



Fernando Antonio Dantas Gomes Pinto

Extração e Triplificação de Atos Públicos

INF2102 - Projeto Final de Programação.

:



Fernando Antonio Dantas Gomes Pinto

Extração e Triplificação de Atos Públicos

INF2102 - Projeto Final de Programação.

Prof. Marcos Kalinowski
Departamento de Informática – PUC-Rio

Prof. Edward Hermann Haeusler
Departamento de Informática – PUC-Rio

Rio de Janeiro, Julho 8th, 2020

Sumário

1	Introdução	10
2	Proposta	13
3	Documentação e Testes	19
4	Resultados	26
5	Conclusões e trabalhos futuros	30
6	Referências bibliográficas	31

Lista de Figuras

Figure 1.1	Camadas de protocolos da Web Semântica.	11
Figure 2.1	Porção do arquivo de auditoria com as informações extraídas dos diários.	17
Figure 3.1	Diagrama de Classes do DO2RDF.	19
Figure 3.2	Modelo do banco de dados.	20
Figure 3.3	Arquitetura da ferramenta <i>DO2RDF</i> .	21
Figure 3.4	Diário Oficial do Rio de Janeiro.	23
Figure 3.5	Ato de nomeação e exoneração de servidores.	24
Figure 3.6	Ato de designar servidor.	24
Figure 3.7	Resultado dos testes unitários.	25
Figure 4.1	Consulta das nomeações para cargos com iniciais "ADM".	28
Figure 4.2	Consulta de exonerações.	29

Lista de Tabelas

Table 2.1	Quantitativo de tipos de atos extraídos nos diários.	18
-----------	--	----

List of Algorithms

Lista de Códigos

Code 1	Método <i>Request</i> para download dos diários.	13
Code 2	<i>Scripts</i> com expressões regulares para atos "Dispensar".	15
Code 3	Expressão regular para tratar datas ausentes ou inconsistentes	17
Code 4	Processo de serialização em <i>RDF/XML</i>	26
Code 5	Processo de serialização em <i>RDF/XML</i>	27

1

Introdução

Dado o rápido desenvolvimento de novas tecnologias de apoio à informação, o volume de dados gerados tem também crescido de forma expressiva. 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 da Internet. A participação social no acesso a estes dados já era prevista desde a constituição de 1988, mas limitado às ferramentas de TICs disponíveis a época. De acordo com a Constituição Federativa do Brasil (2), no seu Art. 5, inciso XXXIII, é dito que todo órgão público é obrigado a ceder informações geradas por suas atividades a indivíduo com interesse particular ou coletivo, sob pena de responsabilidade àqueles que não o cumprirem dentro dos prazos exigidos por lei, respeitando aquelas informações que sejam classificadas como imprescindível à segurança da sociedade e do Estado. Apenas em 2011, dada a grande demanda social por transparência nestes dados governamentais, o governo federal elaborou a Lei de Acesso à Informação (3), provocando debates sobre o tratamento destas informações. Hoje, com o decreto 7.724 de 2012 (BRASIL, 2012), que regulamenta a LAI, o governo federal exige que os entes governamentais publiquem na Internet um conjunto mínimo de informações governamentais. Sabe-se que muitas informações ainda não estão disponíveis para acesso online, por ainda não contar com um mecanismo que automatize o processo de extração de certas informação. Uma outra forma de promover transparência é a extração e publicação dos atos públicos contidos nos Diários Oficiais. A publicidade dos atos públicos em Diários Oficiais é interpretada como um ato legal obrigatório nas esferas públicas, além do fato motivador aos princípios legais da transparência.

Sendo assim, este trabalho pretende apresentar o *DO2RDF*, uma ferramenta que tem como proposta a 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 o manuseio das informações produzidas por entes públicos.

O escopo inicial deste projeto ficou restrito aos processos de movimentação de servidores em cargos comissionados, funções gratificadas e empregos de confiança, dos últimos 7 anos.

Esta ferramenta, além da proposta de tornar os atos públicos mais transparentes para a sociedade, está fortemente apoiada sobre os conceitos

da Web Semântica e dados conectados. Nosso objetivo é permitir que pessoas ou máquinas tenham a oportunidade de fazer consultas a estes conjuntos de dados de forma estruturada na web. E para isso, *DO2RDF* foi especificada e desenvolvida sobre as principais pilhas de protocolo da Web Semântica. Neste caso, a ferramenta expõe os dados no modelo *RDF/XML*.

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 e que faz parte da pilha de protocolos definidos pela w3c para a Web Semântica, Figura 1.1.

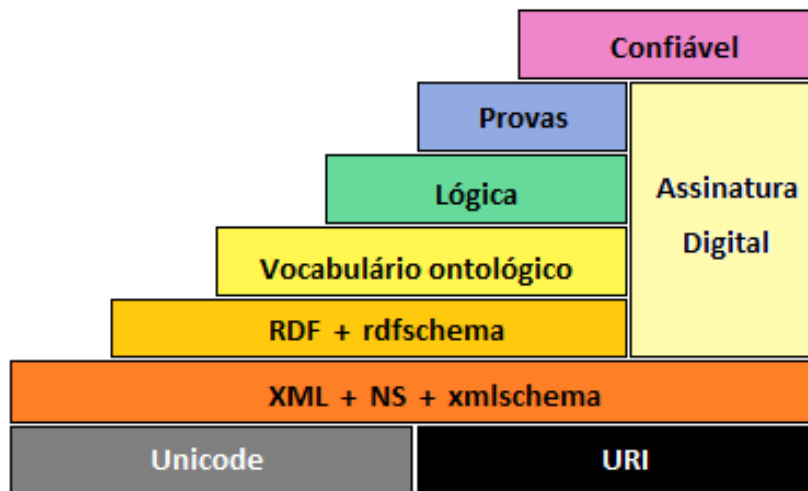


Figure 1.1: Camadas de protocolos da Web Semântica.

Segundo (5), *RDF* estende 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>*. Com esse simples modelo, o *RDF* permite que dados estruturados e semiestruturados sejam combinados, expostos e compartilhados por diferentes tecnologias.

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) (6). Recursos podem ser vistos como qualquer informação disponível, por exemplo, um documento, uma pessoa, etc. Neste caso, a cada recurso é atribuído um elemento identificador *IRI* (*Internationalized Resource Identifier*).

Inicialmente, a proposta do padrão *RDF* não definiu formalmente uma semântica(7)(9). Hoje, para expressar a semântica através do modelo *RDF*, é necessário a definição de um complemento ao *RDF*. Algumas propostas foram desenvolvidas como, por exemplo, *RDF-Schema* (*RDF-S*). *RDF-S* é a especificação que define classes, propriedades e seus relacionamentos, que podem ser usados para descrever triplas (6). Na prática, o esquema *RDF* é uma camada para definição de ontologias. Esta ontologia é composta por duas partes: uma que define as relações entre classes de objetos e outra parte lógica, composta por um conjunto de regras de inferência para os dados.

Sendo assim, este trabalho está dividido nas seguintes seções: Capítulo 1, esta introdução, apresentamos a nossa motivação da pesquisa. O Capítulo 2, a proposta da pesquisa com maiores detalhes sobre a ferramenta *DO2RDF*. O Capítulo 3, apresentamos os resultados gerados pela ferramenta: artefatos e consultas *SPARQL* em uma base *RDF*, e finalmente o Capítulo 4, onde faremos uma breve conclusão da pesquisa e trabalhos futuros.

2 Proposta

Nossa proposta é a extração de dados de atos públicos publicados no Diário Oficial Municipal (DOM) da Prefeitura do Rio de Janeiro. 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 7 anos de governo. 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 agilizar consultas dará maior transparência aos setores de auditoria e, conseqüentemente, a sociedade.

O primeiro desafio foi capturar todos os diários oficiais a partir de 1^o de janeiro de 2013, primeiro dia da gestão do atual prefeito. 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*.

O código apresentado no código 1, representa o que foi desenvolvido como primeira etapa deste projeto. Utilizando o método *requests* do *Python* foram baixados e tratados 896 diários oficiais. Nesta etapa, foram descartados diários (edições normais e suplementares) com apenas capas de publicação.

