represented by using the vocabulary of the ontology in gUFO presented in the previous section. In the next session, we will present some kinds of SPARQL queries demonstrating the possible implications of our resemantization process.

5. Querying legal data

In order to present the potential applications of data stored in terms of the operational ontology in the triplestore, we show here four kinds of SPARQL queries returning insights about the legal data extracted from TNU’s portal.

5.1. Basic SPARQL query

Supported by the vocabulary of our ontology, we are able to query the TNU’s data through Virtuoso’s endpoint. In Listing 5, we present a basic SPARQL query (with `PREFIX` declarations omitted for brevity) to retrieve the names of all TNU rapporteurs in the triplestore.

[13] https://rdflib.readthedocs.io/

```
SELECT    ?rapporteurLabel
WHERE {   ?rapporteur rdf:type   ors:Rapporteur ;
                      rdfs:label ?rapporteurLabel .   }
```

Listing 5: Basic SPARQL query.

5.2. Reasoning-based SPARQL query

Loading a gUFO renderization of OntoRS in Virtuoso, the queries can benefit from semantic reasoning by using subsumption inferences for instance. In Listing 6, we present a reasoning-based SPARQL returning the complete set of URIs representing parties in a legal process (instances of **`Party`**). Because of subsumption inferences, instances of **`Appellant`** and **`Appellee`** are retrieved with this query.

```
SELECT    ?party
WHERE {   ?party rdf:type ors:Party .   }
```

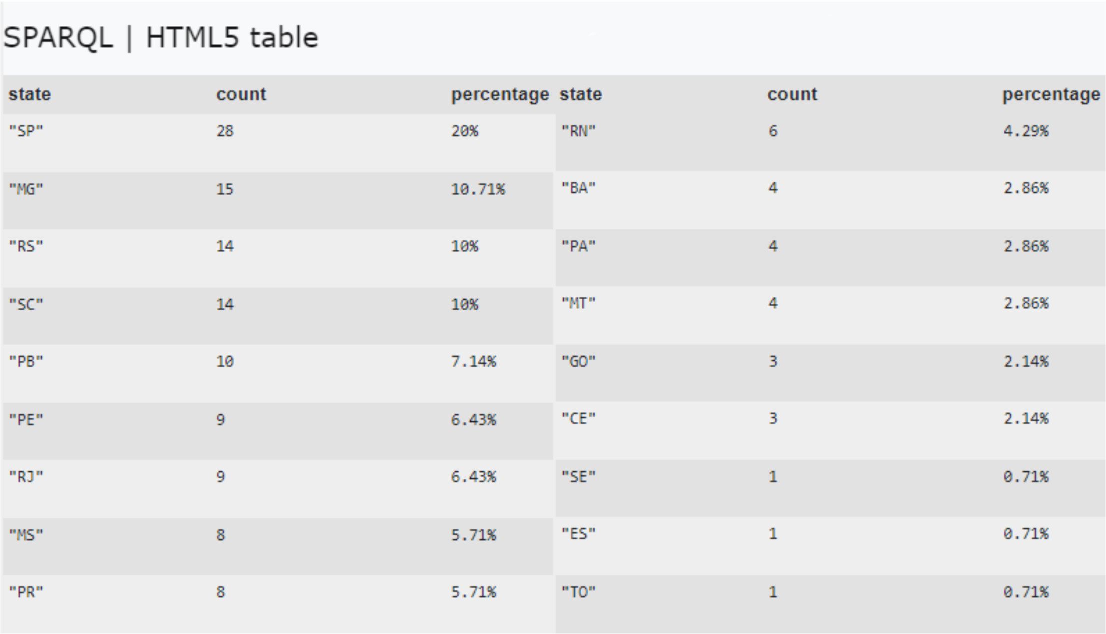
Listing 6: Reasoning-based SPARQL query.

