Transparência pública automatizada a partir da gramática do diário oficial

Fernando Antonio D. G. Pinto[1], **Edward Hermann Haeusler**[1], **Sérgio Lifschitz**[1]

[1]Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
Rio de Janeiro – RJ – Brazil

{fpinto, hermann, sergio}@inf.puc-rio.br

Abstract. *Multiple sources and the growing volume of data lead to the need for new information management models. One of the main sources is in the official journals and its extraction, as well as its interpretation, present itself as an interesting alternative. These diaries are generally available in a closed format, which makes it difficult to extract information. This article presents DO2RDF, a new mechanism that helps transparency by building a knowledge base of public data according to the grammar of the official journal. The scope was restricted to publications involving labor relations in the last 9 years. The result was a Knowledge Base RDF of 24,745 public acts for SPARQL queries.*

Resumo. *Múltiplas fontes e o crescente volume de dados levam à necessidade por novos modelos de gestão da informação. Uma das principais fontes está nos diários oficiais e sua extração, bem como sua interpretação, se apresentam como uma interessante alternativa. Geralmente, estes diários são disponibilizados em formato fechado, o que dificulta a extração das informações. Este artigo apresenta DO2RDF, um novo mecanismo que auxilia a transparência ao construir uma base de conhecimento de dados públicos de acordo com a gramática do diário oficial. O escopo ficou restrito às publicações que envolvem movimentações de pessoal nos últimos 9 anos. O resultado foi uma Base de Conhecimento RDF de 24.745 atos públicos para consultas SPARQL.*

1. Introdução

Como consequência do desenvolvimento de novas soluções tecnológicas de apoio à produção, armazenamento e transmissão da informação, observamos um crescimento expressivo no volume de dados. Essa grande oferta de dados tem ocasionado uma revolução, nos meios tecnológicos e sociais, por novos modelos de gestão da informação. Um exemplo desse modelo é a participação da sociedade nas decisões governamentais com base na análise de dados públicos disponíveis na Internet. A participação social no acesso a estes dados já era prevista desde a constituição de 1988, mas limitado às ferramentas de TICs disponíveis à época.

De acordo com a Constituição Federativa do Brasil [Brasil 1995], no seu Art. 5, inciso *XXXIII*, é dito que todo órgão público é obrigado a ceder informações geradas por suas atividades ao indivíduo com interesse particular ou coletivo, sob pena de responsabilidade àqueles que não o cumprirem dentro dos prazos exigidos por lei.

Apenas em 2011, dada a grande demanda social por transparência nestes dados governamentais, o governo federal elaborou a Lei de Acesso à Informação (LAI) [Brasil 2011], provocando debates sobre o tratamento destas informações. Sabe-se que muitas informações ainda não estão disponíveis, para acesso online, por ainda não contarem com um mecanismo que automatize o processo de extração em certas fontes de informações. Dito isto, uma outra forma de promover a transparência é a extração dos atos públicos contidos nos Diários Oficiais.

Diante do apresentado, esta pesquisa tem como objetivo apresentar um mecanismo (*DO2RDF*) que auxilia a criação de base de conhecimento (*knowledge base*) das decisões públicas de acordo com a gramática[1] formal apresentada no Diário Oficial.

Sendo assim, *DO2RDF* trata da abertura e transparência de dados públicos a partir dos atos publicados em diários oficiais. Geralmente, estes documentos são disponibilizados no formato PDF, um formato fechado, o que dificulta seu processamento. Além de tornar os atos públicos mais transparentes, a base de conhecimento produzida está apoiada sobre os conceitos da Web Semântica e dados conectados [Isotani and Bittencourt 2015]. A ideia é permitir que pessoas ou máquinas tenham a oportunidade de fazer consultas a estes conjuntos de dados de forma estruturada na web.

Este projeto compõe uma outra iniciativa de análise e validação de normas jurídicas e decisões públicas do governo executivo [Haeusler et al. 2011, Delfino et al. 2017] que busca representar não só o conhecimento extraído da principal publicação do poder executivo, nas suas diversas esferas (municipal, estadual e federal), mas também representar conhecimento das normas e leis atuantes nos respectivos poderes.

Este artigo está dividido nas seguintes seções: Seção 1, esta introdução, apresentamos a nossa motivação da pesquisa. A Seção 2, referencial teórico e trabalhos correlatos. Seção 3, desafios e proposta da pesquisa com maiores detalhes sobre a ferramenta *DO2RDF*. A Seção 4, apresentamos os resultados gerados: artefatos e consulta a base de conhecimento, e finalmente a Seção 5 onde faremos uma breve conclusão da pesquisa e trabalhos futuros.

Uma importante observação é que, devido ao que dispõe a Lei Geral de Proteção de Dados (LGPD) [Brasil 2018, Brasil 2019], alguns dados apresentados nas imagens deste trabalho foram devidamente anonimizados.

2. Referencial teórico e trabalhos correlatos

2.1. Resource Description Framework

O *RDF* (*Resource Description Framework*) é um modelo de dados (metadados) utilizado para representação de informação de modo a facilitar a busca por recursos na Web.

Segundo [Group 2014, Barzdins et al. 2019], *RDF* incrementa a estrutura de *Links* da Web ao usar *URIs* para nomear o relacionamento entre os recursos na web, disponíveis na forma de uma tripla de elementos <*sujeito*>, <*predicado*> e <*objeto*>.

