CENTRO UNIVERSITÁRIO DO PARÁ - CESUPA
ESCOLA DE NEGÓCIOS, TECNOLOGIA E INOVAÇÃO - ARGO
CURSO DE ENGENHARIA DA COMPUTAÇÃO

JHERSON HARYSON ALMEIDA PEREIRA
LUIS ENRIQUE GOMES PORTUGAL

AUTOMATIZANDO A EXTRAÇÃO DE DADOS PÚBLICOS: MINERAÇÃO DE DADOS APLICADA AO DIÁRIO OFICIAL DO ESTADO DO PARÁ

BELÉM
2020

JHERSON HARYSON ALMEIDA PEREIRA
LUIS ENRIQUE GOMES PORTUGAL

AUTOMATIZANDO A EXTRAÇÃO DE DADOS PÚBLICOS: MINERAÇÃO DE DADOS APLICADA AO DIÁRIO OFICIAL DO ESTADO DO PARÁ

Trabalho de conclusão de curso apresentado à Escola de Negócios, Tecnologia e Inovação do Centro Universitário do Estado do Pará como requisito para obtenção do título de Bacharel em Engenharia da Computação na modalidade MONOGRAFIA.

Prof. M.e Moshe Dayan Sousa Ribeiro

BELÉM
2020

JHERSON HARYSON ALMEIDA PEREIRA
LUIS ENRIQUE GOMES PORTUGAL

AUTOMATIZANDO A EXTRAÇÃO DE DADOS PÚBLICOS: MINERAÇÃO DE DADOS APLICADA AO DIÁRIO OFICIAL DO ESTADO DO PARÁ

Trabalho de conclusão de curso apresentado à Escola de Negócios, Tecnologia e Inovação do Centro Universitário do Estado do Pará como requisito para obtenção do título de Bacharel em Engenharia da Computação na modalidade MONOGRAFIA.

Data da aprovação: / /
Nota final aluno I: _____
Nota final aluno II: _____

Banca examinadora

_____
Prof. M.e Moshe Dayan Sousa Ribeiro
Orientador e Presidente da banca

_____
Prof. M.a Daniele Moura de Queiroz
Examinador

Dados Internacionais de Catalogação-na-publicação (CIP)
Biblioteca do CESUPA, Belém – PA

Pereira, Jherson Haryson Almeida.

Automatizando a extração de dados públicos: mineração de dados aplicada ao Diário Oficial do Estado do Pará / Jherson Haryson Almeida Pereira, Luís Enrique Gomes Portugal; orientador Moshe Dayan Sousa Ribeiro. – 2020.

Trabalho de Conclusão de Curso (Graduação) – Centro Universitário do Estado do Pará, Engenharia da Computação, Belém, 2020.

1. Mineração de dados. 2. Banco de dados. I. Portugal, Luís Enrique Gomes. II. Ribeiro, Moshe Dayan Sousa, orient. III. Título.

CDD 23ª ed. 005.74

“Dados! Dados! Dados!”, ele gritou impacientemente. “Não posso construir tijolos sem barro”.

-Arthur Conan Doyle

RESUMO

O Governo do Estado do Pará divulga seus atos públicos através de seu Diário Oficial, disponibilizando-os em um formato voltado para o consumo humano. Entretanto, a falta de estrutura do formato disponibilizado dificulta a análise de dados por sistemas computacionais, prejudicando a obtenção de informações relevantes para a fiscalização de modo célere. Para a solução do mesmo, é proposto um modelo computacional para sintetizar dados do Diário Oficial do Estado do Pará, em um modelo bem definido, com o intuito auxiliar a análise sistemática apoiada por ferramentas computacionais. Para isto, este trabalho aborda os processos fundamentais para que um modelo consiga estruturar e extrair informações públicas do Estado do Pará, bem como todo o processo de implementação do mesmo, utilizando técnicas que permitam a extração de informação em dados não estruturados. Por final o modelo proposto conseguiu atingir os objetivos esperados, sintetizando as informações dispostas no Diário Oficial do Estado do Pará e persistindo em um banco de dados, viabilizando e agilizando ao analista, a análise de dados mais rápida e eficaz.

Palavras-chave: Scraping; Mineração de dados; Diário Oficial do Estado do Pará;

ABSTRACT

The Government of the State of Pará discloses its public acts through its Official Gazette, making them available in a format aimed at human consumption. However, the lack of structure of the format provided makes it difficult to analyze data by computer systems, hampering the obtaining of relevant information for inspection in a quick manner. For its solution, a computational model is proposed to synthesize data from the Official Gazette of the State of Pará, in a well-defined model, in order to assist systematic analysis supported by computational tools. For this, this work addresses the fundamental processes for a model to be able to structure and extract public information from the State of Pará, as well as the entire process of implementing it, using techniques that allow the extraction of information from unstructured data. Finally, the proposed model managed to achieve the expected objectives, synthesizing the information provided in the Official Gazette of the State of Pará and persisting in a database, making the analysis of data faster and more efficient and faster.

Keywords: Scraping; Data mining; Official Gazette of the State of Pará;

SUMÁRIO

Capítulo 1

Introdução

Este capítulo apresenta a descrição da motivação da pesquisa executada neste trabalho, os objetivos traçados para a pesquisa, além de mostrar a estrutura organizacional deste documento, contendo a sinopse de cada capítulo.

1 INTRODUÇÃO

O controle da informação tem sido umas das características de maior preocupação do ser humano desde os tempos antigos; podendo tomar como exemplo, a escrita, que nasce como meio de capturar ideias e conhecimentos para a resolução de problemas nas mais diversificadas áreas, como: agricultura, astronomia, caça e pesca (MARTINS, 2014).

Dados são a base para a informação, e consequentemente, para o conhecimento. Levando em consideração o atual contexto da informática e da globalização; dados são criados de forma digital e em larga escala, podendo ser aplicados para a resolução dos mais diversos problemas de uma sociedade moderna (ASKITAS & ZIMMERMANN, 2015), auxiliando a administração pública na gestão de recursos, no controle de gastos, na tomada de decisão e na transparência de seus atos.

Leis como a Lei complementar 131, nascem a fim de gerar maior transparência no âmbito Estatal e Federal, cujo principal objetivo é a disponibilização em tempo real de informações acerca da parte financeira da União[1] dos Estados e dos Municípios, ou seja, essa Lei tem como objetivo tornar órgãos mais transparentes em prol do desenvolvimento da democracia (ANGEL, 2008).

A disponibilização de informações, neste cenário, traz um novo paradigma onde o governo federal brasileiro, contemplando a Lei de disponibilização de informação, vem ampliando e aperfeiçoando os instrumentos que permitem ao cidadão ter participação ativa quanto às ações do Estado (DROPA, 2004), fornecendo-o meios de fiscalizar suas próprias ações.

O acesso às informações sobre as finanças da administração pública trouxe também uma nova abordagem quanto ao papel do cidadão e a identificação de fraudes. Agora, o cidadão é parte fundamental neste processo, cobrando o governo para a correta aplicação de seus recursos. Por outro lado, o Estado também corrobora para a fiscalização, disponibilizando instrumentos como: Caixa Único da União; Plano de Contas Único da Administração Pública Federal; disponibilizando o acesso aos dados financeiros do governo e da fiscalização da contabilidade pelos Tribunais de Contas da União, do Estado e dos Municípios (DROPA, 2004). Atualmente, a divulgação dessas informações, se dá geralmente, por meio de diários oficiais[2], disponibilizados através dos portais de transparência que evidenciem a arrecadação de receitas e despesas diárias (BRASIL, 2009).