5.3. Metrics-valued SPARQL query

Aggregate functions in SPARQL can be used to perform lightweight data analysis tasks. We take here as an example an analysis of the state of registration of the lawyers filing appeals to the TNU. In Brazil, a lawyer's main registration with the Brazilian Bar Association (OAB) must be made at the Sectional Council in which the lawyer intends to establish his/her professional domicile (**`OABRegistrationState`** in the ontology). The query described in Listing 7 retrieves the percentage of lawyers in the triplestore registered in each state as depicted in Figure 4.

```
SELECT ?state
       COUNT(?attorneyCounsel) AS ?count
       CONCAT(STR((ROUND((COUNT(?attorneyCounsel)*100/?totalCount)*100))/100),"%") AS ?percentage
WHERE {  ?attorneyCounsel rdf:type                ors:AttorneyCounsel ;
                          ors:OABRegistrationState ?state .
         {  SELECT   COUNT(?attorneyCounsel) AS ?totalCount
            WHERE {  ?attorneyCounsel rdf:type  ors:AttorneyCounsel .  } } }
GROUP BY ?state ?totalCount
ORDER BY DESC(?count)
```

Listing 7: Metrics-valued SPARQL query.



SPARQL | HTML5 table

state	count	percentage	state	count	percentage
"SP"	28	20%	"RN"	6	4.29%
"MG"	15	10.71%	"BA"	4	2.86%
"RS"	14	10%	"PA"	4	2.86%
"SC"	14	10%	"MT"	4	2.86%
"PB"	10	7.14%	"GO"	3	2.14%
"PE"	9	6.43%	"CE"	3	2.14%
"RJ"	9	6.43%	"SE"	1	0.71%
"MS"	8	5.71%	"ES"	1	0.71%
"PR"	8	5.71%	"TO"	1	0.71%

Figure 4: Results of the SPARQL query in Listing 7 printed from Virtuoso's Conductor interface.

5.4. Graph-valued SPARQL query

In SPARQL, CONSTRUCT queries can return a single RDF graph specified by a graph template. The result is an RDF graph substituting variables in the graph template by queried values and combining the triples into a single RDF graph by set union.

In Listing 8, we present a graph-valued SPARQL query returning a set of triples formed by the instances of the `AttorneyCounsel` class linked to the values of the Brazilian states through the data property `OABRegistrationState`.

```
CONSTRUCT { ?attorneyCounsel ors:OABRegistrationState ?state .   }
WHERE {     ?attorneyCounsel rdf:type                ors:AttorneyCounsel ;
                             ors:OABRegistrationState ?state .   }
```

Listing 8: Graph-valued SPARQL query.

Figure 5 shows a portion of the RDF graph produced by the query in Listing 8 visualized through the pydot library[14].

Finally, note that the complete workflow used for the ETL process, including the algorithms used during the extraction phase, the ontology OntoRS and the SPARQL queries aforementionned are available in a public repository[15].

[14] https://pypi.org/project/pydot/

[15] https://github.com/MelissaZor/ONTOBRAS2023

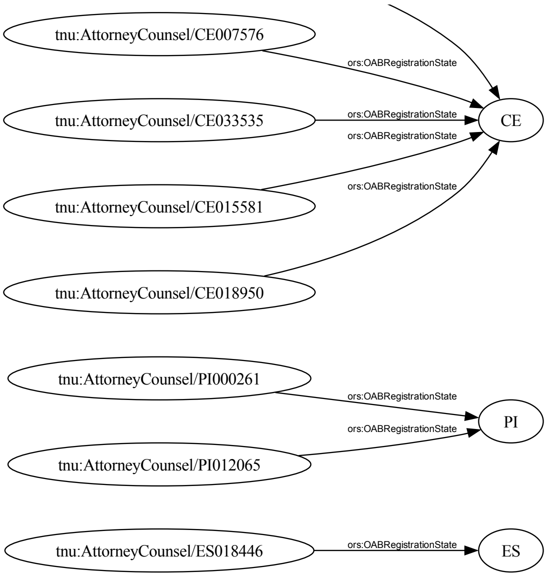

tnu:AttorneyCounsel/CE007576
tnu:AttorneyCounsel/CE033535
tnu:AttorneyCounsel/CE015581
tnu:AttorneyCounsel/CE018950
tnu:AttorneyCounsel/PI000261
tnu:AttorneyCounsel/PI012065
tnu:AttorneyCounsel/ES018446
ors:OABRegistrationState
CE
PI
ES