[1]Uma gramática é um conjunto de regras necessárias para o processo de geração ou reconhecimento de todas as palavras válidas em uma certa linguagem.

Esse modelo *RDF* forma um grafo (Grafo *RDF*) direcionado e rotulado onde as extremidades (nós) representam recursos (Sujeito ou Objeto) que são relacionados por um predicado (aresta) [Isotani and Bittencourt 2015]. Recursos podem ser vistos como qualquer informação, por exemplo, um documento, uma pessoa, etc. Neste caso, a cada recurso é atribuído um elemento identificador *IRI* (*Internationalized Resource Identifier*).

Com esse simples modelo, o *RDF* permite que dados estruturados e semiestruturados sejam combinados, expostos e compartilhados por diferentes tecnologias na web.

Neste trabalho, optamos por construir nossa base de conhecimento na forma de triplas *RDF*, assim como apresentado em [Amato et al. 2008, Najmi et al. 2016].

2.2. Processamento de Linguagem Natural

Técnicas de Processamento de Linguagem Natural (PLN) estão ligadas a problemas de reconhecimento e interpretação automática de informações contidas em documentos onde sua base é a linguagem natural [Friedman et al. 2013]. Para isto, modelos estatísticos e técnicas de aprendizagem de máquina são utilizados para criar sistemas capazes de processar rapidamente as expressões linguísticas presentes nestes documentos.

Neste contexto, a utilização de grandes volumes de dados (*Big Data*) se torna um desafio para a área de PLN. Mais recentemente, observamos um crescente interesse por técnicas como *Small Data-set* e *Transfer Learning* [Barman et al. 2019] que tornam o processo de aprendizagem de modelos PLN mais ágeis, exigindo poucos recursos computacionais.

Umas das tarefas utilizadas na PLN é de detectar e categorizar palavras, ou sequência de palavras, como entidades (*Named-entity Recognition - NER*). Um dos nossos trabalhos futuros é a detecção e categorização automática dos atos do Diário Oficial utilizando estas técnicas de reconhecimento de entidades nomeadas.

2.3. A lógica de descrição iALC

Lógica de Descrição (*Description Logic*) é um formalismo utilizado para representação de conhecimento. Essencialmente, *DL* possui três componentes: Indivíduos (constante) que representam entidades de um certo domínio; Conceitos (predicados unários) que são propriedades dadas aos Indivíduos; Papéis (predicados binários) que são as relações entre os Indivíduos [Baader et al. 2008].

Parte do alfabeto de uma DL consiste em:
– Conjuntos de nomes de Indivíduos, Conceitos e Papéis.
- $\sqcup$ representa a conjunção de Conceitos.
- $\sqcap$ representa a disjunção de Conceitos.
- $\sqsubseteq$ representa a inclusão de Conceitos. A operação $C \sqsubseteq D$ indica que o Conceito C está incluso no Conceito D.
- $\forall$ representa a restrição universal de Conceitos. Assim, $\forall R.C$ representa a restrição universal do Conceito C sob o Papel R.
- $\exists$ representa a restrição existencial de Conceitos utilizando Papéis.
- $\neg$ representa o complemento de Conceitos.
- : um operador que gera asserções do tipo $a : C$ (asserção de Conceitos) e $(a, b) : R$ (asserção de Papéis). Para C (Conceito), R (Papel), e a e b (Indivíduos), estas operações

indicam que os Indivíduos "populam" o Conceito ou Papel a que faz referência.

A lógica *iALC*, derivada de *ALC*, é uma lógica de descrição de caráter intuicionista criada para lidar com textos jurídicos. Na seção 5 será apresentada uma formalização utilizando lógica *iALC* para um caso particular. Para mais informações sobre *iALC* e lógica descrição, sugerimos a leitura do artigo [Haeusler et al. 2011].

2.4. Trabalhos correlatos

Em nossas pesquisas foram encontrados dois trabalhos que apresentam ferramentas de recuperação da informação, mas que se limitam apenas ao processo de leitura das informações contidas em diários oficiais.

Em [Rodríguez, M., Dantas Bezerra, B 2019] os autores utilizam técnicas de Processamento de Linguagem Natural [Friedman et al. 2013] para reconhecer Entidades Nomeadas em portarias do Diário Oficial. Eles utilizam os recursos disponíveis na plataforma *NLTK (Natural Language Toolkit)* para etapas do processo de tokenização até o reconhecimento destas entidades. Uma limitação deste trabalho é que os autores apresentam uma ferramenta que reconhece apenas os nomes dos agentes públicos (servidores) em portarias.

Um outro trabalho [Junior et al. 2018] utiliza técnicas de *Data Mining* para recuperar informações contidas no Diário Oficial do Governo de Pernambuco. Apesar dos esforços, os autores concordam que "caso os órgãos desejem um algoritmo com melhores resultados, é necessário realizar uma padronização mínima dos Diário Oficiais para que a extração seja mais eficiente". Isso evidencia a complexidade no tratamento dos dados contidos no diário oficial de Pernambuco. Neste caso, faz-se necessário um estudo sobre outras estratégias de recuperação de informação.