---

[1] União dos Estados e Municípios constituem a República Federativa do Brasil.

[2] Os diários oficiais são jornais criados, mantidos e administrados por governos para publicar as literaturas dos atos oficiais da administração pública executiva, legislativa e judiciária, sendo este um meio de publicação na

No entanto, Paiva e Revoedo (2016) ressaltam que as disponibilizações dessas informações em portais governamentais não asseguram um aumento efetivo do grau de transparência desses entes, que, devido ao grande volume de dados, aliado à falta de padrões, torna custoso qualquer tipo de acompanhamento sistemático.

Uma possível solução para esse tipo de problema se dá através da aplicação de um sistema computadorizado para ler informações não estruturadas e gerar uma saída estruturada que pode ser tomada como base para análises sistemáticas. Neste cenário, são comumente utilizadas técnicas de tratamentos de dados, sendo necessário um método capaz de a) processar um grande volume de dados e b) que permitam uma visualização consolidada dessas informações, facilitando assim, a sua identificação e análise.

Diante dessa perspectiva, este estudo discorre utilizando como matéria prima o Diário Oficial do Estado do Pará publicado pela Imprensa Oficial do Estado do Pará (IOEPA), cujas características se enquadram ao que foi levantado anteriormente. Este trabalho tem como objetivo o desenvolvimento de um processo computadorizado para a extração de informação do diário citado, a fim de auxiliar os processos de acompanhamento e análise de maneira mais rápida, profícua e eficaz.

1.1 SITUAÇÃO PROBLEMA

Entre os crimes contra a administração pública previstos no Código Penal Brasileiro, estão: o exercício arbitrário ou abuso de poder; a falsificação de papéis públicos; a má-gestão praticada por administradores públicos; a apropriação indébita previdenciária; a lavagem ou ocultação de bens oriundos de corrupção; o emprego irregular de verbas ou rendas públicas; contrabando ou descaminho e a corrupção ativa e passiva (CONSELHO NACIONAL DE JUSTIÇA, 2015).

Padeiro (2017, p. 1), descreve a corrupção como sendo:

> "[...] maléfica porque se trata de um obstáculo que altera os mercados, cria vantagens altamente desiguais, diminui as projeções dos recursos das políticas públicas, implica em deformações no campo dos negócios e fortalece a cultura do superfaturamento com as transgressões e demandas inflacionadas"

De fato, não é novidade que a corrupção distorce especialmente a alocação dos recursos públicos. No Brasil, as principais consequências são facilmente visíveis no

qual são publicadas as Leis, licitações, atas de plenário e todas as demais atividades de uma divisão administrativa brasileira.

crescimento econômico, e no desenvolvimento social, causando escassez e desfalque de recursos, além de outros efeitos sistêmicos em cadeia, implicando em crises financeiras e dívidas (PADEIRO, 2017).

> "A corrupção pode se apresentar de diversas maneiras; se escondendo em obras superfaturadas; em caixa dois para partidos políticos ou na aprovação de Leis que garantem vantagens para uma minoria. Sempre que você tem corrupção você tem um tipo de superfaturamento. Essa gordura, esse preço, é dinheiro do orçamento público que vai para mãos privadas, às vezes, divididos com políticos, burocratas e até mesmo com empresários". (FERREIRA, 2017, p. 8)

Para fiscalizar e garantir a correta aplicação dos recursos públicos, o Estado dispõe de órgãos, como a Controladoria Geral da União (CGU) e a Controladoria-Geral do Estado (CGE) no âmbito da União e o Tribunal de Contas do Estado (TCE) e a Auditoria Geral do Estado (AGE) no âmbito do Estado do Pará. Estes órgãos contam com a participação dos cidadãos, para que de uma maneira ainda mais eficaz, exerçam o controle dos recursos do Estado, possibilitando assim realizar ações de fiscalização concomitante para evitar o dano ao erário ou recuperar recursos desviados.

O Diário Oficial do Estado é um dos instrumentos utilizados para exercer a fiscalização dos atos do governo do estado do Pará. Contudo, essa ação se torna custosa devido ao grande volume de dados gerados diariamente. Os esforços são ainda maiores devido à falta de padronização. Cada órgão insere os seus atos em um formato textual livre, demandando mais esforço do analista para a compreensão, e ainda podendo ser passível de interpretações. Este cenário aumenta ainda mais a dificuldade da rápida identificação de Informações relevantes para fiscalização. A ilustração 1 mostra o comparativo entre a escrita de atos para dois órgãos distintos.

Ilustração 1 - Comparativo entre o tipo de escrita entre órgãos distintos

INSTITUTO DE ASSISTÊNCIA DOS SERVIDORES DO ESTADO DO PARÁ

TERMO ADITIVO A CONTRATO
PROCESSO Nº 2020/193393
Termo Aditivo:2º
Data da Assinatura:02/04/2020
Vigência: 02/04/2020 a 02 /04/2021
Justificativa: Prorrogação do prazo de vigência.
Contrato: 017
Exercício: 2018
Dotação Orçamentária: 8888- 0261 - 339039
Contratada: REABILITAÇÃO CLÍNICA & SERVIÇOS LTDA-CNPJ n. 27.613.119/0001-96
Endereço: RUA JOSE BARROS DA SILVA nº.608, CAPITÃO POÇO/PA, CEP:68.650-000
Ordenador: BERNARDO ALBUQUERQUE DE ALMEIDA

Protocolo: 557874

SECRETARIA DE ESTADO DA FAZENDA

TERMO ADITIVO A CONTRATO

Secretaria de Estado de Desenvolvimento Urbano e Obras Públicas–CNPJ 03.137.985/0001 90
Pará Segurança LTDA – CNPJ 04.113.174/0001-11
Objeto: Prestação de Serviços de Vigilância./Segurança Patrimonial Armada nos Turnos Diurnos e Noturnos nesta SEDOP, neste Estado. Justificativa: Repactuação de serviços nos termos da cfe. art.65 II, d, da Lei nº 8.666/93 REPACTUAÇÃO: R$ 1.208,49 Data da Assinatura: 01/07/2020 Ordenador Responsável: Benedito Ruy Santos Cabral Secretário de Estado de Desenvolvimento Urbano e Obras Pública

Protocolo: 557787

Fonte: Os autores

Desse modo, é imprescindível o apoio computacional para extração automatizada de dados relevantes para uma fiscalização mais eficiente.

1.2 OBJETIVOS DO ESTUDO

1.2.1 Objetivo Geral

Este trabalho tem como objetivo geral desenvolver um mecanismo de *software* para a obtenção, extração e disponibilização de dados estruturados do Diário Oficial do Estado do Pará.

1.2.2 Objetivos Específicos

- Identificar os tipos de dados disponibilizados pelo Diário Oficial do Estado do Pará.
- Estudar os meios de obtenção de informação do Diário Oficial do Estado do Pará.
- Projetar um modelo capaz de coletar os arquivos do Diário Oficial do Estado do Pará.
- Projetar um modelo capaz de pré-processar os arquivos coletados.
- Projetar um modelo capaz de estruturar os dados pré-processados.
- Projetar um modelo capaz de extrair informações dos dados estruturados.
- Projetar um modelo capaz de persistir as informações extraídas.