Figure 5: RDF graph produced from the SPARQL query in Listing 8

6. Related Work

In the domain of the open linked data, ETL has emerged as a category of processes [16] employed to gather data directly from operational systems, and remodel it through standardized formats to perform automated reasoning in powerful data warehouses [17]. Among the range of techniques used to perform data extraction, browser simulation provides the crawler the ability to simulate interaction behavior of a real user (see [18] for example). The reader is referred to [19], for a large review about the state-of-the-art of artificial intelligence-based approaches supporting unstructured information repositories browsing in legal domain.

In Brazil, the authors in [20] extracted information from public acts available in the Official Gazettes to support process of continuous auditing. Then, some SPARQL queries were built to infer conditions of non-compliance with the legislation. In [4], the authors propose an approach to support exposure and interoperability of public data from governmental publishers. A prototype application was built by providing a SPARQL endpoint to access scholar data from Lattes platform, the well known Brazilian information system managing user contents about science, technology, and innovation.

With respect to legal data, an ETL based framework can be materialized through search engines inputting legal argument in order to browse legal claims [21, 22] in documents. Having a well founded ontology-based knowledge representation allows to support effective tools for legal research [23]. Such approaches are recognized to improve the convenience of legal texts structured storage and avoid a lot of manual labor by professionals in the judicial field [24]. For instance, the authors in [25] present a framework to extract and classify named-entities and relations in Portuguese criminal reports from police investigations. A graph database representation in Neo4J is used to visualize the relations extracted from the documents. In [26], the authors propose a framework supported by the identification of recurring linguistic patterns to extract metadata from Greek Supreme Courts decisions (e.g. names of judges and lawyers, litigant parties, etc). Then, Akoma Ntoso [27] schema was adopted to semantically enrich the legal open data. This aforementioned schema is largely used to support the representation of different kinds of legal knowledge bases (e.g. for laws [28]).

7. Final Considerations

The significance of access to information in Brazil was facilitated by Law n° 12.527/2011, and the role played by open data portals in disseminating data of collective and general interest. However, the challenge of presenting data in semantic formats has limited the digital citizen's ability to fully exercise their newfound rights. In this paper, we proposed a solution to overcome these obstacles through the implementation of a semantic data warehouse, leveraging computational ontologies. By employing a well-founded legal ontology and employing web scraping techniques, we implemented an ETL (Extract, Transform, Load) process involving triplifications of legal decisions from a Brazilian judicial organ website, with an specific focus on the Request for Standardization (RS) of interpretation of federal law. By using a gUFO renderization of the OntoUML ontology OntoRS, the extracted data was transformed into a suitable RDF format and populated into a Virtuoso triple store. The OntoRS ontology proved instrumental in enabling SPARQL queries, which in turn yielded new insights, metrics, and small RDF graphs. Overall, this research demonstrates the practicality and value of employing semantic data approaches, specifically in the legal domain, to enhance the utilization of open data. By bridging the gap between data accessibility and data comprehension, these methodologies contribute to a more informed and empowered digital citizenry by fostering ethical transparency. In future work, we intend to pair conceptual modeling and machine learning to bridge the gap between the Information Retrieval and Semantic Web in the Brazilian legal domain.

Acknowledgements

Thiago B. P. Vieira is currently supported by an undergraduate grant from the Institutional Scientific Initiation Program of the Vila Velha University. João Paulo A. Almeida is funded in part by CNPq (313687/2020-0) and FAPES (281/2021 and 1022/2022).

References

[1] L. C. B. Martins, M. de Carvalho Victorino, M. Holanda, G. Ghinea, T. Grønli, UnBGOLD: UnB government open linked data: semantic enrichment of open data tool, in: Proceedings of the 10th International Conference on Management of Digital EcoSystems, MEDES 2018, Tokyo, Japan, September 25-28, 2018, ACM, 2018, pp. 1–6. doi:10.1145/3281375.3281394.