3. Desafios e Proposta de Sistema

Nossa proposta é a extração de dados de atos públicos publicados no Diário Oficial Municipal (DOM). Para evidenciar as contribuições da ferramenta, iremos nos restringir aqui à Prefeitura do Rio de Janeiro[2], sem prejuízo de falta de generalização. O escopo ficou restrito às publicações que envolvem nomeações, exonerações e designações para cargos comissionados, funções gratificadas e empregos públicos (este último das empresas públicas), dos últimos 9 anos. Dada a grande dificuldade de processamento de arquivos PDF, a ideia é montar uma base histórica de movimentação de pessoal para estes cargos de livre nomeação, o que além de agilizar as consultas dará maior transparência aos setores de auditoria e, consequentemente, a sociedade.

O primeiro desafio foi capturar todos os diários oficiais de um determinado período. Para esta atividade, e outras deste projeto, foram utilizadas a linguagem *Python*, versão 3, como *backend* da ferramenta de extração e geração das triplas *RDF*.

Inicialmente foi desenvolvido um *script* para *download* dos diários utilizados no projeto com o método *requests* do *Python*. No total foram recuperados e tratados 1.126 diários oficiais. De posse destes documentos, a segunda etapa foi construir o extrator de informações contidas em cada diário oficial. Neste processo, foi utilizada a biblioteca *RE*[3]

[2]https://doweb.rio.rj.gov.br/

[3]https://docs.python.org/3/library/re.html

Tabela 1. Scripts com expressão regular para um padrão de ato do tipo "Dispensar".

Padrão	Dispensar Servidor
Ato	*Dispensar[,|\s]*
Nome	(?P<nome>[A-ZÉÁÍÓÚÇÃÊÔÕÀÜ\s]+)
Matrícula	(?P<matricula>[0-9\./-]+)
Cargo Efetivo	(?P<cargoEfetivo>[A-ZÉÁÍÓÚÇÃÊÔÕÀÜa-záêéóíçãâôú\-\s]+)
Dia	(?P<dia>[0-9]+)
Mês	(?P<mes>[J | j]aneiro | [F | f]evereiro | [...] | [D | d]ezembro)
Ano	(?P<ano>[0-9]+)
Cargo Comissionado	(?P<cargo>[A-ZÉÁÍÓÚÇÃÊÔÕÀÜa-záêéóíçãâôú\-\s]+)
Símbolo	(?P<simbolo>[A-Z\-0-9\/\s]+)

do *Python* e um conjunto de padrões de expressões regulares para compor um conjunto de regras baseadas na gramática destes documentos PDF (diários). Naturalmente, algoritmos de expressões regulares tendem a ter seu comportamento guloso e extrair dados tornou-se uma atividade de grande custo dado o grande número de padrões de publicação presentes nos diários. Esta falta de padronização fez com que fosse definida uma arquitetura capaz de interpretar cada padrão de forma única, evitando assim sobreposição de padrões.

Para exemplificar a utilização de nossa ferramenta, decidimos tratar um subconjunto de informações de movimentação de pessoal nos cargos de livre nomeação, que são cargos que não necessitam de concurso público para seu efetivo exercício, cabendo ao gestor a escolha dos profissionais que devem ocupá-lo. Baseado neste escopo, foram desenvolvidos os extratores de padrões utilizando técnicas de expressões regulares aplicadas a cada documento *PDF*.

Expressão Regular (ER) é uma notação para especificar padrões de lexemas. Sua construção sintática é composta de símbolos atômicos (caracteres), união, concatenação e fecho de *Kleene* de outras expressões regulares. Leitores não familiarizados com o conceito e terminologia relacionados às expressões regulares podem consultar o livro [Aho et al. 2006].

A Tabela 1, apresenta os grupos de uma expressão regular, onde o ato do gestor foi de "Dispensar" o servidor público. Podemos observar as informações a serem recuperadas: nome do servidor, matrícula, cargo efetivo, dia, mês, ano, cargo comissionado e seu símbolo (código) na folha de pagamento.

Até o presente momento da escrita deste trabalho, foram identificados 33 (trinta e três) padrões em expressões regulares nos atos públicos.

O resultado da execução dos padrões de ER foi a extração das informações contidas em 24.745 atos públicos. Para facilitar o processo de análise dos dados, a ferramenta gera um arquivo de auditoria (Figura 1) com as principais informações recuperadas. Este registro, que apresenta dados agrupados por data de publicação do diário, contém atos do tipo"Nomear", "Exonerar" e "Designar" servidores. Por exemplo, como destacado na imagem, houve a publicação de dois atos para o servidor Antonio Barbosa: uma exoneração do cargo comissionado de Assessor II (DAS-08) e sua nomeação para o