1.3 JUSTIFICATIVA

O presente trabalho justifica-se não pela falta de dados provenientes do Diário Oficial do Estado do Pará, mas sim devido à dificuldade em analisá-los rapidamente e obter informações relevantes para a fiscalização e o controle, sendo a raiz desse problema a falta de uma estruturação na disponibilização de informações, dificultando a análise sistemática apoiada por ferramentas computacionais.

Diante desses fatos, este trabalho propõe uma ferramenta para extração de dados do Diário Oficial do Estado do Pará, e disponibilização em um novo modelo bem definido para facilitar o consumo por quem tenha interesse em fiscalizar atos da administração pública do estado do Pará.

1.4 METODOLOGIA DA PESQUISA

Inicialmente, realizou-se um estudo acerca do atual cenário da fiscalização dos atos da administração pública. Diante do resultado encontrado, foi iniciado o processo de entendimento acerca de tecnologias envolvidas na obtenção das informações dos portais de transparência e nos tipos de dados disponibilizado pelo mesmo. Logo após, analisaram-se as técnicas para a manipulação da informação não estruturada e modelos capazes de representar estruturas de dados bem definidas. Subsequentemente, uma pesquisa bibliográfica foi desenvolvida, com o intuito de identificar as principais tecnologias utilizadas para a realização de técnicas de *data scraping*[3] na *web* moderna. Ademais, foi desenvolvida uma pesquisa voltada a identificar soluções arquiteturais de aplicações de *software* consolidadas no mercado.

Devido à complexidade dos assuntos abordado, também foram consultadas outras fontes de informações tais como: artigos, jornais, livros, monografias, dissertações e teses de doutorado, a fim de se obter um maior entendimento quanto ao cenário.

Posteriormente, com base nos estudos, pesquisas desenvolvidas e pontos levantados, utilizou-se o método hipotético-dedutivo para formular um processo computacional, capaz de entender informações não estruturadas disponibilizadas pelo Diário Oficial do Estado do Pará e estruturá-las, disponibilizando um novo modelo que possibilite a utilização de processos computacionais na análise sistemática. A metodologia proposta foi baseada e estruturada em 5 etapas, sendo elas: obtenção de dados, marcação, estruturação, raspagem e persistência.

1.5 ESTRUTURA DO TRABALHO

O trabalho consiste em um primeiro capítulo contendo a introdução, onde são apresentados o tema, o problema de pesquisa, a pergunta que norteou o estudo e os objetivos. No segundo capítulo, o referencial teórico aborda acerca do – Cenário atual e Produção de dados; Dados, Informação e Conhecimento; *Big data* e Extração de conhecimento; Tipos de dados; Transferência de dados e *Application Programming Interface* (API); Extração de dados e *scraping*; Mineração de dados; Inteligência artificial; Lei da transparência e controle social – enquanto que no terceiro capítulo, apresenta-se o desenvolvimento da solução,

[3] Web Scraping é uma técnica utilizada para extrair informações na Internet com o intuito de fazer análises para tomadas de decisão (TECHNOPEDIA, 2020).

elencando o processo de estruturação da informação, arquitetura de software e as tecnologias da implementação. O quarto capítulo discorre sobre a análise dos dados, enquanto que o quinto, sobre a discussão dos dados; seguido das considerações finais.

Capítulo 2

Referencial Teórico

Este capítulo apresenta a fundamentação teórica relacionada à proposta deste estudo, abordando os conceitos básicos para o entendimento deste trabalho.

2 REFERENCIAL TEÓRICO

A seguir, o presente trabalho aborda diversos assuntos importantes para a correta construção da solução proposta, conceitos que perpassam pelo entendimento e desenvolvimento do modelo.

2.1 REVISÃO DA LITERATURA

Neste capítulo é apresentada a fundamentação teórica relacionada à proposta deste estudo, abordando os conceitos básicos sobre: o cenário atual e a produção de dados; *Big Data* e extração de conhecimento; dados, informação e conhecimento; tipos de dados; transferência de dados e APIs; extração de dados e *scraping*; mineração de dados; inteligência artificial, além de uma discussão sobre Lei da transparência e controle social.

2.1.1 Cenário atual e Produção de dados

A Revolução Digital, também conhecida como a Terceira Revolução Industrial, refere-se aos processos associados à passagem da tecnologia eletrônica mecânica e analógica para a eletrônica digital, iniciada entre o final dos anos 1950 e o final dos anos 1970, com expansão do uso de computadores digitais e a constituição de arquivos digitais, processo que segue até os dias atuais, marcando o início da Era da Informação (SCHOENHERR, 2004).

Dados sempre foram produzidos e consumidos ao longo da história. No entanto, nem sempre foram de fácil acesso ou facilmente interpretáveis, como hoje. A produção de dados tem aumentado constantemente, isso deve-se a grande abundância de dispositivos digitais presentes em casas, ambientes corporativos e espaços públicos, mídias sociais e da Internet das Coisas – revolução tecnológica que tem como objetivo conectar os itens usados do dia a dia à rede mundial de computadores (KITCHIN, 2014).

Ao longo das últimas décadas, a quantidade de dados gerados tem crescido de forma exponencial. O surgimento da Internet aumentou de forma abrupta a quantidade de dados produzidos, a popularização da Internet das Coisas ocasionou rapidamente a mudança da era do *terabyte* para a era do *petabyte*. Em 2015, a junção de todos os dados gerados no planeta Terra chegou a marca dos *zetabytes*[4], e segundo a International Business Machines

[4] Zettabyte é uma unidade de informação ou memória, correspondendo a 1180591620717411303424 Bytes.

Corporation, desde 2008, é gerado mais de 2,5 quintilhões de *bytes* todos os dias (MULLER *et al*. 2013).

A maneira mais profícua de disponibilizar os dados na Internet é através de um formato específico para análise, contudo, atualmente, dados são divulgados de forma diferente. Portanto, combiná-los com outros dados ou explorá-los de maneira independente é uma tarefa crucial, e de difícil execução para a computação moderna.

Em meio a essa avalanche de dados provinda das mais variadas origens como *smartphones*, *tablets*, *TVs*, computadores, *notebooks*, *smartwatches*, nasce o termo *Big Data*, com o intuito de extrair valor dessa quantidade, cada vez mais crescente, de dados não estruturados gerados exponencialmente a cada segundo.

2.1.2 Dados, Informação e Conhecimento

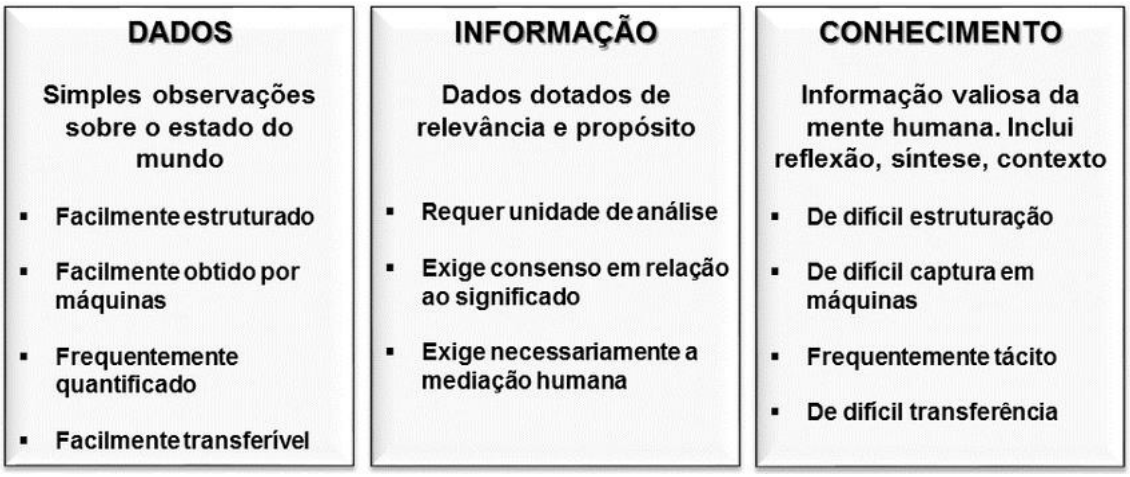
Segundo Davenport e Prusak (1998) e Moreira (2005), os dados são fatos distintos e objetivos ou eventos isolados. No entanto, quando é discorrido em relação à sua importância, os dados em si não são dotados de relevância, propósito e significado. Todavia, esses dados são importantes porque são a matéria-prima essencial para a criação da informação (SANTOS, 2001).