Code 1: Método *Request* para download dos diários.

```
1 count = 0
2 while count < 5000:
3     print(count)
4     filename = Path('pdf/' + str(count) + '.pdf')
5     print(filename)
6     url = 'http://doweb.rio.rj.gov.br/portal/edicoes/download/' +
7         str(count)
8     print(url)
9     response = requests.get(url)
10    filename.write_bytes(response.content)
11    count = count + 1
```

De posse destes documentos, a segunda etapa foi construir o extrator de informações contidas em cada diário. Neste processo, foi utilizada a biblioteca *RE*, do Python, para implementar técnicas de extração de dados, com um conjunto de expressões regulares, de forma automática nos documentos PDF. Naturalmente, algoritmos de expressões regulares tendem a ter seu comportamento guloso e extrair dados, de documentos PDF, tornou-se uma atividade de grande custo, dado o grande número de padrões de publicações presentes nos

diários. Esta falta de padronização fez com que fosse definida uma arquitetura de sistema capaz de interpretar cada padrão de forma única, evitando assim sobreposição de padrões.

Para este projeto, como mencionado, foram tratadas apenas as informações de movimentação de pessoal nos cargos de livre nomeação. Cargos de livre nomeação são cargos que não necessitam de concurso público para seu efetivo exercício, cabendo a gestor a escolha dos profissionais que devem ocupar tal cargo. Inicialmente, estamos coletando as seguintes informações dos diários:

- Número do diário;
- Data do diário;
- Tipo do diário (normal ou suplementar);
- Governo;
- Número da portaria de publicação;
- Data da portaria de publicação;
- Gestor da portaria (secretaria);
- Data do efetivo ato;
- Tipo do ato (resolução, portaria ou decreto);
- Tipo da movimentação (nomear, exonerar, designar e dispensar);
- Matrícula;
- Nome do servidor;
- Cargo Efetivo;
- Cargo Comissionado;
- Função Gratificada;
- Tipo do cargo.

Definido estes objetos de extração, foi desenvolvido os extratores de padrão utilizando técnicas de expressões regulares aplicadas a cada documento PDF. Como podemos ver no Código 2, para o caso do tipo de movimentação "Dispensar", atribuído às "funções gratificadas" e "empregos de confiança", foram identificados 8 (oito) formas diferentes de formalizar o ato no diário, consequentemente definimos também 8 (oito) padrões de expressões.

Até o presente momento da escrita deste trabalho, foram identificados 33 (trinta e três) diferentes padrões de atos.

Code 2: *Scripts com expressões regulares para atos "Dispensar".*

```

1      def atos_dispensar(self , buffer_local , Detalhe):
2          servidor = []
3          dispensar_pattern1 = re.compile(u'[\.\|\s]*Dispensar[,|\s
          ]*a\s*pedido[,|\s]*(?P<nome>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛ\s]+)[,|\s
          ]*(?P<cargoEfetivo>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛa-záêéóíçãâôú\-\s
          ]+),\s*matrícula\s*(?P<matricula>[0-9\.\-]+),\s*com\s
          *eficácia\s*a\s*contar\s*de\s*(?P<dia>[0-9]+).*\s*de\s
          *(?P<mes>[J|j]aneiro|[F|f]evereiro|[M|m]arço|[A|a]
          bril|[M|m]aio|[J|j]unho|[J|j]ulho|[A|a]gosto|[S|s]
          etembro|[O|o]utubro|[N|n]ovembro|[D|d]ezembro)\s*de\s
          *(?P<ano>[0-9]+),\s*da\s*[F|f]unção\s*[G|g]ratificada
          \s*(de)*\s*(?P<cargo>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛa-záêéóíçãâôú\-\s
          ]+),\s*símbolo\s*(?P<simbolo>[A-Z\0-9\|\s]+),')
4      dispensar_pattern2 = re.compile(u'[\.\|\s]*Dispensar[,|\s
          ]*a\s*pedido[,|\s]*(?P<nome>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛ\s]+)[,|\s
          ]*(?P<cargoEfetivo>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛa-záêéóíçãâôú\-\s
          ]+),\s*matrícula\s*(?P<matricula>[0-9\.\-]+),\s*da\s
          *[F|f]unção\s*[G|g]ratificada\s*de\s*(?P<cargo>[A-ZÉÁ
          ÍÓÚÇÃÊÕÕÀÛa-záêéóíçãâôú\-\s]+),\s*símbolo\s*(?P<
          simbolo>[A-Z\0-9\|\s]+),')
5      dispensar_pattern3 = re.compile(u'[\.\|\s]*Dispensar[,|\s
          ]*(?P<nome>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛ\s]+),*\s*matrícula\s*(?P<
          matricula>[0-9\.\-]+),\s*da\s*[F|f]unção\s*de\s*(?P<
          cargo>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛa-záêéóíçãâôú\-\s]+),*\s*sí
          mbolo\s*(?P<simbolo>[A-Z\0-9\|\s]+),')
6      dispensar_pattern4 = re.compile(u'[\.\|\s]*Dispensar[,|\s
          ]*(?P<nome>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛ\s]+)[\s|,]+com[\s|,]+
          validade[\s|,]+de[\s|,]+(?P<dia>[0-9]+).*\s*de\s*(?P<
          mes>[J|j]aneiro|[F|f]evereiro|[M|m]arço|[A|a]bril|[M|
          m]aio|[J|j]unho|[J|j]ulho|[A|a]gosto|[S|s]etembro|[O|
          o]utubro|[N|n]ovembro|[D|d]ezembro)\s*de\s*(?P<ano
          >[0-9]+)[\s|,]+(?P<cargoEfetivo>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛa-záê
          éóíçãâôú\-\s]+)[\s|,]+matrícula\s*(?P<matricula
          >[0-9\.\-]+)[\s|,]+da\s*[F|f]unção\s*[G|g]ratificada
          \s*(de)*\s*(?P<cargo>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛa-záêéóíçãâôú\-\s
          ]+)[\s|,]+símbolo[\s|,]+(?P<simbolo>[A-Z\0-9\|\s]+)
          ,')
7      dispensar_pattern5 = re.compile(u'[\.\|\s]*Dispensar[,|\s
          ]*(?P<nome>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛ\s]+)[\s|,]+(?P<
          cargoEfetivo>[A-ZÉÁÍÓÚÇÃÊÕÕÀÛa-záêéóíçãâôú\-\s]+)[\s
          |,]+matrícula\s*(?P<matricula>[0-9\.\-]+)[\s|,]*[\s
          |,]*com\s*eficácia\s*a\s*contar\s*de\s*(?P<dia
          >[0-9]+).*\s*de\s*(?P<mes>[J|j]aneiro|[F|f]evereiro|[
          M|m]arço|[A|a]bril|[M|m]aio|[J|j]unho|[J|j]ulho|[A|a]
          gosto|[S|s]etembro|[O|o]utubro|[N|n]ovembro|[D|d]