cargo comissionado de Assessor I (DAS-09). Ambas no dia 11/01/2013.

```
1  (PUC-RIO/TECMF)::PROCESSAMENTO DO DIÁRIO:: ANO: 26 No.:000201 TIPO:NORMAL * RIO DE JANEIRO * ARQUIVO: 1974.PDF SEQ.: 0001 26/02/2021 21:16:22
2
3  RESOLUÇÃO   0418 11/01/2013 NOMEAR     ARETUSA  PAULA                    01/01/2013 DIRETOR I              DAS-09 CC
4  RESOLUÇÃO   0419 11/01/2013 NOMEAR     LILIANE  CESAR                    01/01/2013 ASSISTENTE TÉCNICO     DAS-07 CC
5  RESOLUÇÃO   0420 11/01/2013 NOMEAR     JUSSARA                           01/01/2013 DIRETOR IV             DAS-06 CC
6  RESOLUÇÃO   0421 11/01/2013 EXONERAR   60/25    -6  ANTONIO  BARBOSA      02/01/2013 ASSESSOR II            DAS-08 CC
7  RESOLUÇÃO   0422 11/01/2013 NOMEAR     60/25    -6  ANTONIO  BARBOSA      02/01/2013 ASSESSOR I             DAS-09 CC
8  RESOLUÇÃO   0425 11/01/2013 NOMEAR     LUIZ  PIRES                       30/12/2012 ASSISTENTE I           DAS-06 CC
9  RESOLUÇÃO   0426 11/01/2013 EXONERAR   PAULA  CARMO                       01/01/2013 ASSISTENTE I           DAS-06 CC
10 RESOLUÇÃO   0427 11/01/2013 EXONERAR   PAULO  PINTO                       02/01/2013 ASSISTENTE I           DAS-06 CC
11 RESOLUÇÃO   0428 11/01/2013 NOMEAR     RENATO  LAGRIMANTE                 02/01/2013 ASSISTENTE I           DAS-06 CC
12 RESOLUÇÃO   0435 11/01/2013 NOMEAR     MARIA  SANSON                      02/01/2013 ASSESSOR II            DAS-08 CC
13 RESOLUÇÃO   0440 11/01/2013 EXONERAR   ALEXANDRE  BAPTISTA                01/01/2013 ASSESSOR III           DAS-07 CC
14 RESOLUÇÃO   0441 11/01/2013 EXONERAR   ALEXANDRE  OLIVEIRA                01/01/2013 ASSISTENTE I           DAS-06 CC
15 RESOLUÇÃO   0443 11/01/2013 EXONERAR   HELIO  FURTADO                     01/01/2013 COORDENADOR I          DAS-09 CC
16 RESOLUÇÃO   0445 11/01/2013 EXONERAR   SUEDENI  OLIVEIRA                   01/01/2013 ASSESSOR II            DAS-08 CC
17 RESOLUÇÃO   0449 11/01/2013 NOMEAR     SIMONE  GUILHERMINO                01/01/2013 ASSESSOR III           DAS-07 CC
18 RESOLUÇÃO   0450 11/01/2013 NOMEAR     MARCIA  SOUSA                      01/01/2013 ASSESSOR III           DAS-07 CC
19 RESOLUÇÃO   0451 11/01/2013 NOMEAR     ZORAIDE  COSTA                     01/01/2013 ASSISTENTE I           DAS-06 CC
20 RESOLUÇÃO   0361 07/01/2013 NOMEAR     ISABEL                             01/01/2013 ASSESSOR CHEFE         DAS-08 CC
21 RESOLUÇÃO   0361 07/01/2013 NOMEAR     MARIA  SILVA                       01/01/2013 ASSESSOR CHEFE         DAS-08 CC
22 RESOLUÇÕES  0095 00/00/2013 NOMEAR     VITOR  ALMEIDA                     04/01/2013 DIRETOR IV             DAS-06 CC
23 RESOLUÇÕES  0096 00/00/2013 DESIGNAR   VANESSA  FERREIRA                   XX/XX/XXXX ASSISTENTE II          DAI-06 FG
24 RESOLUÇÕES  0100 00/00/2013 DESIGNAR   ANA  HORTA                         XX/XX/XXXX COORDENADOR PEDAGÓGICO DAI-06 FG
25 RESOLUÇÕES  0101 00/00/2013 DESIGNAR   FERNANDA  SILVA                    XX/XX/XXXX DIRETOR-ADJUNTO        DAI-06 FG

EXONERAR  60/25    -6  ANTONIO  BARBOSA
NOMEAR    60/25    -6  ANTONIO  BARBOSA
```

Figura 1. Fragmento do arquivo de auditoria com informações recuperadas dos diários.

Esta etapa de auditoria é um passo anterior a etapa de persistência dos dados no banco de dados, podendo o usuário desativá-la a qualquer momento.

Finalmente podemos apresentar a arquitetura da ferramenta *DO2RDF* (Figura 2), onde:

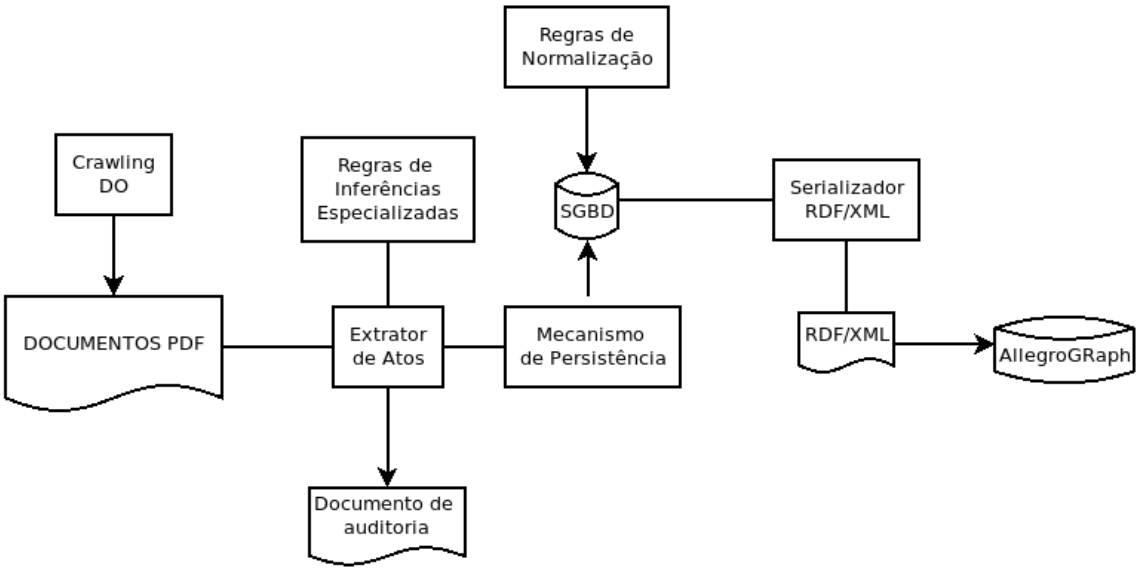

Figura 2. Arquitetura da ferramenta *DO2RDF*.

Os módulos ficam assim representados:

- **Módulo de *Crawling*** - É responsável por recuperar os Diários Oficiais, de forma automática, do repositório da prefeitura do Rio de Janeiro.

- **Módulo Extrator de Atos** - Conjunto de regras que utiliza expressões regulares para extrair conteúdos dos diários oficiais. Também é responsável por um processo de geração de arquivos de auditoria.
- **Módulo de Regras de Inferências Especializadas** - Módulo que contém as regras gramaticais para cada ente público (governos estaduais, municipais e federal). Opera em conjunto com o Módulo Extrator de Atos.
- **Mecanismo de Persistência de dados** - Responsável por manter uma interface de gravação dos dados extraídos com o Banco de Dados *PostgreSQL* [Douglas and Douglas 2003].
- **Módulo de Regras de Normalização** - Responsável pela leitura e aplicação de regras de normalização nos dados previamente armazenados.
- **Módulo de Serializador** - Este módulo faz a leitura do banco de dados e serialização *RDF*. Neste trabalho optamos pelo formato *RDF/XML*.

O código fonte e documentação do projeto podem ser obtidos em: https://github.com/fernandoantoniodantas/DO2RDF.

4. Resultados

Nesta seção, apresentaremos os resultados do processo de extração das informações contidas nos diários e o passo de triplificação em dados *RDF/XML* para ser executado em um ambiente de consultas em linguagem *SPARQL* [Buil-Aranda et al. 2013].

Como apresentado na seção anterior, de posse das informações previamente validadas, foi realizada uma carga em um banco de dados *PostgreSQL* versão 12.1 em ambiente Ubuntu GNU/Linux.

Com o banco de dados devidamente "populado", iniciamos o processo de triplificação das informações. Para esta etapa foi utilizada a biblioteca *RDFLib* [Team 2013] para *Python*. A biblioteca possui interfaces que tornam simples e facilitam a implementação dos nós *RDF*. Como opção, ela inclui analisadores e serializadores para *RDF/XML*, *N3*, *NTriples*, *N-Quads*, *Turtle*, *TriX*, *RDFa* e *Microdata*. Ela implementa uma interface *Graph* ao qual podemos armazenar informações dos grafos em memória ou de armazenamento persistente. Também é possível executar consultas e atualizações na linguagem *SPARQL*.

Neste trabalho optamos por fazer a serialização dos dados no formato *RDF/XML* para carga em um ambiente de consultas. Neste caso, por afinidade, foi utilizado o *AllegroGraph* [AllegroGraph 2019]. Um passo importante do projeto foi quanto a ontologia a ser utilizada na definição dos significados aos conteúdos do Diário Oficial. Para este momento, como prova básica de conceito, optamos por definir uma ontologia adaptada e genérica em *"Friend of a Friend" (FOAF)* [Brickley and Miller 2014].

O Código 1 apresenta uma porção do arquivo *DO2RDF.rdf*, resultado do processo de serialização do *RDFLib*, para carga no *AllegroGraph*.

```
1 <?xml version="1.0" encoding="utf-8"?>
2 <rdf:RDF
3   xmlns:foaf="http://xmlns.com/foaf/0.1/"
4   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
5 >
6   <foaf:Funcionario rdf:nodeID="N7654fa3513984ea89c0ade256a971939">
7     <foaf:acao>NOMEAR</foaf:acao>
```

```
8     <foaf:dataPublicacao>2017-01-19</foaf:dataPublicacao>
9     <foaf:tipoCargo>CC</foaf:tipoCargo>
10    <foaf:simbolo>DAS-06</foaf:simbolo>
11    <foaf:dataEfeito>2017-01-01</foaf:dataEfeito>
12    <foaf:matricula>60/210917-1</foaf:matricula>
13    <foaf:cargo>ASSISTENTE I</foaf:cargo>
14    <foaf:nome>MARIA CRISTINA DOS SANTOS BASTOS</foaf:nome>
15  </foaf:Funcionario>
```