A informação é um conjunto de dados organizados em um contexto. Se os dados agrupados geram sentido para quem o analisa, eles passam a ser o valor de um determinado evento ao qual se refere. O valor de uma informação está de acordo com a qualidade em que é disponibilizada, reduzindo ou aumentando a probabilidade de interpretação ambígua pelo emitente. Quanto mais precisa, mais valiosa ela se torna (OECD, 2008).

O conhecimento é resultado de várias informações organizadas de forma lógica o suficiente para criar um evento, tornar possível um evento ainda não conhecido ou entender um evento e suas causas. O conhecimento é uma informação valiosa sendo produto de reflexão e síntese (OECD, 2008).

> "O conhecimento, refere-se à habilidade de criar um modelo mental que descreve o objeto e indique as ações a implementar, e decisões a tomar" (REZENDE, 2003).

Ilustração 2 - Dados, informação e conhecimento



Fonte: Davenport e Prusak (1998)

Os dados geralmente são assumidos como o conceito menos abstrato, a informação o seguinte e o conhecimento, o mais abstrato (MITRA, 2011). O processo de transformação de dados em informações se dá por meio da interpretação; por exemplo, a temperatura da Antártida é comumente considerada "dados"; suas características geológicas e climáticas podem ser consideradas "informação"; já as considerações acerca de formas mais eficazes para a diminuição do degelo neste local são consideradas "conhecimento".

2.1.3 Big Data e Extração de conhecimento

Atualmente, existem várias definições para o termo *Big Data*. Manyika *et al*. (2013) consideram *Big Data* como sendo o conjunto de dados cujo tamanho é maior do que a capacidade que as ferramentas de *software* de banco de dados[5] tradicionais têm para capturar, armazenar, gerenciar e analisar. Taurion (2013) define *Big Data* como sendo grandes volumes de dados que chegam de uma variedade de fontes, em alta velocidade, com veracidade e que agregam algum tipo de valor a um negócio. Gupta *et al.* (2012) resumem *Big Data* como sendo um grande volume de dados que excede a capacidade de processamento dos bancos de dados convencionais. No entanto, independentemente de qualquer definição formal, o fato é que a escalabilidade do volume e variedade de dados vem trazendo dificuldades para o processamento dessas informações, portanto a necessidade abrupta da extração de informação de qualquer meio que o produz, pode ser definido como big data.

[5] Banco de dados é uma coleção de dados inter-relacionados, representando informações sobre um domínio específico, Korth (2004).

Neste cenário também nascem outros conceitos como o *Data Science* que eleva a capacidade de extração de conhecimento, tomando como base os dados. *Data Science* envolve princípios, processos e técnicas para compreender fenômenos por meio da análise (automatizada) de dados, cujo objetivo primordial é o aprimoramento da tomada de decisão, onde a informação de uma fonte pode ser cruzada com uma infinidade de fatores externos a fim de trazer algum conhecimento que possa ser levado como vantagem em um dado negócio (PROVOT & FAWCETT, 2016).

Portanto, ao se pensar na geração de conhecimento de dados, o real problema não se resume somente ao poder de processamento em massa de uma quantidade gigantesca de dados, mas também se preocupa no processo de geração de valor agregado aos dados.

Extrair informações realmente úteis neste cenário é extremamente difícil, tanto pela diversidade de dados, quanto pelas suas origens, onde informações podem estar em diversos bancos de dados, ou em infinidades de arquivos de texto como *Portable Document Format* (PDF), Word ou Excel, além de arquivos de mídias como, imagens, vídeos e áudios. Tentar encontrar algo relevante em meio a este oceano de informações é de fato uma tarefa extremamente difícil e custosa até mesmo para processos computacionais.

2.1.4 Tipos de dados

Dados podem ser classificado em três categorias de acordo com sua estrutura, sendo elas: dados estruturados, semiestruturados e não estruturados (MONTEIRO, 2019). Os dados estruturados compõem a menor parcela de informações geradas, nesta categoria, uma das propriedades principais é a definição de uma estrutura padronizada pouco flexível e bem definida, como arquivos de banco de dados ou arquivos *Comma-separated values* (CSV)[6]. Um formulário de cadastro é um exemplo de estrutura rígida, onde os campos são bem definidos. O campo **nome** de um formulário pode ser definido como "textual"; já o campo **idade** pode ser definido do tipo "numérico e inteiro"; a validação para o campo **e-mail** deve possuir conteúdo "textual admitindo caracteres especiais como arroba" e por final, o campo **CPF** é obrigatoriamente "textual admitindo somente números e limitado a 11 caracteres" (MONTEIRO, 2019).

Dados não estruturados compõem a maior parcela das informações existentes e se comportam de maneira completamente oposta aos dados estruturados, dessa forma, nesta

[6] CSV é um formato de texto que estrutura seus valores separando-os por vírgulas.

categoria os dados não possuem um padrão bem definido, sendo difíceis de serem processados ou serializados[7]; texto, imagens, vídeos e áudio são alguns exemplos de dados não estruturados. Segundo o artigo *Unstructered Data and The 80 Percent Rule* (2008), "80% das informações mais relevantes para os negócios estão neste formato [...] e para identificar bastando olhar para os lados e verificar quantas horas são gastas escrevendo e-mail, redigindo relatórios ou artigos, em conversas, ouvindo áudios e vídeos".

A categoria de dados semiestruturados é o meio termo entre os dados estruturados e os não estruturados, onde arquivos no formato *eXtensible Markup Language* (XML)[8] ou *JavaScript Object Notation* (JSON)[9] são exemplos de dados semiestruturados, apresentando uma flexibilidade intermediária para a manipulação de dados (MELLO, 2000).

Ilustração 3 – Diferença entre as Estruturas de Dados

Fonte: Os autores

A ilustração 3 demostra o comparativo de uma lista de empregados em três tipos de estruturas de dados, onde como exemplo de dados não estruturados é demostrado um texto simples com formatação; no exemplo de dados estruturados é demostrado um esquema de

[7] Serialização é o processo de transformar um objeto em uma sequência de *bytes* para o envio para outro sistema computadorizado de modo que na recuperação do objeto preserve seu estado original (OLIVEIRA *et al.*, 2002).
[8] XML é uma recomendação para gerar linguagens de marcação para necessidades especiais. Sendo um formato para a criação de documentos com dados organizados de forma hierárquica (W3C, 2008).
[9] JSON é um formato compacto para troca de dados simples e rápida entre sistemas no formato atributo-valor, criado em 2000 (INTERNATIONAL ORGANIZATION OF NORMALIZATION, 2017).

um banco de dados seguido de suas restrições; e por último são exemplificados dados semiestruturados através dos exemplos de JSON e XML respectivamente.