```

```

ezembro)\s*de\s*(?P<ano>[0-9]+) [\s|,]+da\s*[F|f]unção
\s*[G|g]ratificada\s*(de)*\s*(?P<cargo>[A-ZÉÁÍÓÚÇÃÊÕÕ
ÃÛa-záêéóíçãâôú\-\s]+) [\s|,]+sím[-|\n]*bolo [\s|,]+(?P
<simbolo>[A-Z\ -0-9\/\s]+),')
8 dispensar_pattern6 = re.compile(u'[\.\|\s]*Dispensar[,|\s
]*(?P<nome>[A-ZÉÁÍÓÚÇÃÊÕÕÃÛ\s]+) [,|\s]*matrícula\s*(?
P<matricula>[0-9\.\|/-]+) [,|\s]*da\s*[F|f]unção\s*de\s
*[C|c]onfiança\s*de\s*(?P<cargo>[A-ZÉÁÍÓÚÇÃÊÕÕÃÛa-záê
éóíçãâôú\-\s]+) [,|\s]*código\s*(?P<simbolo>[A-Z\ -0-9\
s\/]+),')
9 dispensar_pattern7 = re.compile(u'[\.\|\s]*Dispensar[,|\s
]*(?P<nome>[A-ZÉÁÍÓÚÇÃÊÕÕÃÛ\s]+) [,|\s]*registro\s*(N|
n)*\s*[0-9]+[,|\s]*com(\s)*(a)*(\s)*validade(\s)*(a)
*(\s)*partir(\s)*de(\s)*(?P<dia>[0-9]+) .*(\s)*de(\s)
*(?P<mes>[J|j]aneiro|[F|f]evereiro|[M|m]arço|[A|a]
bril|[M|m]aio|[J|j]unho|[J|j]ulho|[A|a]gosto|[S|s]
etembro|[O|o]utubro|[N|n]ovembro|[D|d]ezembro)\s*de\s
*(?P<ano>[0-9]+) [,|\s]*do\s*[E|e]mprego\s*de\s*[C|c]
onfiança\s*de\s*(?P<cargo>[A-ZÉÁÍÓÚÇÃÊÕÕÃÛa-záêéóíçãâ
ôú\-\s]+) [,|\s]*[C|c]ategoria\s*(?P<simbolo>[A-Z
-0-9\ â \-\/\s]+),')
10 dispensar_pattern8 = re.compile(u'[\.\|\s]*Dispensar[,|\s
]*(?P<nome>[A-ZÉÁÍÓÚÇÃÊÕÕÃÛ\s]+) [,|\s]*(?P<
cargoEfetivo>[A-ZÉÁÍÓÚÇÃÊÕÕÃÛa-záêéóíçãâôú\-\s]+) [,|\s
]+matrícula\s*(?P<matricula>[0-9\ â \-\/-]+) [,|\s]*
com\s*a*\s*validade\s*a*\s*partir\s*de\s*(?P<dia
>[0-9]+) .*\s*de\s*(?P<mes>[J|j]aneiro|[F|f]evereiro|[
M|m]arço|[A|a]bril|[M|m]aio|[J|j]unho|[J|j]ulho|[A|a]
gosto|[S|s]etembro|[O|o]utubro|[N|n]ovem[-|\n]*bro|[D
|d]ezembro)\s*de\s*(?P<ano>[0-9]+) [,|\s]*da\s*[F|f]un
ção\s*[G|g]ratificada\s*(de)*\s*(?P<cargo>[A-ZÉÁÍÓÚÇÃ
ÊÕÕÃÛa-záêéóíçãâôú\-\s]+),\s*(código)*\s*[0-9]*[,|\s
]*[S|s]ím[-|\n]*bolo[,|\s]*(?P<simbolo>[A-Z\ â
\ -0-9\/\s]+),')
11
12 disp1 = dispensar_pattern1.search(buffer_local)
13 disp2 = dispensar_pattern2.search(buffer_local)
14 disp3 = dispensar_pattern3.search(buffer_local)
15 disp4 = dispensar_pattern4.search(buffer_local)
16 disp5 = dispensar_pattern5.search(buffer_local)
17 disp6 = dispensar_pattern6.search(buffer_local)
18 disp7 = dispensar_pattern7.search(buffer_local)
19 disp8 = dispensar_pattern8.search(buffer_local)

```

Como resultado da execução deste algoritmo, até o presente momento, foram extraídos 14.826 atos públicos. Para facilitar o processo de análise

dos dados extraídos foi gerado um arquivo de auditoria (Figura 2.1) com as principais afirmações processadas. Este passo é necessário para garantir a qualidade da informação antes de ser inserida no banco de dados. Em em segundo momento, quando a ferramenta já estiver homologada este passo poderá ser reavaliado.

```
(PUC-RIO/TECMF)  ::PROCESSAMENTO DO DIÁRIO:: ANO: 31 No.: 000226 TIPO: NORMAL          * RIO DE JANEIRO * ARQUIVO: 3672.PDF SEQ.: 0010

(PADRAO 2.3)  RESOLUÇÕES  2284 21/02/2018 DESIGNAR 10/116630-5 REGINA COELI VIEIRA FERNANDES 01/02/2018 COORDENADOR PEDAGÓGICO
(PADRAO 2.3)  RESOLUÇÕES  2287 21/02/2018 DESIGNAR 10/231922-6 SABRINA MARTINS DA ROCHA 09/01/2018 DIRETOR-ADJUNTO
(PADRAO 2.3)  RESOLUÇÕES  2290 21/02/2018 DESIGNAR 10/266309-4 JACQUELINE CARDELLY JOORIS 01/02/2018 COORDENADOR PEDAGÓGICO
(PADRAO 2.3)  RESOLUÇÕES  2292 21/02/2018 DESIGNAR 10/264537-2 CAROLINE GONZALEZ VIVAS 01/02/2018 COORDENADOR PEDAGÓGICO
(PADRAO 2.3)  RESOLUÇÕES  2293 21/02/2018 DESIGNAR 10/177378-7 ANA LÚCIA OLIVEIRA DO COUTO 01/02/2018 COORDENADOR PEDAGÓGICO
(PADRAO 2.3)  RESOLUÇÕES  2294 21/02/2018 DESIGNAR 10/241890-3 MÔNICA JANETE CARVALHO TEIXEIRA 01/02/2018 COORDENADOR PEDAGÓGICO
(PADRAO 2.3)  RESOLUÇÕES  2302 21/02/2018 DESIGNAR 10/286152-4 FERNANDA DA SILVA CALDEIRA VASCONCELOS 01/02/2018 COORDENADOR PEDAGÓGICO
(PADRAO 2.3)  RESOLUÇÕES  2304 21/02/2018 DESIGNAR 10/171387-4 ANDRÉA MARIA RANGEL MARQUES 01/02/2018 COORDENADOR PEDAGÓGICO
(PADRAO 2.3)  RESOLUÇÕES  2305 21/02/2018 DESIGNAR 10/299736-9 CARLA AMORIM BDIRÁ ALVES 01/02/2018 COORDENADOR PEDAGÓGICO
(PADRAO 2.3)  RESOLUÇÕES  2306 21/02/2018 DESIGNAR 10/275922-3 JOICE MIRANDA TOLENTINO MENDES 01/02/2018 COORDENADOR PEDAGÓGICO
(PADRAO 2.4)  RESOLUÇÕES  2285 21/02/2018 DESIGNAR 10/259392-9 TERESA CRISTINA BORGES FERREIRA XX/XX/XXXX COORDENADOR PEDAGÓGICO
(PADRAO 2.4)  RESOLUÇÕES  2286 21/02/2018 DESIGNAR 10/285513-8 ERNANE CARRANO JANN XX/XX/XXXX COORDENADOR PEDAGÓGICO
(PADRAO 2.4)  RESOLUÇÕES  2289 21/02/2018 DESIGNAR 10/260465-0 ALESSANDRA RIBEIRO CAMPOS DE LIMA XX/XX/XXXX COORDENADOR PEDAGÓGICO
(PADRAO 2.4)  RESOLUÇÕES  2298 21/02/2018 DESIGNAR 10/127986-8 LÚCIA SOBRAL ROCHA SILVEIRA XX/XX/XXXX COORDENADOR PEDAGÓGICO
(PADRAO 2.6)  RESOLUÇÕES  2281 21/02/2018 DISPENSAR 12/235713-5 SILVIA DE JESUS PEREIRA XX/XX/XXXX COORDENADOR PEDAGÓGICO
(PADRAO 2.6)  RESOLUÇÕES  2282 21/02/2018 DISPENSAR 12/172270-1 SONIA MARIA DA SILVA MÁXIMO XX/XX/XXXX DIRETOR-ADJUNTO

(PUC-RIO/TECMF)  ::PROCESSAMENTO DO DIÁRIO:: ANO: 33 No.: 000079 TIPO: NORMAL          * RIO DE JANEIRO * ARQUIVO: 4207.PDF SEQ.: 0011

(PUC-RIO/TECMF)  ::PROCESSAMENTO DO DIÁRIO:: ANO: 32 No.: 000007 TIPO: NORMAL          * RIO DE JANEIRO * ARQUIVO: 3707.PDF SEQ.: 0012

(PADRAO 1.16) RESOLUÇÃO  1825 22/03/2018 EXONERAR 60/210987-4 MARCELO MOREIRA PESSOA 19/03/2018 DIRETOR I
(PADRAO 1.4)  RESOLUÇÃO  1827 22/03/2018 NOMEAR 60/260220-9 CLÁUDIO MARCIO FERREIRA DA SILVA 19/03/2018 GERENTE I
(PADRAO 1.16) RESOLUÇÃO  1828 22/03/2018 EXONERAR 60/260220-9 CLÁUDIO MARCIO FERREIRA DA SILVA 19/03/2018 ASSESSOR I
(PADRAO 1.3)  RESOLUÇÃO  1830 22/03/2018 NOMEAR XXXXXXXXXX RONALDO BRAZIL PEREIRA 20/02/2018 GERENTE II
(PADRAO 1.4)  RESOLUÇÃO  1831 22/03/2018 NOMEAR 60/292047-8 MANUELITO DE SOUSA REIS JUNIOR 01/02/2018 ASSISTENTE I
(PADRAO 1.18) RESOLUÇÃO  1835 22/03/2018 EXONERAR 11/215102-5 ELIAS SILVA DE OLIVEIRA 12/03/2018 COORDENADOR II
(PADRAO 1.18) RESOLUÇÃO  1837 22/03/2018 EXONERAR 11/237115-1 CARMEM LUCIA PEREIRA LOPES 14/03/2018 GERENTE II
```

Figure 2.1: Porção do arquivo de auditoria com as informações extraídas dos diários.