[2] C. Bittencourt, J. Estima, G. Pestana, Open Data Initiatives in Brazil, in: 14th Iberian Conference on Information Systems and Technologies (CISTI 2019), IEEE, 2019, pp. 1–4. doi:10.23919/CISTI.2019.8760592.

[3] S. Martin, M. Foulonneau, S. Turki, 1-5 Stars: Metadata on the Openness Level of Open Data Sets in Europe, in: Metadata and Semantics Research - 7th Research Conference, MTSR 2013, Thessaloniki, Greece, November 19-22, 2013. Proceedings, volume 390 of *Communications in Computer and Information Science*, Springer, 2013, pp. 234–245. doi:10.1007/978-3-319-03437-9_24.

[4] K. de Faria Cordeiro, F. F. de Faria, B. de Oliveira Pereira, A. Freitas, C. E. Ribeiro, J. V. V. B. Freitas, A. C. Bringuente, L. de Oliveira Arantes, R. Calhau, V. Zamborlini, et al., An approach for managing and semantically enriching the publication of linked open governmental data, in: Proceedings of the 3rd Workshop in Applied Computing for Electronic Government (WCGE), SBBD, 2011, pp. 82–95.

[5] F. C. Pellison, R. P. C. L. Rijo, V. C. Lima, R. R. de Lima, R. Martinho, R. J. C. Correia, D. Alves, Development and evaluation of an interoperable system based on the semantic web to enhance the management of patients' tuberculosis data, Procedia Computer Science 121 (2017) 791–796. doi:10.1016/j.procs.2017.11.102.

[6] V. Borges, N. Queiroz de Oliveira, H. F. Rodrigues, M. L. M. Campos, G. R. Lopes, A Platform to Generate FAIR Data for COVID-19 Clinical Research in Brazil, in: Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 1: ICEIS, INSTICC, SciTePress, 2022, pp. 218–225. doi:10.5220/0011066800003179.

[7] L. B. R. da Fonseca, A. A. Detoni, J. P. A. Almeida, R. de Almeida Falbo, Uma proposta de ontologia de referência para autorização orçamentária e execução da despesa pública, in: Proceedings of the IX ONTOBRAS Brazilian Ontology Research Seminar, Curitiba, Brazil, October 3rd, 2016, volume 1862 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016, pp. 210–215. URL: https://ceur-ws.org/Vol-1862/paper-22.pdf.

[8] M. A. J. de Santa Cruz Oliveira, Reforming the Brazilian Supreme Federal Court: A Comparative Approach, Wash. U. Global Stud. L. Rev. 5 (2006) 99. URL: https://openscholarship.wustl.edu/law_globalstudies/vol5/iss1/5.

[9] M. Z. Costa, G. Guizzardi, J. P. A. Almeida, On capturing legal knowledge in ontology and process models combined, in: Legal Knowledge and Information Systems - JURIX 2022: The Thirty-fifth Annual Conference, Saarbrücken, Germany, 14-16 December 2022, volume 362 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2022, pp. 267–272. doi:10.3233/FAIA220478.

[10] J. L. Williams, Automation Is Not Hacking: Why Courts Must Reject Attempts to Use the CFAA as an Anti Competitive Sword, BUJ Sci. & Tech. L. 24 (2018) 416–450. URL: https://www.bu.edu/jostl/files/2018/10/Williams-FINAL.pdf.

[11] J. E. Christensen, The Demise of the CFAA in Data Scraping Cases, Notre Dame JL Ethics & Pub. Pol'y 34 (2020) 529–547. URL: https://jlepp.org/2020/07/26/volume-34-issue-2/.

[12] M. P. Goodyear, Circumscribing the spider: Trademark law and the edge of data scraping, U. Kan. L. Rev. 70 (2021) 295.