2.1.5 Transferência de dados e APIs

A *Application Programing Interface* (API) é definida por Fraga *et al*. (2011) como uma interface projetada por um sistema, com o objetivo de permitir que outros sistemas possam interagir entre si, possibilitando assim, a sua comunicação; em outras palavras pode-se definir uma API como uma interface de comunicação, que fornece um conjunto de operações, ferramentas e protocolos, acessíveis por meio de um protocolo bem estruturado e pré-definido, e cujo usuário (outro sistema) deve conhecer a assinatura de suas funções, gerando uma comunicação de sistema para outros sistemas.

Uma API pode ser empregada em qualquer contexto de desenvolvimento de *software* cujo objetivo de sua utilização é abstrair o processo interno de um dado sistema para com outro; em outras palavras, toda a complexidade do processamento é abstraída e somente é exposto o resultado final através da interface, funcionando semelhante a duas caixas pretas cujo único conhecimento que uma caixa possui da outra são suas interfaces.

Assumindo que uma API é uma camada de *software* semelhante a um contrato não flexível, cujos sistemas interessados em uma dada comunicação devem cumpri-lo, pode-se afirmar que devido à grande complexidade de desenvolvimento de *softwares* e dependendo da natureza do problema cujo sistema busca resolver, não é recomendado e muito menos viável que um *software* seja programado em todas as camadas de desenvolvimento e dessa forma, o *software* reuse outros *softwares* para executar uma dada ação. Um exemplo disso é a utilização do *login* com as credenciais do Facebook ou da Google por diversos sites, que ao invés de implementar todo um sistema de gestão e autenticação de usuários, utilizam uma API disponibilizada pelo Google ou Facebook, para obter as informações de um usuário. Também pode-se exemplificar o processo da utilização de API em nível de linguagens de programação, onde o processo de armazenamento de dados é feito por outro Sistema Gerenciador de Banco de Dados (SGBD). Nesse exemplo a linguagem de programação "não conhece" o banco de dados e utiliza uma API disponibilizada pelo SGBD para requisitar armazenamentos e recuperações de informações dentro do banco de dados. Ainda pode-se citar outros exemplos em mais baixo nível, como uma linguagem de programação de alto nível como Java, Python, Kotlin, Swift, Javascript e outras, não se comunicando diretamente

com o *hardware* necessitando de linguagens de baixo nível para fazer a disponibilização de memória e recursos, geralmente chamadas de *Software Development Kit* (SDK)[10].

Ao acessar um sistema *web* é quase invisível a utilização de APIs na comunicação entre o *front-end*[11] e o *back-end*[12] para o usuário que o utiliza. Apesar disso, APIs são comumente empregadas neste aspecto, cuja transferência de dados entre os sistemas é feita utilizando uma estrutura de dados bem definida, protocolos rigidamente estruturados, mantendo a ambiguidade ao mínimo, sendo adequadas para processos automatizados por computadores (SUSANNE, 2015).

2.1.6 Extração de dados e *Scraping*

"A coleta automatizada de dados na Internet é tão antiga quanto a própria Internet" (MITCHELL, 2015). Na *web* moderna, nem sempre existe a disponibilização de dados através de uma API. Isso se dá devido à Internet ser uma mídia, principalmente de conteúdo visual, feita e projetada para consumidores finais humanos (SUSANNE, 2015), ou seja, não é esperado que todos os dados disponíveis na Internet possam ser facilmente utilizados por outros *softwares*, e sim facilmente consumidos por pessoas. Neste cenário, onde a minoria dos aplicativos *web* disponibilizam APIs ou outros meios facilitadores para a comunicação com outros sistemas, a raspagem de dados (*data scraping*) é uma solução frequentemente utilizada nestes casos.

Scraping é definido por Mitchell (2015) como sendo a prática de coletar dados através de qualquer meio que não seja um programa interagindo com uma API. Geralmente, a raspagem de dados é utilizada como último recurso, por ser classificada como uma técnica deselegante e somente utilizada quando não existe outro mecanismo de intercâmbio de dados (FREITAS, 2020), pois os *displays*[13] de saída (destinados ao consumo humano) geralmente alteram a estrutura com frequência. Seres humanos podem lidar facilmente com alterações em posicionamento, tamanho, cores e mudança de fluxos. Todavia, um programa de

[10] *Software Development Kit* é um conjunto de ferramentas de desenvolvimento de software que permite a criação de aplicativos para um certo pacote de software, plataforma de hardware, sistema de computador, console de videogame ou sistema operacional. (BEAL, 2016)

[11] Front-end é o meio ao qual o usuário interage com o sistema, associada a Interface Gráfica (SOUTO, 2019).

[12] Back-end é a camada de servidor, associado a regra de negócio, processamento e acesso a banco de dados (SOUTO, 2019).

[13] Displays são tipo de interface humano-computador que transmite informação de modo visual ou táctil, adquirida, armazenada ou transmitida sob várias formas (LEMLEY, 2012). Na web, pode ser tomado como qualquer formato de visualização de conteúdo.

computador pode não conseguir obter resultados se a posição ou o formato forem minimamente alterados.

O *data scraping* é uma técnica computacional na qual um programa extrai dados de uma saída legível somente para humanos, proveniente de um serviço ou aplicativo. Os dados extraídos geralmente são minerados e estruturados em um formato padrão como CSV, XML ou JSON, para possibilitar manipulação posteriormente (SUSANNE, 2015).

O *web scraping* é uma modalidade de *data scraping*, sendo essencialmente uma forma de mineração de dados aplicado em páginas *web* com o intuito de extrair dados textuais úteis a partir de arquivos de texto, tais como HTML (*Hyper Text Markup Language*) e XHTML (*Extensible Hypertext Markup Language*) (TECHNOPEDIA, 2020). Entretanto, a maioria das páginas *web* são projetadas para usuários humanos e não para uso automatizado, sendo um obstáculo para a obtenção de maneira facilitada.

Para executar *web scraping* de maneira automatizada pela *web*, geralmente se utiliza um rastreador de rede (*web crawler*) para indexar páginas *web* de maneira metódica e automatizada (PATIL, 2016). O *web crawler*, também conhecido como *bots*, *web spiders*, *web robot* ou *web scutter*, é uma modalidade de programa de computador que navega pela Internet indexando páginas visitadas para um pós-processamento por outro agente (GLOBALAD, 2017). Os rastreadores também podem ser utilizados para as tarefas de manutenção automatizadas em um *site* da rede, como verificar os *links* ou validar códigos HTML.

2.1.7 Mineração de dados

No trabalho de Camilo *et al*. (2009), a Mineração de Dados foi considerada uma área interdisciplinar, tendo três pilares fundamentais: Estatística, Banco de Dados e Aprendizado de Máquina, cujos embasamentos são:

a) Segundo a perspectiva estatística de Hand *et al*. (2001), que define a Mineração de Dados como sendo análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensível ao dono dos dados.
b) Segundo a perspectiva de banco de dados de Cabena *et al*. (1998), cuja definição de Mineração de Dados é dada em um campo interdisciplinar que junta técnicas de reconhecimento de padrões, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados.

c) Segundo a perspectiva do aprendizado de máquina de Fayyad *et al*. (1996), onde é definida como sendo um passo no processo de Descoberta de Conhecimento, consistindo na realização da análise dos dados e de seus padrões e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem e identificam um conjunto de padrões de dados.