Código 1. Fragmento da serialização em *RDF/XML.*

Para consultas *SPARQL* foi utilizado o *AllegroGraph* na versão 6.6.0 [AllegroGraph 2019], em ambiente virtual com Ubuntu GNU/Linux. Na sequência, foi criado o repositório *DO2RDF* para receber a carga do arquivo *DR2RDF.rdf*. Para ilustrar o uso da ferramenta, mostramos na Figura 3 o resultado de uma consulta *SPARQL* para os casos de nomeação para todos os cargos que iniciam com a expressão "ADM", cargo do tipo "CC"[4] e ação "NOMEAR". Neste exemplo as informações do nome do servidor, cargo, tipo do cargo e ação (ato de nomear ou exonerar) foram recuperadas.

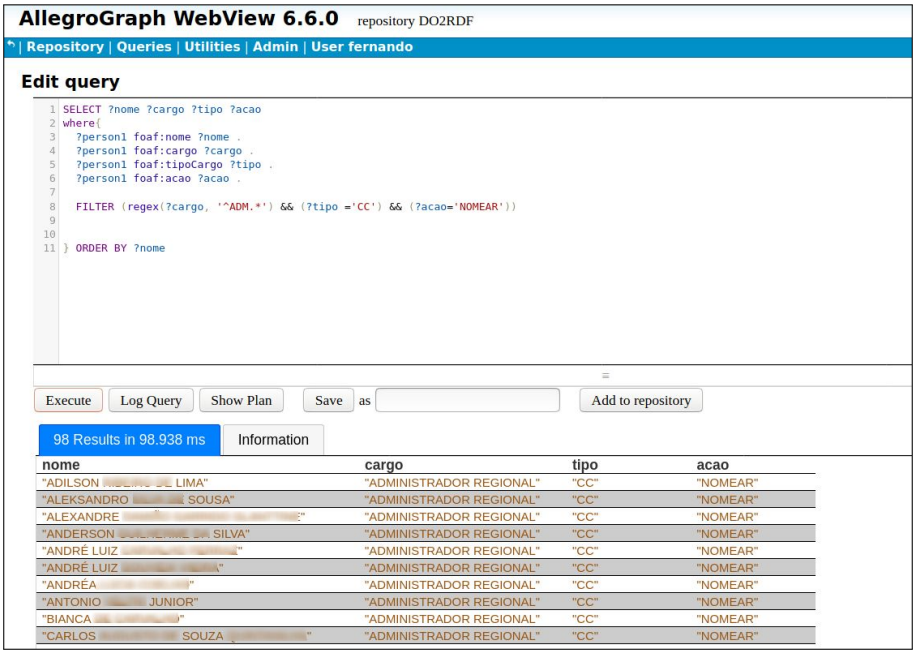

Figura 3. Consulta das nomeações para cargos com iniciais "ADM".

5. Conclusões e trabalhos futuros

Apresentamos como a transparência pública poderá ser dada por processos automatizados de leitura, extrações de informações e publicidade, em Base de Conhecimento, dos atos públicos disponíveis em Diários Oficiais.

De posse dos diários oficiais, e definida uma arquitetura de estilos de padrões, a técnica de extração com expressões regulares apresentou ser simples e eficiente.

Cabe observar que a utilização da biblioteca *RDFLib* apresentou ser pouco flexível para criação de ontologias genéricas. Inicialmente tentamos modificar seus *scripts* para

[4]Sigla de Cargo Comissionado.

que fossem incorporados novos conceitos. Uma solução mais simples foi especializar a ontologia *FOAF* para as necessidades deste trabalho. Superada esta parte, a serialização em formato *RDF/XML* se tornou bastante eficiente para o propósito da pesquisa. Consultas *SPARQL* foram realizadas para demonstrar a razoabilidade da ferramenta.

Também está em andamento uma pesquisa e desenvolvimento da viabilidade da técnica com *PLN* para extração de informações contidas no Diário Oficial. Para isto, foi construído um modelo que aplica os conceitos de *Small Dataset* e *Transfer Learning* na tarefa de reconhecimento das entidades nomeadas.

Um exemplo deste trabalho é apresentado na Figura 4 onde as entidades que representam o tipo de ato (resolução), número e data da resolução, a ação (nomear), nome do servidor, bem como sua matrícula e cargo foram devidamente identificadas em nosso modelo *PLN*.


RESOLUÇÃO**RESOLUCAO** "P"
Nº 326 **NUMRESOLUCAO** DE 7 **DIARESOLUCAO** DE JANEIRO **MESRESOLUCAO** DE 2013 **ANORESOLUCAO.** O SECRETÁRIO CHEFE DA SECRETARIA MUNICIPAL DA CASA CIVIL, no uso das atribuições que lhe são conferidas pela legislação em vigor, RESOLVE Nomear **ACAO** ANA [redacted] ALMEIDA NIGRO **SERVIDOR**, matrícula 60/[redacted]18-9 **MATRICULA**, para exercer o Cargo em Comissão de Assessor III **CARGO**, símbolo DAS-07, código 029846, da XVI Administração Regional, da Coordenadoria Especial da Área de Planejamento 4, da Subsecretaria de Integração das Áreas de Planejamento, da Secretaria Municipal de Governo.