Um ponto a se observar no arquivo de auditoria é a qualidade das informações publicadas nos diários. Muitas datas estavam com ausência de informação e poucos casos com problemas na sua tipografia. Assim, para este momento do projeto, foi definido uma data "fictícia" para compor esta informação. No código 3 apresenta uma expressão regular que trata esta informação.

Code 3: Expressão regular para tratar datas ausentes ou inconsistentes

```
1  def converteData(sef, dates):
2      date_pattern = re.compile(u'(?P<data>[X]+|00/|0000)')
3      date = date_pattern.search(dates)
4      if (date):
5          vlr = '01/01/1900'
6      else: vlr = dates
7      return vlr
```

Para resumir esta atividade a Tabela 2.1 apresenta uma síntese da movimentação de pessoal na prefeitura do Rio de Janeiro:

Quantidade	Tipo da ação
5.311	Nomear
1.994	Designar
5.698	Exonerar
1.823	Dispensar

Table 2.1: Quantitativo de tipos de atos extraídos nos diários.

3

Documentação e Testes

Neste capítulo iremos apresentar a documentação do DO2RDF com base em artefatos UML (*Unified Modeling Language*), concentrados nos diagramas de classe, módulo e componentes. Também iremos apresentar o modelo de dados. Na sequência iremos apresentar um plano de testes e sua execução para validar a qualidade do modelo proposto.

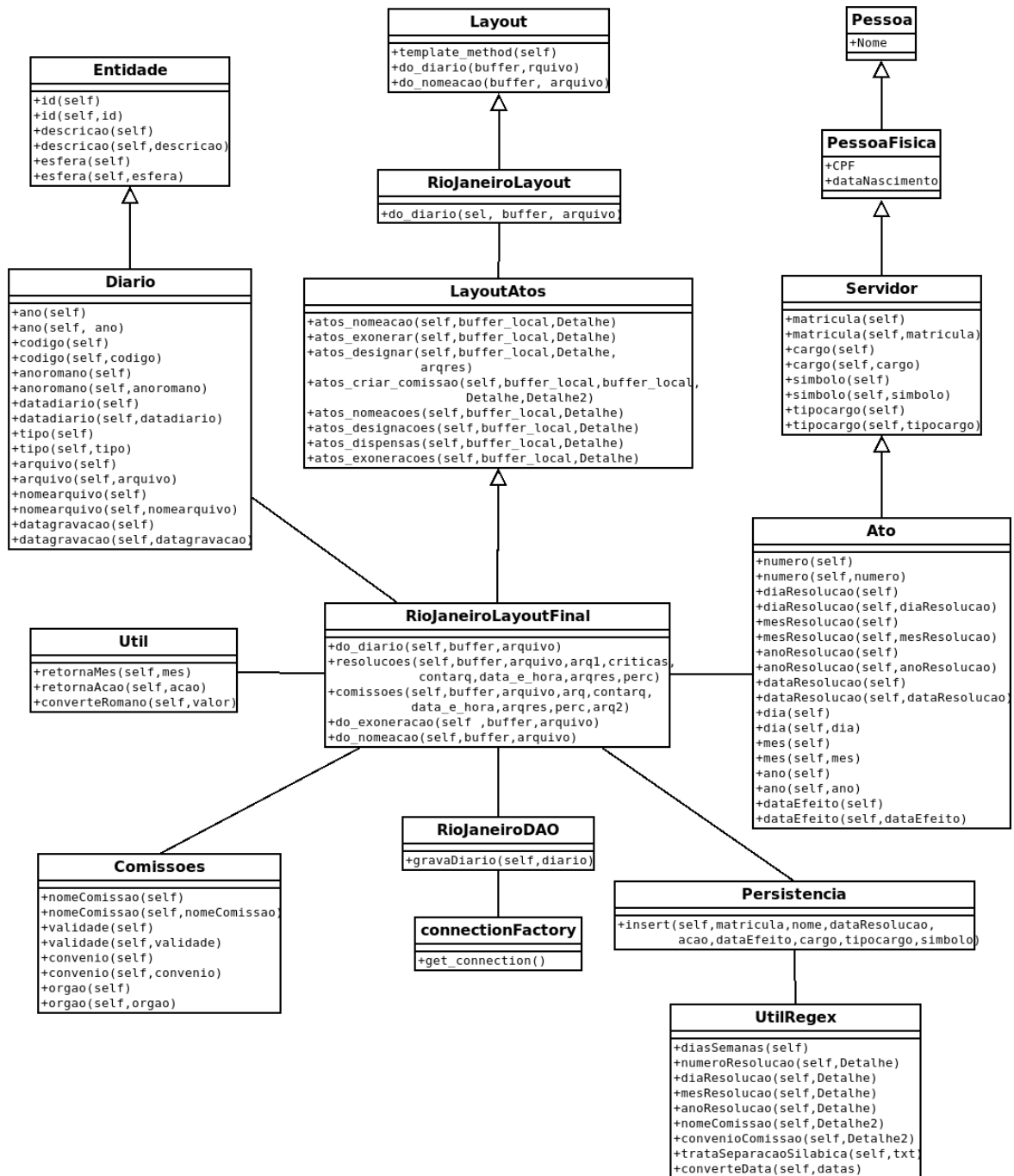


Figure 3.1: Diagrama de Classes do DO2RDF.

Como apresentado na seção anterior, de posse das informações previamente verificadas, foi realizada uma carga em um banco de dados *PostgreSQL* versão 12.1 em ambiente Ubuntu GNU/Linux. Estas informações são então tratadas por um módulo que normaliza os dados. A tabela de dados ficou assim organizada e apresentada na Figura 3.2:

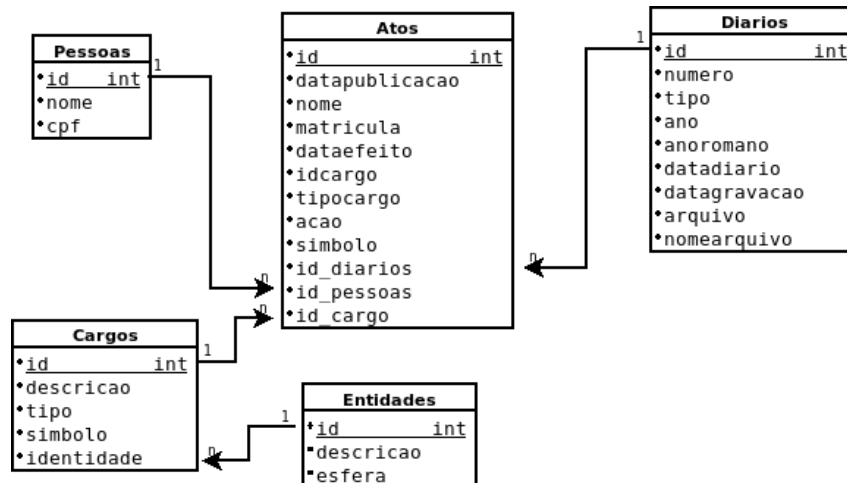


Figure 3.2: Modelo do banco de dados.