2.1.8 Inteligência artificial

Apesar de que seja comumente associado inteligência artificial (IA) aos mais modernos algoritmos de aprendizado de máquina, como *deep learning*, redes neurais artificiais e algoritmos genéticos, a definição inicial de inteligência artificial introduzida por John McCarthy na famosa conferência de Dartmouth em 1956 foi exemplificada como sendo "fazer uma máquina comportar-se de tal forma que seja chamada inteligente caso fosse este o comportamento de um ser humano" (MEDEIROS, 2018).

A maior parte das definições acerca de IA podem ser categorizadas em sistemas que: "pensam como um humano; agem como um humano; pensam racionalmente ou agem racionalmente" (RUSSEL *et al*., 2003). Existe também uma separação entre duas perspectivas de inteligência artificial, a IA Forte e IA Fraca com aplicações bastante distintas entre as mesmas (GRANATYR, 2017).

Em resumo, a inteligência artificial Fraca não é capaz de raciocinar e basicamente simula a inteligência, porém, não é de fato inteligente devido a necessidade de especialistas humanos para fornecer o conhecimento para que o software consiga executar e tomar suas decisões (GRANATYR, 2017).

Por outro lado, a inteligência artificial Forte está relacionada à criação de máquinas que tenham autoconsciência e que possam pensar; e não somente simular raciocínios. Este modelo ainda está desenvolvendo-se no mercado, sendo que ainda existem controvérsias sobre a sua existência (GRANATYR, 2017).

2.1.9 Lei da transparência e Controle social

O Estado Democrático nasce para suprir direitos básicos da população, tais como: saúde, segurança, educação e moradia. O exercício de tais direitos, demanda a disponibilidade de recursos financeiros que são arrecadados da população por meio dos impostos, taxas e contribuições (CONCEIÇÃO, 2010).

No entanto, devido à natureza complexa da administração do Estado como um todo, e todas as suas subdivisões e departamentos, torna-se extremamente difícil assegurar que de fato todos os direitos citados serão ofertados sem desperdícios, erros ou desvios. Dessa forma, torna-se imprescindível que o Controle Social recaia principalmente sobre os recursos financeiros, pois neles está indicada, ou não, a implementação destes direitos.

O controle social conta com vários dispositivos legais implantados, entre eles estão a Constituição Federal de 1988, denominada de Constituição Cidadã, e a Lei Complementar nº 101, de 04 de maio de 2000, também chamada de Lei de Responsabilidade Fiscal – LRF (CONCEIÇÃO, 2010).

No artigo 1º da Carta Magna[14] é previsto uma sociedade participativa de decisões políticas de diversas formas, este pensamento é principalmente identificado no parágrafo que descreve: "todo poder emana do povo, que o exerce por meio de representantes eleitos ou diretamente, nos termos desta constituição. " (BRASIL, 2008). Logo, como parte indispensável desta participação, o exercício representativo do poder pelo povo é a mais comum forma de controle social (CONCEIÇÃO, 2010).

> Percebe-se que a base dos princípios relativos aos instrumentos que propiciam o controle social encontra-se gravada no Título II da Lei Maior. Assim sendo, pode ser abstraído então que o Controle Social é um direito fundamental do cidadão e deve ser garantido pelo Estado (CONCEIÇÃO, 2010, p. 11).

Por sua vez, a Lei de Responsabilidade Fiscal trata principalmente da gestão dos recursos públicos nos três níveis de governo: Municipal, Estadual e Federal. A LRF também denomina as Leis orçamentárias de "instrumentos de transparência da gestão fiscal" onde devem ser amplamente divulgadas (CONCEIÇÃO, 2010). O art. 48 da LRF destaca principalmente a participação popular e disponibilidade da informação, determinando o incentivo à participação popular" por intermédio de audiências públicas, e a "liberação ao pleno conhecimento e acompanhamento da sociedade, em tempo real, de informações circunstâncias sobre a execução orçamentária e financeira, em meios eletrônicos de acesso público. (CASA CIVIL, 2000).

Lei de Responsabilidade Fiscal tem como base a transparência, que se revela como um instrumento democrático que busca principalmente o fortalecimento da cidadania, servindo para tornar mais eficiente o sistema de controle das contas públicas (CONCEIÇÃO, 2010), ou seja, servindo de requisito para o controle social.

[14] Carta Magna é a Constituição Federal de 1988.

A LRF também retrata a transparência como sendo um dos princípios da gestão fiscal, e admite a **publicidade** e a **compreensibilidade** das informações sendo fatores **essenciais** para o mesmo. Tal como é ressaltado por Conceição (2010), onde é defendido que a mera divulgação do conteúdo não compreensível para a sociedade não é transparência, como também não é informação compreensível.

Em síntese, a Controladoria Geral da União (2009) resume a discussão descrita nos parágrafos anteriores como sendo uma das obrigações da administração pública:

> "[...] dar transparência aos seus gastos, disponibilizando acesso à informação sobre recursos públicos transferidos ao Estado e aos municípios. Os Estados e municípios, de posse das informações acerca das despesas públicas, têm o direito e o dever de fiscalizar a sua regularidade e sua eficiência. A população também deve ter acesso as despesas públicas e deve ser estimulada a participar da fiscalização, mediante controle social. A prefeitura deve incentivar a participação popular na discussão de planos e orçamentos. Suas contas devem ficar disponíveis para qualquer cidadão segundo a Casa Civil com a Lei de Responsabilidade Fiscal, art. 48 e 49 do ano 2000".

Finalmente, pode ser concluído que todos os meios e ferramentas citadas visam informar o cidadão acerca dos atos de gestão de recursos públicos a fim de torna-lo parte essencial do processo de fiscalização, e este ato para o Controle Social significa transparência.

2.2 TRABALHOS RELACIONADOS

Existem várias linhas de pesquisa para solucionar problemas relacionados a extração de informações, Paiva e Revoedo (2016) propõem a aplicação de técnicas de tratamento de dados que permitam a estruturação de forma mais clara e elucidativa para a população. A aplicação de técnicas de tratamento de dados, permite a estruturação dos mesmos em forma mais claras e sintetizadas, seguindo a linha de pesquisa, que toma como base o desenvolvimento de ferramentas capazes de processar grandes volumes de dados e que permita uma visualização condensada dos mesmos, propondo uma metodologia de tratamento de informação para a obtenção de indicadores sobre gastos públicos, utilizando técnicas de programação paralela baseada no paradigma MapReduce[15].

[15] MapReduce é um modelo de programação desenhado para processar grandes volumes de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes. (DEAN; GHEMAWAT, 2004)

Nos trabalhos de Carvalho *et al*. (2013, 2014) utilizam uma metodologia para aumentar a informatividade do Portal de Transparência do Governo Federal, sugerindo a criação de um banco a partir do tratamento textual e da utilização de ferramentas *Extract, Transform e Load* (ETL)[16], para auxiliar o tratamento dos dados e em seguida, a aplicação de técnicas de mineração de dados.

Por outro lado, todos os trabalhos citados partem da premissa que os dados já foram obtidos e armazenados, focando principalmente no processamento de dados para agregar valor, não utilizando qualquer processo de extração de dados na Internet.