[13] G. Guizzardi, G. Wagner, Towards ontological foundations for agent modelling concepts using the unified fundational ontology (UFO), in: Int. Bi-Conference Workshop on Agent-Oriented Information Systems, Springer, 2004, pp. 110–124. doi:10.1007/11426714_8.

[14] J. P. A. Almeida, R. A. Falbo, G. Guizzardi, T. P. Sales, gUFO: A Lightweight Implementation of the Unified Foundational Ontology (UFO), 2020. URL: https://purl.org/nemo/doc/gufo.

[15] J. A. de Oliveira Lima, LexML – portal especializado em informação jurídica e legislativa,

in: L. Sayão, L. B. Toutain, F. G. Rosa, C. H. Marcondes (Eds.), Implantação e gestão de repositórios institucionais, Editora da Universidade Federal da Bahia, 2009, pp. 249–260.

[16] B. Oliveira, V. Santos, O. Belo, Pattern-based ETL conceptual modelling, in: Model and Data Engineering - Third International Conference, MEDI 2013, Amantea, Italy, September 25-27, 2013. Proceedings, volume 8216 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 237–248. doi:10.1007/978-3-642-41366-7_20.

[17] P. Vassiliadis, A survey of extract-transform-load technology, in: Integrations of Data Warehousing, Data Mining and Database Technologies - Innovative Approaches, Information Science Reference, 2011, pp. 171–199. doi:10.4018/978-1-60960-537-7.ch008.

[18] S. I. Bhat, T. Arif, M. B. Malik, A. A. Sheikh, Browser simulation-based crawler for online social network profile extraction, Int. J. Web Based Communities 16 (2020) 321–342. doi:10.1504/IJWBC.2020.111377.

[19] C. Sansone, G. Sperlí, Legal information retrieval systems: State-of-the-art and open issues, Inf. Syst. 106 (2022) 101967. doi:10.1016/j.is.2021.101967.

[20] F. A. D. Pinto, S. Lifschitz, E. H. Haeusler, A graph knowledge-base for auditing human resources public management, in: Anais do X Workshop de Computação Aplicada em Governo Eletrônico, SBC, 2022, pp. 61–72.

[21] M. A. Martija, J. Domoguen, P. C. Naval, How deep is your law? predicting associations between cases in philippine jurisprudence, in: TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, October 17-20, 2019, IEEE, 2019, pp. 886–891. doi:10.1109/TENCON.2019.8929425.

[22] H. Jamil, Semantic querying of knowledge rich legal digital libraries using prism, in: Legal Knowledge and Information Systems - JURIX 2022: The Thirty-fifth Annual Conference, Saarbrücken, Germany, 14-16 December 2022, volume 362 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2022, pp. 63–72. doi:10.3233/FAIA220449.

[23] G.-K. J. Li, C. V. Trappey, A. J. C. Trappey, A. A. S. Li, Ontology-based knowledge representation and semantic topic modeling for intelligent trademark legal precedent research, World Patent Information 68 (2022) 102098. doi:10.1016/j.wpi.2022.102098.

[24] Y. Ren, J. Han, Y. Lin, X. Mei, L. Zhang, An Ontology-Based and Deep Learning-Driven Method for Extracting Legal Facts from Chinese Legal Texts, Electronics 11 (2022). doi:10.3390/electronics11121821.

[25] G. Carnaz, V. B. Nogueira, M. Antunes, A Graph Database Representation of Portuguese Criminal-Related Documents, Informatics 8 (2021) 37. doi:10.3390/informatics8020037.

[26] J. D. Garofalakis, K. Plessas, A. Plessas, P. Spiliopoulou, Application of an ecosystem methodology based on legal language processing for the transformation of court decisions and legal opinions into open data, Inf. 11 (2020) 10. doi:10.3390/info11010010.

[27] M. Palmirani, F. Vitali, Akoma-ntoso for legal documents, Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management (2011) 75–100.

[28] F. A. C. Silva, J. E. L. Gayo, Legislative document content extraction based on semantic web technologies - A use case about processing the history of the law, in: The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings, volume 11503 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 558–573. doi:10.1007/978-3-030-21348-0_36.