Figura 4. Entidades classificadas em um ato de nomeação.

No momento, a pesquisa com *PLN* está em fase de análise experimental e o intuito é avaliar (acurácia e precisão) a ferramenta em contraste ao uso de expressões regulares, para um mesmo cenário. Preliminarmente, dada as características do Diário Oficial, a utilização apenas de *PLN*, para a extração das informações, não se apresentou ser tão efetiva. Existem muitos padrões combinados nas edições dos diários e esta diversidade torna o processo com *PLN* propenso a significativas taxas de erro de classificação. De forma ainda preliminar, mas não definitiva, estamos propensos a declinar por um modelo hibrido, com uso de expressões regulares e *PLN*. A expressão regular assegura que o Diário seja segmentado em grupos para análise do nosso modelo *PLN*. Em resumo, a tarefa de reconhecimento das entidades nomeadas será aplicada a grupos e não a todo o diário. Acreditamos que com isso a ferramenta consiga ganhos em sua capacidade de generalização e precisão.

Com foco na segmentação do diário em grupos, esta proposta será extrair outras informações públicas como pagamento de diárias, licitações, pagamento de custeios, participações de servidores em comissões remuneradas, etc. Todas estas informações são passíveis de extrações, bastando definir seus padrões de publicação, na forma de expressões regulares, e as entidades nomeadas.

Um outro ponto a ser explorado é a definição de uma ontologia para o diário oficial. Neste trabalho foi utilizado um modelo genérico de ontologia para fins de execução da posposta. Para uma atividade fim, faz-se necessário definir o significado de cada dado presente no modelo *RDF*. Isso facilitará o processo de captura e processamento dos dados armazenados nas triplas *RDF*, por máquinas ou usuários, através de consultas *SPARQL*.

Como já dito na introdução deste trabalho, este projeto faz parte de uma iniciativa

de análise e validação da normas jurídicas de atos e decisões governamentais. Uma das iniciativas foi a definição de uma versão Intuicionista de uma Lógica de Descrição (DL) chamada *iALC*, apresentada na seção 2.3. Por se tratar de uma lógica de descrição temos que uma futura implementação de um raciocinador gerará construtivamente uma base de conhecimento de leis nesta linguagem. Como exemplo, apresentamos a modelagem de uma pequena parte da norma jurídica da Prefeitura do Rio de Janeiro, referente a manutenção dos cargos do poder executivo.

$$Lei16801991 = Art26 \sqcap Art27 \sqcap \ldots \sqcap Art_{n-1} \sqcap Art_n \tag{1}$$

Na Fórmula 1, é demonstrado em *iALC* a composição de artigos da Lei nº 1680, de 26 de março de 1991 que define a criação dos cargos:

$$\begin{aligned} &CargoPublico \sqsubseteq Art26 \\ &CargoComissionado \sqcap CargoEfetivo \sqcap FuncaoGratificada \sqsubseteq CargoPublico \\ &CargoComissionado \sqsubseteq \neg FuncaoGratificada \\ &FuncaoGratificada \sqsubseteq \neg CargoComissionado \end{aligned} \tag{2}$$

As regras funcionais, previstas no Art. 26, que definem o quadro permanente de pessoal do Poder Executivo (cargos efetivos, cargos comissionados, funções gratificadas) são modeladas na Fórmula 2. Por ela, não é possível que um servidor ocupante de cargo em comissão possa receber uma função gratificada, assim como o inverso também não é possível.

Também neste mesmo artigo é definido que os cargos comissionados são divididos em cargos da Administração Direta e Indireta, Fórmula 3:

$$\begin{aligned} &CargoComissionadoAD \sqsubseteq CargoComissionado \\ &CargoComissionadoAI \sqsubseteq CargoComissionado \end{aligned} \tag{3}$$

Nesta mesma lei, e de acordo com o Art. 27, é vedada a transferência de cargos de provimento efetivo e de provimento em comissão e funções gratificadas da administração direta para a administração indireta sem lei que a determine ou autorize. Essa característica é modelada nas Fórmulas 4 e 5:

$$\begin{aligned} &a : CargoComissionadoAD \\ &b : CargoComissionadoAI \\ &ad : OrgaoAD \\ &ai : OrgaoAI \end{aligned} \tag{4}$$

$$\begin{aligned} &a\mathbf{T}ad, \text{ similar a forma prefixa } \mathbf{T}(a, ad) \\ &b\mathbf{T}ai, \text{ similar a forma prefixa } \mathbf{T}(b, ai) \end{aligned} \tag{5}$$

Onde **T** é uma Papel que normatiza a transferência de servidores entre órgãos da administração direta ou indireta.

Especificações da normativa legal em *iALC* serão traduzidas para uma linguagem de consulta em Base de Conhecimento (*BC*) das normas legais bem como na *BC* com os registros recuperados dos diários oficiais. Assim, uma máquina poderá inferir sobre a real aplicação da legislação, validando ou não os atos em decorrência da lei.