Uma observação é que para cada ato de nomeação, por exemplo, uma pessoa poderá ter uma nova matrícula associada. Este comportamento motivou esta modelagem das informações sobre matrículas. Quanto as outras informações não persistidas no banco é justificado por ser um ambiente de desenvolvimento e testes. Para uma versão final, em um ambiente com maior disponibilidade física de hardware, as demais informações farão parte do esquema.

3.0.1

Passos de implantação:

Para executar o DO2RDF é necessário um ambiente de desenvolvimento *Python 3* com os seguintes módulos:

- *psycopg2* - Módulo PostgreSQL;
- *tika* - Extrator de metadados e textos;
- *rdflib* - Biblioteca RDF;

Após preparar o ambiente de desenvolvimento, iremos preparar o ambiente de execução do DO2RDF. Para isto é necessário realizar os seguintes passos:

1. Instalar o PostgreSQL em uma versão ≥ 10.0 ;

2. Executar o script *Script_DO2RDT_PGSQL.sql* para construção do banco de dados.
3. Criar uma pasta *pdf/* dentro do diretório de fontes Python;
4. Criar uma pasta *rdf/* dentro do diretório de fontes Python;
5. Realizar o download dos diários oficiais da prefeitura do Rio de Janeiro executando o script *CrawlingDO.py*;
6. Executar o script *Principal.py*;
7. Executar o script *DO2RDF.py*;
8. Instalar o *AllegroGraph WebView versão 6.6.0*;
9. Fazer *deploy* das triplas *RDF* no *AllegroGraph*.

No final da execução do script *Principal.py* teremos um arquivo de auditoria gerado com o nome *AUDIT_DD-MM-YYYY_HH:MM:SS.txt*. Com este arquivo (Figura 2.1) podemos ver o resultado da extração das informações contidas nas seções dos diários. Também será armazenado no banco *postgres* as informações necessárias para o último passo, a triplificação *RDF*. Finalmente, após a análise dos dados do arquivo de auditoria podemos executar o script *DO2RDF.py*. Este script é responsável por extrair os dados do banco e construir as triplas *RDF*. Essas triplas podem ser vistas no arquivo *DO2RDF_13.rdf* na pasta *rdf/*.

Finalmente podemos apresentar a arquitetura da ferramenta *DO2RDF* (Figura 3.3), onde:

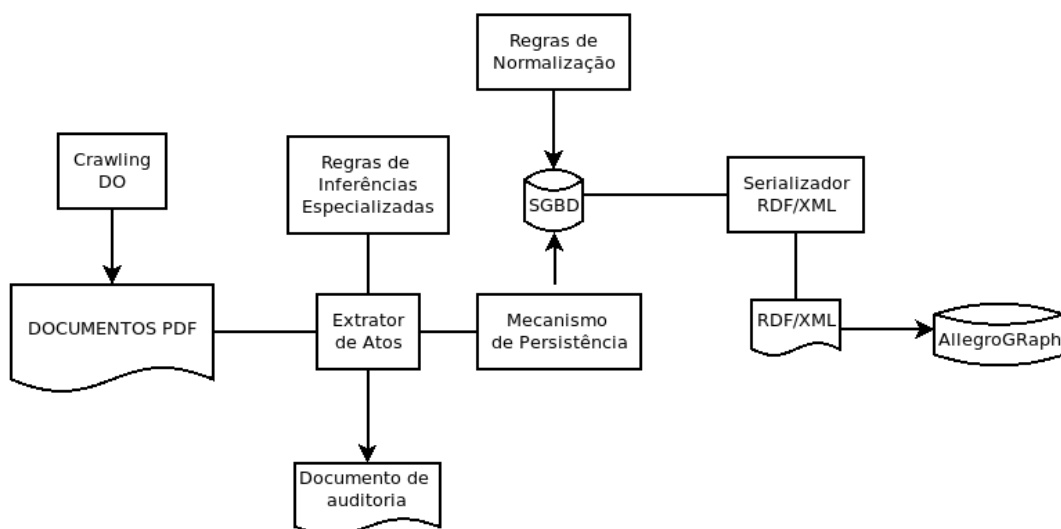


Figure 3.3: Arquitetura da ferramenta *DO2RDF*.


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*.
- **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*.

3.0.2

Testes:

A classe *LayoutAtos* é responsável por reconhecer os textos extraídos dos diários realizando o casamento dos padrões com uso de expressões regulares. O processo é bastante sensível pois uma pequena diferença no texto poderá acarretar no descasamento destes padrões implementados. Dito isto, achamos importante realizar testes unitários nesta classe, especificamente nos padrões que reconhecem atos públicos de Nomeação, Exoneração e Designações. Um pequeno exemplo de atos podem ser vistos nas figuras 3.4, 3.5 e 3.6.



D.O. RIO
Diário Oficial do Município do Rio de Janeiro

ATOS DO PREFEITO

DECRETO Nº 36694 DE 03 DE JANEIRO DE 2013

Altera a redação do Decreto nº 28324, de 16 de agosto de 2007.

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

DECRETA:
Art. 1º. O art. 1º do Decreto nº 28324, de 16 de agosto de 2007, passa a vigorar com a seguinte redação:

"Art. 1º. A gestão do Parque da Vitória e Ramo (Placido de Ramo)" será coordenada pela Secretaria Municipal de Esportes e Lazer - SMEL, que administrará os serviços públicos municipais a serem prestados pelos órgãos subordinados às Secretarias Municipais de Obras - SMO, Cultura - SMC, Trabalho e Emprego - SMT, Habitação - SMH, Secretaria Municipal de Pesca com Orlinhas - SMP, assim como a Companhia Municipal de Limpeza Urbana - COMLURB e a Guarda Municipal do Rio de Janeiro - GM-RIO".

Art. 2º. Este Decreto entra em vigor na data de sua publicação.
Rio de Janeiro, 03 de janeiro de 2013, 448º ano da fundação da Cidade.
EDUARDO PAES

DECRETO Nº 36695 DE 03 DE JANEIRO DE 2013

Declara de utilidade pública, para fins de desapropriação, o imóvel que menciona.

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso de suas atribuições legais e tendo em vista o disposto nos artigos 5º, alínea "I", e o 6º do Decreto Lei nº 3.365, de 21 de junho de 1941, com a redação que lhe deu a Lei Federal nº 9.785, de 29 de janeiro de 1999,

DECRETA:
Art. 1º Fica declarado de utilidade pública, para fins de desapropriação, o imóvel abaixo relacionado, necessário à implantação do Corredor Transolímpico:

Desapropriação Total:
+ Estrada das Bandeirantes, lote 3 da Quadra 2 do PAL 23.929 (nº 5.511)

Desapropriação Parcial:
+ Avenida Salvador Alende, lote 20 da Quadra 2 do PAL 28.733 (nº 180)

Art. 2º Este Decreto entra em vigor na data de sua publicação.
Rio de Janeiro, 03 de janeiro de 2013, 448º ano da fundação da Cidade.
EDUARDO PAES

DECRETO Nº 36696 DE 03 DE JANEIRO DE 2013

Modifica o Decreto nº 36.269 de 24 de setembro de 2012.

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso de suas atribuições legais e tendo em vista o disposto nos artigos 5º, alínea "I", e o 6º do Decreto Lei nº 3.365, de 21 de junho de 1941, com a redação que lhe deu a Lei Federal nº 9.785, de 29 de janeiro de 1999,

Art. 1º Modificar o Decreto nº 36.269 de 24 de setembro de 2012, de desapropriação total para desapropriação parcial:

+ Estrada do Catumbi nº 2.491.

Art. 2º Este Decreto entra em vigor na data de sua publicação.
Rio de Janeiro, 03 de janeiro de 2013, 448º ano da fundação da Cidade.
EDUARDO PAES

DECRETO Nº 36697 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE:
Nomear **ANA WAKSBERG GUERRINI**, matrícula 60/272.922-6, com validade a partir de 1º de janeiro de 2013, para exercer o Cargo em Comissão de Gestor 1746, símbolo DAS-10.A, código 035978, da Central de Teletendimento da PCRU - Central 1746, da Secretaria Municipal da Casa Civil.