Apesar deste trabalho ter como objetivo o desenvolvimento de um modelo para extração de dados em informações não estruturadas, é comumente associado a esta tarefa o processo de análise, visto que a análise é executada partindo da premissa que existe a posse de uma base de dados. Logo, ferramentas de mercado neste sentido detém recursos que possibilitam estruturação de dados para um modelo computacional, bem como também possibilitam manipulação e análises dos dados.

[16] ETL é um processo onde a extração, carregamento e transformação é uma abordagem alternativa, projetada para execução de processamento externo ao banco de dados, de modo a aprimorar a performance.

Capítulo 3

Desenvolvimento da solução

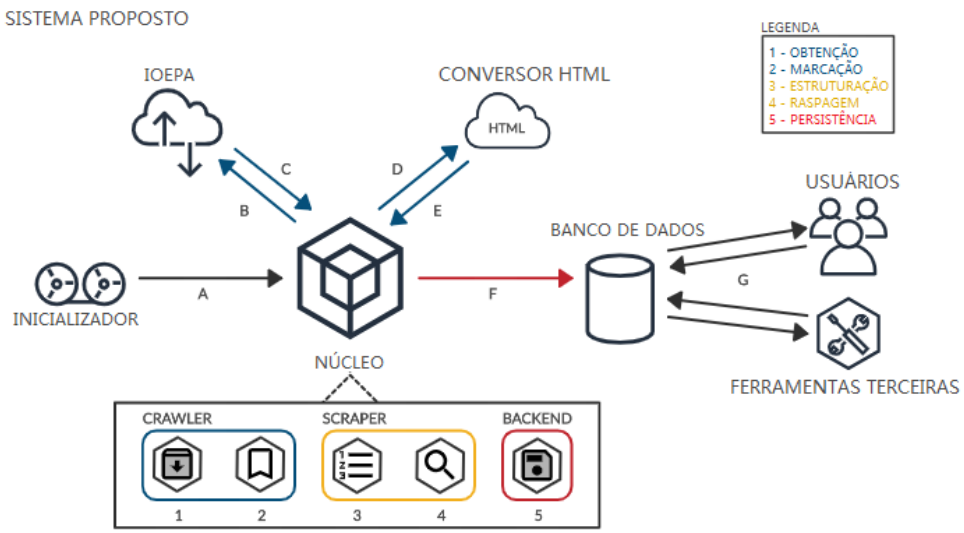
Este capítulo apresenta os processos fundamentais para a estruturação das informações dispostas no Diário Oficial do Estado do Pará, a Arquitetura de Software modelada para abstrair os processos fundamentais e as tecnologias tomadas como base para a implementação do mesmo.

---

3 DESENVOLVIMENTO DA SOLUÇÃO

Neste capítulo é proposto uma de ferramenta para extração de dados do Diário Oficial do Estado do Pará utilizando um modelo computacional baseado em 5 processos fundamentais, sendo eles: obtenção de dados, marcação, estruturação, raspagem e persistência, sendo delegado para 3 sistemas: *Crawler*, *Scraper*, *Persistence*. Os componentes de *software*, suas propriedades externas e seus relacionamentos com outros *softwares* é apresentado no diagrama abaixo.

Ilustração 4 - Componentes de *software*



Fonte: Os autores

3.1 PROCESSO DE ESTRUTURAÇÃO DA INFORMAÇÃO

O processo de estruturação da informação foi elaborado seguindo o percurso que a informação percorre dentro do sistema até atingir o resultado esperado (a estruturação da informação). Tal fluxo se caracteriza em gerenciar todas as etapas do pré-processamento e estruturação dos dados, interligando-as e garantindo a execução de todos os processos de forma autônoma.

Ilustração 5 - Etapas da solução proposta

Fonte: Os autores

3.1.1 Obtenção de dados

Os dados utilizados nos sistemas estão em formato único, disponibilizados em PDF pelo IOEPA (Imprensa Oficial do Estado do Pará) disponível para download no portal digital http://www.ioepa.com.br/portal/ através das publicações diárias do Diário Oficial.

3.1.2 Marcação

Após a obtenção do arquivo em PDF no Diário Oficial, é necessário convertê-lo para outro meio, visando o melhor "entendimento" do mesmo por um computador. Nesta fase pode-se converter o arquivo originalmente em PDF para uma infinidade de tipos de arquivos. Uma possibilidade é a conversão para arquivo de texto simples (.txt), entretanto, este formato dificultaria o pré-processamento das informações, uma vez que o Diário Oficial possui uma distribuição de informação bastante complexa, variando tamanhos, fontes, cores, posicionamento e sentido.

Para considerar tal distribuição complexa, foi adotado o formato *Hypertext Markup Language* (HTML) por fazer uso de diversas propriedades de estilização e marcação das informações, além da facilidade de manusear tais atributos, permitindo mais flexibilidade quanto a identificação dos dados. Para executar a conversão do formato PDF para HTML utilizou-se o conversor online disponível em https://www.pdftohtml.net/.

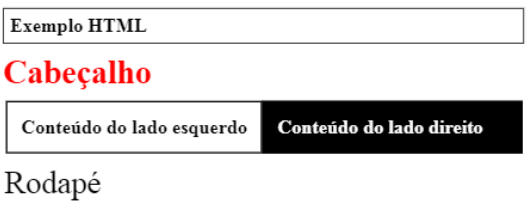
A linguagem HTML utiliza de *tags* (marcadores ou *tokens*) que sinalizam o início e o fim de um determinado dado disposto em um documento HTML, denotando uma semântica estrutural para textos em geral, como parágrafos, listas, botões e outros, sendo seu diferencial a possibilidade de codificar hipertexto como: estilização, *links*, fotos, vídeos e *scripts*.

Ilustração 6 - Trecho de código HTML

CÓDIGO HTML

```
<!DOCTYPE html>
<html>
  <head> </head>
  <body>
    <h1 style="border: 1px solid black; padding: 8px;">Exemplo HTML</h1>
    <h3 style="color: red; font-size: 32pt;">Cabeçalho</h3>
    <div style="display: flex; margin: 16px; display: flex; flex: 1;">
      <div style="font-weight: bolder;border: 2px solid #000; flex: 1;padding: 18px;">
        Conteúdo do lado esquerdo
      </div>
      <div style="font-weight: bolder; border: 2px solid #000;
                  background-color:  #000; flex: 1; color: #fff; padding: 18px;">
        Conteúdo do lado direito
      </div>
    </div>
    <div style="position: absolute; bottom: 8px; font-size: 20px;">Rodapé</div>
  </body>
</html>
```

RESULTADO



Fonte: Os autores

A estratégia nessa etapa para a marcação do arquivo PDF do Diário Oficial é a conversão deste para um arquivo HTML, possibilitando através das propriedades de estilização e *tags*, a identificação dos diversos padrões nas informações dispostas do documento e o fácil acesso a interpretação dos dados marcados.

3.1.3 Estruturação

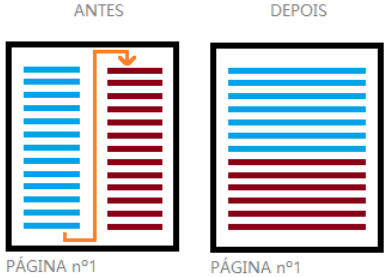
A etapa de estruturação consiste na organização do conteúdo e na estruturação dos dados. O Diário Oficial é dividido em páginas e cada página pode ser dividida em uma ou mais colunas como a ilustração 7 demonstra.