A Figura 5 é um exemplo deste processo de inferência na base de conhecimento. Neste caso, apresentamos uma prova utilizando o sistema dedutivo para *iALC* [Haeusler et al. 2010] de que a servidora Ana não pode assumir uma Função Gratificada ao mesmo tempo em que já possui Cargo em Comissão. Por questões de legibilidade, definimos os conceitos "CargoComissionado" e "FuncaoGratificada" por suas respectivas siglas: "CC" e "FG".

$$\dfrac{\dfrac{ana : CC \qquad CC \sqsubseteq \neg FG}{ana : \neg FG} \qquad [ana : FG]}{\dfrac{\bot}{\neg(ana : FG)}}$$

Figura 5. Demonstração da impossibilidade do acúmulo de cargos.

Com isto, no futuro uma máquina poderá inferir automaticamente sobre uma base de conhecimento e detectar eventuais inconformidades com a norma legal.

Referências

Aho, A. V., Lam, M. S., Sethi, R., and Ullman, J. D. (2006). *Compilers: Principles, Techniques, and Tools (2nd Edition)*. Addison-Wesley Longman Publishing Co., Inc., USA.

AllegroGraph (2019). Allegrograph - The Enterprise Knowledge Graph. `https://allegrograph.com`. Acessado: 11-12-2019.

Amato, F., Mazzeo, A., Penta, A., and Picariello, A. (2008). Building rdf ontologies from semi-structured legal documents. In *2008 International Conference on Complex, Intelligent and Software Intensive Systems*, pages 997–1002.

Baader, F., Horrocks, I., and Sattler, U. (2008). Chapter 3 description logics. In van Harmelen, F., Lifschitz, V., and Porter, B., editors, *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 135–179. Elsevier.

Barman, R., Deshpande, S., Agarwal, S., Inamdar, U., Devare, M., and Patil, A. (2019). Transfer learning for small dataset. In *Proceedings of National Conference on Machine Learning*, pages 132–137.

Barzdins, G., Gosko, D., Barzdins, P. F., Lavrinovics, U., Bernans, G., and Celms, E. (2019). Rdf* graph database as interlingua for the textworld challenge. In *2019 IEEE Conference on Games (CoG)*, pages 1–2.

Brasil (1995). Emenda Constitucional nº 9, de 9 de novembro de 1995. *Diário Oficial [da] República Federativa do Brasil*, 59:1966.

Brasil (2011). Lei nº. 12.527. Lei de Acesso à Informação. *Diário Oficial [da] República Federativa do Brasil*.

Brasil (2018). Lei nº. 13.709. Lei Geral de Proteção de Dados pessoais. *Diário Oficial [da] República Federativa do Brasil*.

Brasil (2019). Lei nº. 13.853. Lei Geral de Proteção de Dados pessoais. *Diário Oficial [da] República Federativa do Brasil*.

Brickley, D. and Miller, L. (2014). FOAF Vocabulary Specification 0.99. `http://www.foaf-project.org/`. Acessado: 03-12-2019.

Buil-Aranda, C., Hogan, A., Umbrich, J., and Vandenbussche, P.-Y. (2013). Sparql web-querying infrastructure: Ready for action? In Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J. X., Aroyo, L., Noy, N., Welty, C., and Janowicz, K., editors, *The Semantic Web – ISWC 2013*, pages 277–293, Berlin, Heidelberg. Springer Berlin Heidelberg.

Delfino, P., Cuconato, B., Haeusler, E. H., and Rademaker, A. (2017). Passing the brazilian OAB exam: Data preparation and some experiments. In Wyner, A. Z. and Casini, G., editors, *Legal Knowledge and Information Systems - JURIX 2017: The Thirtieth Annual Conference, Luxembourg, 13-15 December 2017*, volume 302 of *Frontiers in Artificial Intelligence and Applications*, pages 89–94. IOS Press.

Douglas, K. and Douglas, S. (2003). *PostgreSQL*. New Riders Publishing, USA.

Friedman, C., Rindflesch, T. C., and Corn, M. (2013). Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of Biomedical Informatics*, 46(5):765–773.

Group, R. W. (2014). Resource Description Framework (RDF). `https://www.w3.org/RDF/`. Acessado: 11-10-2019.

Haeusler, E. H., d. Paiva, V., and Rademaker, A. (2010). Intuitionistic logic and legal ontologies. *Conference: Legal Knowledge and Information Systems - JURIX 2010*.

Haeusler, E. H., d. Paiva, V., and Rademaker, A. (2011). Intuitionistic description logic and legal reasoning. In *2011 22nd International Workshop on Database and Expert Systems Applications*, pages 345–349.

Isotani, S. and Bittencourt, I. (2015). *Dados Abertos Conectados: em Busca da Web do Conhecimento*. Novatec.

Junior, R., Melo, W., Fagundes, R., and Maciel, A. (2018). Extração de informação e mineração de dados no diário oficial de pernambuco. *Revista de Engenharia e Pesquisa Aplicada*, 3.

Najmi, E., Malik, Z., Hashmi, K., and Rezgui, A. (2016). Conceptrdf: An rdf presentation of conceptnet knowledge base. In *2016 7th International Conference on Information and Communication Systems (ICICS)*, pages 145–150.

Rodríguez, M., Dantas Bezerra, B (2019). Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias). *Revista de Engenharia e Pesquisa Aplicada*, 5(1):67–77.

Team, R. (2013). RDFLib. `https://rdflib.readthedocs.io/en/stable/#`. Acessado: 13-11-2019.