DECRETO Nº 36698 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE:
Exonerar, a pedido, **GUSTAVO MIRANDA RODRIGUES**, matrícula 110158.389-6, Fiscal de Atividades Econômicas, com validade a partir de 1º de janeiro de 2013, do Cargo em Comissão de Gestor 1746, símbolo DAS-10.A, código 035978, da Central de Teletendimento da PCRU - Central 1746, da Secretaria Municipal da Casa Civil.

DECRETO Nº 36699 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE:
Exonerar, a pedido, **MICHELE SANTOS DA SILVA**, matrícula 110240.806-8, Assistente Social, com validade a partir de 28 de dezembro de 2012, do Cargo em Comissão de Coordenador Geral, símbolo DAS-10.B, código 019398, da Coordenadoria Geral de Gestão do Sistema Municipal de Assistência Social, da Secretaria Municipal de Assistência Social.

DECRETO Nº 36700 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE:
Exonerar **ANDERSON PINHEIRO LOPES**, matrícula 110218.894-2, Agente de Administração, com validade a partir de 1º de janeiro de 2013, do Cargo em Comissão de Assessor Especial, símbolo DAS-10.B, código 017224, da Secretaria Especial do Envelhecimento Saudável e Qualidade de Vida.

DECRETO Nº 36701 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE:
Exonerar **SANDRA REGINA JULIÃO**, matrícula 110337.495-8, Professor II, com validade a partir de 1º de janeiro de 2013, do Cargo em Comissão de Subsecretário, símbolo DAS-10.A, código 029537, da Subsecretaria, da Secretaria Especial do Envelhecimento Saudável e Qualidade de Vida.

DECRETO Nº 36698 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE:
Nomear **ADELMO FELICIANO DA SILVA**, matrícula 110190.263-4, Técnico de Contabilidade, com validade a partir de 1º de janeiro de 2013, para exercer o Cargo em Comissão de Subsecretário, símbolo DAS-10.A, código 029537, da Subsecretaria, da Secretaria Especial do Envelhecimento Saudável e Qualidade de Vida, mantendo-se inalterável sua vinculação com o Sistema de Controle Interno.

DECRETO Nº 36699 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE:
Designar **ADELMO FELICIANO DA SILVA**, Subsecretário, símbolo DAS-10.A, para, sem prejuízo de suas funções, substituir **Adelmo Feliciano da Silva**, Subsecretário, símbolo DAS-10.A, código 029537, da Subsecretaria, da Secretaria Especial do Envelhecimento Saudável e Qualidade de Vida, em seus impedimentos eventuais e afastamentos legais.

DECRETO Nº 36700 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE:
Designar **MARCUS VINICIUS AZEVEDO DA SILVA**, Assessor Especial, símbolo DAS-10.B, para, sem prejuízo de suas funções, substituir **Adelmo Feliciano da Silva**, Subsecretário, símbolo DAS-10.A, código 029537, da Subsecretaria, da Secretaria Especial do Envelhecimento Saudável e Qualidade de Vida, em seus impedimentos eventuais e afastamentos legais.

DECRETO Nº 36701 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE:
Nomear **MARCUS VINICIUS AZEVEDO DA SILVA**, matrícula 110227.408-2, Administrador, com validade a partir de 1º de janeiro de 2013, para exercer o Cargo em Comissão de Assessor Especial, símbolo DAS-10.B, código 017224, da Secretaria Especial do Envelhecimento Saudável e Qualidade de Vida.

DECRETO Nº 36702 DE 03 DE JANEIRO DE 2013

O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,


RESOLVE:
Nomear **TATIANA DA COSTA SILVA**, para exercer o Cargo em Comissão de Subsecretário, símbolo DAS-10.A, código 032019, da Subsecretaria de Esportes e Lazer, da Secretaria Municipal de Esportes e Lazer.

DESPACHOS DO PREFEITO

EXPEDIENTE DE 28/12/2012

18/000.862/2012 | OFÍCIO SMH/GAB Nº 824/2012

Assinatura.



D.O. RIO
Diário Oficial do Município do Rio de Janeiro
<http://www.rio.rj.gov.br/diario>

PREFEITURA DA CIDADE DO RIO DE JANEIRO
Secretaria Municipal da Casa Civil
Empresa Municipal de Artes Gráficas S/A
Imprensa da Cidade
Diretor Presidente: Franck Correa da Oliveira
Diretor de Administração e Finanças: Roberto M. Pereira
Diretor Industrial: Frederico Dias Martins De Paiva

A CAPA DO DIÁRIO OFICIAL É PRODUZIDA PELA COORDENADORIA ESPECIAL DE COMUNICAÇÃO DA SECRETARIA MUNICIPAL DA CASA CIVIL.

FOTOLITO E IMPRESSÃO:
GRÁFICA EDITORA JORNAL DO COMÉRCIO S.A.

AVISO

A Imprensa da Cidade comunica aos órgãos e entidades municipais que a Agência do D.O. Rio não aceitará a publicação de extrato de contrato que esteja em desacordo com o § 2º do art. 441 do RGCAF.

Preço das publicações (centímetro de coluna)	R\$ 5,00
Empresas Públicas, Fundações e Sociedades de Economia Mista do Município	R\$ 80,00
Terceiros (entidades e demais do Município)	R\$ 120,00

Os textos para publicação devem ser apresentados em cd, pendrive, digitais em fonte Arial, corpo 12, em linhas de 13 centímetros de largura, acompanhados de uma cópia com assinatura e identificação do responsável.

As páginas do Diário Oficial são formadas por três colunas de 68 centímetros.

Preço do Diário Oficial	R\$ 0,90
Exemplar avulso (vende na Agência D.O. RIO)	R\$ 1,20
Exemplar atrasado (sujeito à disponibilidade)	R\$ 250,00
Assinatura semestral (incluindo o frete)	R\$ 170,00

Entrega de matérias para publicação e forma de pagamento: A entrega das matérias, os pagamentos de publicações e a aquisição de exemplares atrasados devem ser efetuados diretamente na Agência D.O. Rio - Centro Administrativo São Sebastião - CASS, Rua Afonso Cavalcanti, 455 - Térreo - Cidade Nova. Tel: 2075-2284. As contratações ou renovações de assinaturas deverão ser efetuadas pelo telefone: 2075-2284. As assinaturas serão pagas por intermédio de boletos emitidos pela empresa e entregues aos assinantes. Para reclamações sobre publicações dirija-se Agência D.O. Rio - Centro Administrativo São Sebastião - CASS, Rua Afonso Cavalcanti, 455 - Térreo - Cidade Nova. Tel: 2075-2284, através do e-mail pdoficial@pcrj.rj.gov.br no prazo de 10 dias da data de circulação.

Ano XXVI • Nº 195 • Rio de Janeiro 3 Sexta-feira, 04 de Janeiro de 2013

Figure 3.4: Diário Oficial do Rio de Janeiro.

DECRETO “P” Nº 83 DE 03 DE JANEIRO DE 2013
O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE

Nomear **ANA WAKSBERG GUERRINI**, matrícula 60/272.922-6, com validade a partir de 1º de janeiro de 2013, para exercer o Cargo em Comissão de Gestor 1746, símbolo DAS-10.A, código 035078, da Central de Teleatendimento da PCRJ - Central 1746, da Secretaria Municipal da Casa Civil.

DECRETO “P” Nº 84 DE 03 DE JANEIRO DE 2013
O PREFEITO DA CIDADE DO RIO DE JANEIRO, no uso das atribuições que lhe são conferidas pela legislação em vigor,

RESOLVE

Exonerar, a pedido, **GUSTAVO MIRANDA RODRIGUES**, matrícula 11/156.395-6, Fiscal de Atividades Econômicas, com validade a partir de 1º de janeiro de 2013, do Cargo em Comissão de Gestor 1746, símbolo DAS-10.A, código 035078, da Central de Teleatendimento da PCRJ - Central 1746, da Secretaria Municipal da Casa Civil.

Figure 3.5: Ato de nomeação e exoneração de servidores.

nº 014 - Designar **DENISE NOGUEIRA BARCELLOS**, PROFESSOR II, matrícula 10/100429-0, para exercer a Função Gratificada de ASSISTENTE II da Gerência de Educação, símbolo DAI-06, código 005367, setor 10721 da 2ª Coordenadoria Regional de Educação, desta Secretaria, em vaga decorrente da dispensa de **ROSILENE ADRIANO MATTOS**. (ref. ao processo nº 07/02/004164/2012).

Figure 3.6: Ato de designar servidor.

Os testes unitários foram realizados através da classe *TestesUnitarios.py* com auxílio do pacote *unittest*. Esta classe detecta se as expressões regulares estão reconhecendo os padrões implementados na classe *Layout.py*. Além do casamento de padrões, foram testados, através da propriedade *self.assertEqual* de *unittest*, se as listas retornam as informações corretas como: nome, cargo e matrícula nos atos públicos. A figura 3.7 apresenta o resultados dos testes unitários aplicados à classe *LayoutAtos*.

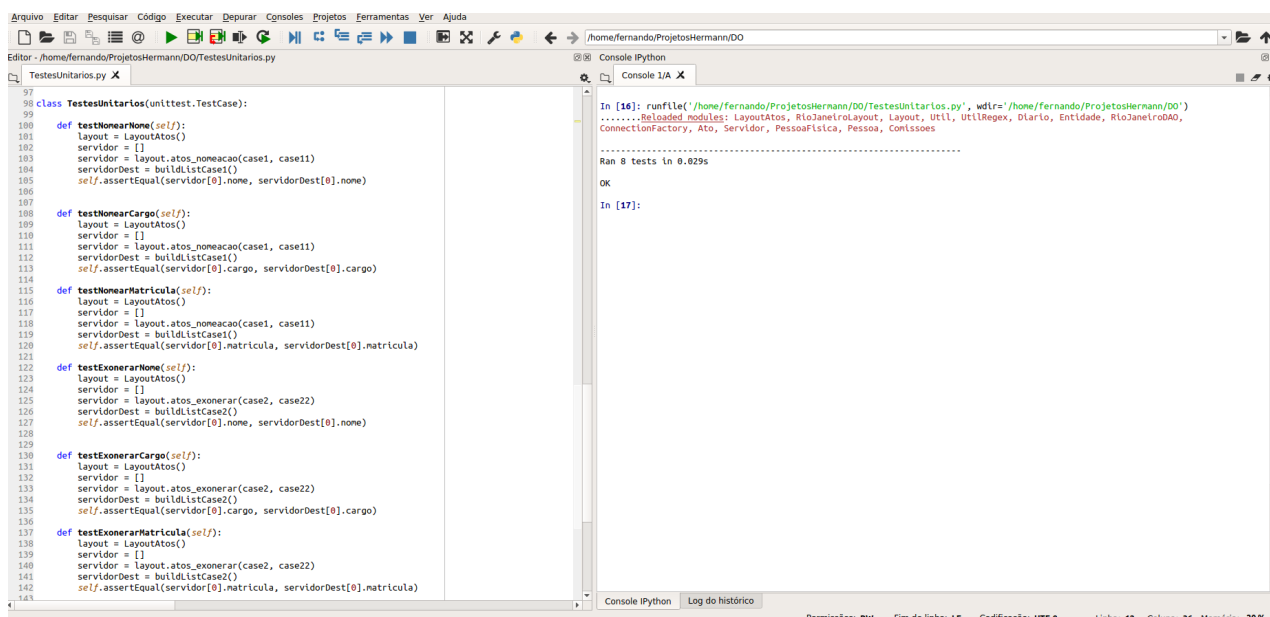


Figure 3.7: Resultado dos testes unitários.

As entradas dos testes unitários são apresentados abaixo e representam a redação normativa comum nos atos do diário oficial do Rio de Janeiro.

1. case1 = 'Nomear ANA WAKSBERG GUERRINI, matrícula 60/272.922-6, com validade a partir de 1o de janeiro de 2013, para exercer o Cargo em Comissão de Gestor, símbolo DAS-10.A, código 035078, da Central de Teleatendimento da PCRJ - Central 1746, da Secretaria Municipal da Casa Civil.'
2. case2 = 'Exonerar, a pedido, GUSTAVO MIRANDA RODRIGUES, matrícula 11/156.395-6, Fiscal de Atividades Econômicas, com validade a partir de 1o de janeiro de 2013, do Cargo em Comissão de Gestor, símbolo DAS-10.A, código 035078, da Central de Teleatendimento da PCRJ Central 1746, da Secretaria Municipal da Casa Civil.'
3. case3 = 'No 497 - Designar MAURÍCIO OLIVEIRA CHAVES, PROFESSOR DE ENSINO FUNDAMENTAL, matrícula 10/285208-5, para exercer, com eficácia a contar de 1o de fevereiro de 2019, a Função Gratificada de COORDENADOR PEDAGÓGICO, símbolo DAI-06, código 7835, setor 11075 da 8a Coordenadoria Regional de Educação, desta Secretaria. (ref. ao processo n o 07/08/000375/2019).'

Mais detalhes da implementação destes testes podem ser vistos na classe de testes unitários disponível em: <https://github.com/fernandoantoniodantas/DO2RDF/blob/master/TestesUnitarios.py>

4

Resultados

Neste capítulo iremos apresentar os resultados do processo de extração das informações contidas nos diários. Neste etapa, iremos apresentar o passo de triplificação em dados *RDF/XML* para ser executado em um ambiente de consultas *SPARQL*.

Como apresentando na seção anterior, de posse das informações previamente verificadas, foi realizada uma carga em um banco de dados *PostgreSQL* versão 12.1 em ambiente Ubuntu GNU/Linux. Estas informações são então tratadas por um módulo que normaliza os dados.

Com o banco de dados devidamente "populado", iniciamos o processo de triplificação das informações. Para esta etapa foi utilizada a biblioteca *RDFLib* (8) 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 com ela é possível executar consultas e atualizações *SPARQL*. Neste trabalho optamos por fazer a serialização dos dados no formato *RDF/XML* pois os demais formatos (*N3*, *NTriples* e *Turtle*) não foram possíveis de carga no ambiente do *AllegroGraph*. 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 FOAF (4).

O código 4 apresenta kernel do módulo de serialização dos dados do diário oficial e no código 5 podemos observar uma porção do arquivo *DO2RDF.rdf*, resultado do processo de serialização do *RDFLib*, para carga no *AllegroGraph*.

Code 4: Processo de serialização em *RDF/XML*

```
1 for row in funcionarios_records:
2
3     seqa+=1
4     idP = seqa
5     idP = BNode()
6
7     store.add((idP, RDF.type, FOAF.Funcionarios))
8     store.add((idP, FOAF.matricula, Literal(row[0].strip())))
9     store.add((idP, FOAF.nome, Literal(row[1].strip())))
10    store.add((idP, FOAF.dataPublicacao, Literal(row[2])))
```

```

11     store.add((idP, FOAF.acao, Literal(row[3].strip())))
12     store.add((idP, FOAF.dataEfeito, Literal(row[4])))
13     store.add((idP, FOAF.cargo, Literal(row[5].strip())))
14     store.add((idP, FOAF.tipoCargo, Literal(row[6].strip())))
15     store.add((idP, FOAF.simbolo, Literal(row[7].strip())))
16
17     # Serialize the store as RDF/XML to the file DO2RDF.rdf.
18     store.serialize("RDF/DO2RDF.rdf", format="pretty-xml",
19                    max_depth=3)
19     print('RDF Serializations:', seqa, 'De', size)

```

Nesta etapa, o módulo gerador apresentou lentidão para a serialização dos 14.826 registros de atos. Sua completa geração está levando em torno de 26 horas neste ambiente virtual (Intel core i7, 8 GB de memória RAM). A solução foi particionar sua execução em lotes de 1.000 registros, que levou pouco mais de 2 horas no total.

Code 5: Processo de serialização em *RDF/XML*

```

1
2 <?xml version="1.0" encoding="utf-8"?>
3 <rdf:RDF
4   xmlns:foaf="http://xmlns.com/foaf/0.1/"
5   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
6 >
7   <foaf:Funcionario rdf:nodeID="
8       N7654fa3513984ea89c0ade256a971939">
9       <foaf:acao>NOMEAR</foaf:acao>
10      <foaf:dataPublicacao>2017-01-19</foaf:dataPublicacao>
11      <foaf:tipoCargo>CC</foaf:tipoCargo>
12      <foaf:simbolo>DAS-06</foaf:simbolo>
13      <foaf:dataEfeito>2017-01-01</foaf:dataEfeito>
14      <foaf:matricula>60/210917-1</foaf:matricula>
15      <foaf:cargo>ASSISTENTE I</foaf:cargo>
16      <foaf:nome>MARIA CRISTINA DOS SANTOS BASTOS</foaf:nome>
17    </foaf:Funcionario>
18    <foaf:Funcionario rdf:nodeID="
19        Nde2047713a3f431e91f6f2c4471025ee">

```

```

18   <foaf:simbolo>DAS-10</foaf:simbolo>
19   <foaf:dataEfeito>1900-01-01</foaf:dataEfeito>
20   <foaf:nome>RICARDO VIEIRA SILVA</foaf:nome>
21   <foaf:cargo>SUBSECRETÁRIO</foaf:cargo>
22   <foaf:matricula>11/087325-7</foaf:matricula>
23   <foaf:acao>NOMEAR</foaf:acao>
24   <foaf:dataPublicacao>2019-02-21</foaf:dataPublicacao>
25   <foaf:tipoCargo>CC</foaf:tipoCargo>
26   </foaf:Funcionario>

```

Para executar as consultas *SPARQL* foi instalado o *AllegroGraph* na versão 6.6.0 (1), em ambiente virtual com Ubuntu GNU/Linux. Na sequência, foi criado o repositório *DO2RDF* para receber a carga do arquivo *DR2RDF.rdf*.

Podemos observar na Figura 4.1 o resultado de uma simples consulta *SPARQL* para os casos de nomeação para todos os cargos que iniciam com a expressão "ADM".

The screenshot shows the AllegroGraph WebView 6.6.0 interface. The top navigation bar includes links for Repository, Queries, Utilities, Admin, and User fernando. The main area is titled 'Edit query' and contains a SPARQL query. Below the query editor are buttons for 'Execute', 'Log Query', 'Show Plan', 'Save as', and 'Add to repository'. The 'Execute' button has been clicked, resulting in a table of 98 results in 98.938 ms. The table has four columns: nome, cargo, tipo, and acao. The results list names of individuals, all with the cargo 'ADMINISTRADOR REGIONAL', tipo 'CC', and acao 'NOMEAR'.

nome	cargo	tipo	acao
"ADILSON RIBEIRO DE LIMA"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"
"ALEKSANDRO SILVA DE SOUSA"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"
"ALEXANDRE DAMIÃO GARRIDO GLANTTINE"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"
"ANDERSON GUILHERME DA SILVA"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"
"ANDRÉ LUIZ CARVALHO FERRAZ"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"
"ANDRÉ LUIZ GOUVEIA VIEIRA"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"
"ANDRÉA LUCIA COELHO"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"
"ANTONIO VELTRI JUNIOR"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"
"BIANCA DE CARVALHO"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"
"CARLOS AUGUSTO DE SOUZA QUINTANILHA"	"ADMINISTRADOR REGIONAL"	"CC"	"NOMEAR"

Figure 4.1: Consulta das nomeações para cargos com iniciais "ADM".

Para concluir, na Figura 4.1, podemos observar o resultado de uma consulta para os casos todas os registros de exonerações.

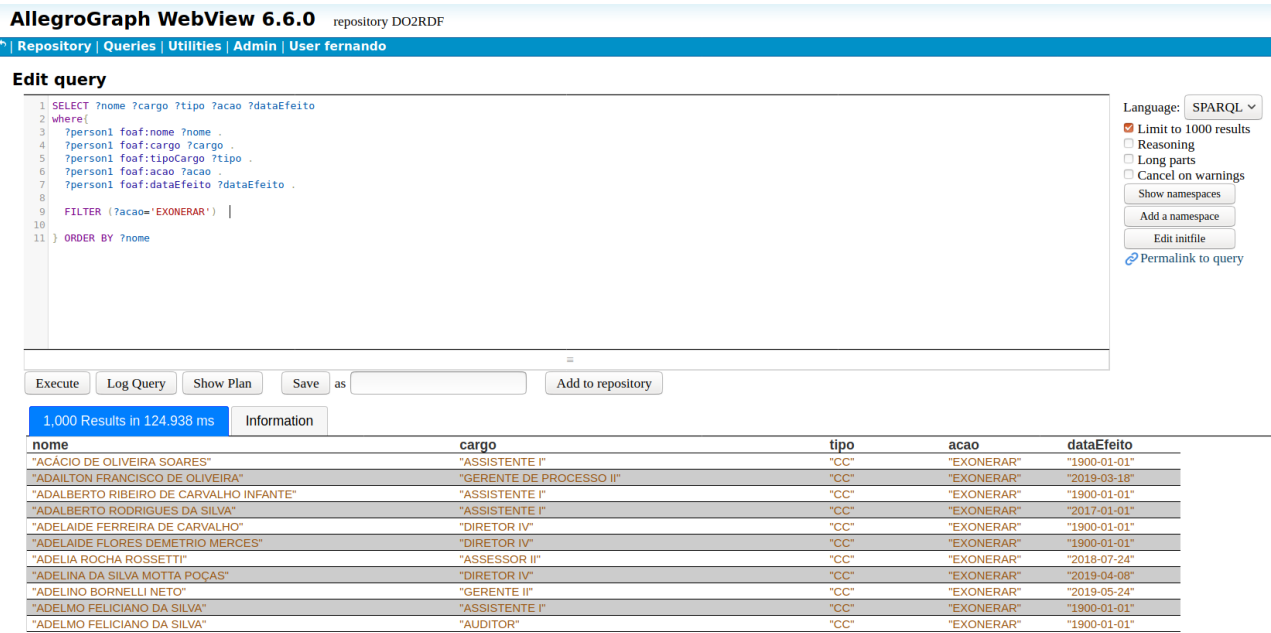


Figure 4.2: Consulta de exonerações.

5

Conclusões e trabalhos futuros

Apresentamos como a transparência pública poderá ser dada por processos automatizados de leitura e extrações de informações de atos públicos disponíveis em Diários Oficiais. De posse dos diários, e definido uma arquitetura de estilos de padrões, o processo de extração utilizando expressões regulares apresentou-se de forma simples e eficiente.

Um outro ponto a se observar foi quanto a utilização da biblioteca *RDFLib* que apresentou ser pouco flexível para criação de ontologias genéricas. Inicialmente tentamos modificar seus *scripts* para que fossem incorporados novos conceitos. Uma solução mais simples foi adaptar a ontologia *FOAF* para nossas necessidades.

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.

Como trabalhos futuros, podemos 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 extração bastando definir seus padrões de publicação na forma de expressões regulares. 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*.

Referências bibliográficas

- [1] ALLEGROGRAPH. **Allegrograph - The Enterprise Knowledge Graph**. <https://allegrograph.com>. Acessado: 11-12-2019. 4
- [2] BRASIL. **Emenda Constitucional nº 9, de 9 de novembro de 1995**. Diário Oficial [da] República Federativa do Brasil, 59:1966. 1
- [3] BRASIL. **Lei nº. 12.527. Lei de Acesso à Informação**. Diário Oficial [da] República Federativa do Brasil. 1
- [4] BRICKLEY, D.; MILLER, L.. **FOAF Vocabulary Specification 0.99**. <http://www.foaf-project.org/>, 2014. Acessado: 03-12-2019. 4
- [5] GROUP, R. W.. **Resource Description Framework (RDF)**. <https://www.w3.org/RDF/>, 2014. Acessado: 11-10-2019. 1
- [6] ISOTANI, S.; BITTENCOURT, I.. **Dados Abertos Conectados: em Busca da Web do Conhecimento**. 08 2015. 1
- [7] LASSILA, O.; SWICK, R. R.. **Resource Description Framework (RDF) Model and Syntax Specification**. W3c recommendation, W3C, February 1999. 1
- [8] TEAM, R.. **RDFLib**. <https://rdflib.readthedocs.io/en/stable/#>, 2013. Acessado: 13-11-2019. 4
- [9] YANG, G.; KIFER, M.. **On the Semantics of Anonymous Identity and Reification**. In: ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS, 2002 - DOA/COOPIS/ODBASE 2002 CONFEDERATED INTERNATIONAL CONFERENCES DOA, COOPIS AND ODBASE 2002, p. 1047–1066, Berlin, Heidelberg, 2002. Springer-Verlag. 1