Ilustração 7 - Folhas de um arquivo disponibilizado pelo Diário Oficial

Fonte: Os autores

Na ilustração 7 é possível analisar claramente que o *layout* das páginas do arquivo é composto majoritariamente por duas colunas e separado em páginas. Dessa forma a primeira etapa do processo de estruturação é unificar as colunas colocando o conteúdo do lado direito abaixo do conteúdo do lado esquerdo a fim de manter o fluxo de leitura da informação.

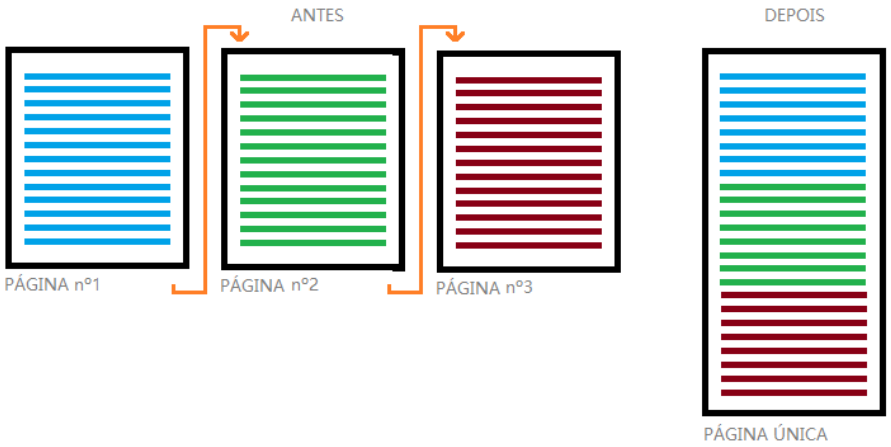
Ilustração 8 - Unificação de colunas em uma página



Fonte: Os autores

Para cada página no documento HTML, é unificado as colunas como mostra a ilustração 8. Todavia, somente esta técnica não assegura que todo o conteúdo analisado estará na mesma página, podendo parte do mesmo estar no final da coluna da direita e o restante na próxima página. Dessa forma, para solucionar esse problema foi necessário unificar as páginas como demonstra a ilustração 9.

Ilustração 9 - Processo de unificação de página de um arquivo



Fonte: Os autores

Ao aplicar a técnica descrita na ilustração 9, obtém-se uma única página HTML com uma única coluna, onde a mesma já se encontra minimamente estruturada, facilitando a leitura e identificação de seus conteúdos.

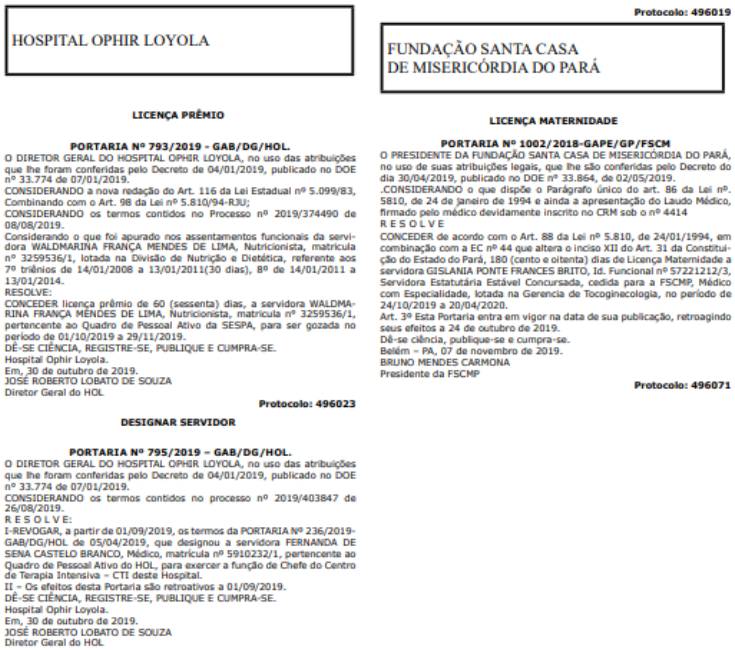
3.1.4 Raspagem

Esta etapa consiste na extração da informação da página HTML gerada no tópico anterior. Para isso, primeiramente, é necessário identificar os dados a serem contabilizados, bem como a disposição desses dados no Diário Oficial do Estado do Pará.

Sabendo que o Diário Oficial é o meio de comunicação pelo qual a Imprensa Oficial do Estado torna público todo e qualquer assunto acerca do âmbito do estado, pode-se formular um processo capaz de identificar padrões e extrair informações do mesmo.

A primeira etapa do processo identifica todos os órgãos e secretaria contidos no documento. Na ilustração 10 é mostrado um exemplo da disposição das informações no arquivo original.

Ilustração 10 - Divisão de órgãos e secretarias no arquivo PDF



HOSPITAL OPHIR LOYOLA

LICENÇA PRÊMIO

PORTARIA Nº 793/2019 - GAB/DG/HOL.

O DIRETOR GERAL DO HOSPITAL OPHIR LOYOLA, no uso das atribuições que lhe foram conferidas pelo Decreto de 04/01/2019, publicado no DOE nº 33.774 de 07/01/2019.
CONSIDERANDO a nova redação do Art. 116 da Lei Estadual nº 5.099/83, Combinando com o Art. 98 da Lei nº 5.810/94-RJU;
CONSIDERANDO os termos contidos no Processo nº 2019/374490 de 08/08/2019.
Considerando o que foi apurado nos assentamentos funcionais da servidora WALDMARINA FRANÇA MENDES DE LIMA, Nutricionista, matrícula nº 3259536/1, lotada na Divisão de Nutrição e Dietética, referente aos 7º triênios de 14/01/2008 a 13/01/2011(30 dias), 8º de 14/01/2011 a 13/01/2014.
RESOLVE:
CONCEDER licença prêmio de 60 (sessenta) dias, a servidora WALDMARINA FRANÇA MENDES DE LIMA, Nutricionista, matrícula nº 3259536/1, pertencente ao Quadro de Pessoal Ativo da SESPA, para ser gozada no período de 01/10/2019 a 29/11/2019.
DÊ-SE CIÊNCIA, REGISTRE-SE, PUBLIQUE E CUMPRA-SE.
Hospital Ophir Loyola.
Em, 30 de outubro de 2019.
JOSÉ ROBERTO LOBATO DE SOUZA
Diretor Geral do HOL

Protocolo: 496023

DESIGNAR SERVIDOR

PORTARIA Nº 795/2019 – GAB/DG/HOL.

O DIRETOR GERAL DO HOSPITAL OPHIR LOYOLA, no uso das atribuições que lhe foram conferidas pelo Decreto de 04/01/2019, publicado no DOE nº 33.774 de 07/01/2019.
CONSIDERANDO os termos contidos no processo nº 2019/403847 de 26/08/2019.
R E S O L V E:
I-REVOGAR, a partir de 01/09/2019, os termos da PORTARIA Nº 236/2019-GAB/DG/HOL de 05/04/2019, que designou a servidora FERNANDA DE SENA CASTELO BRANCO, Médico, matrícula nº 5910232/1, pertencente ao Quadro de Pessoal Ativo do HOL, para exercer a função de Chefe do Centro de Terapia Intensiva – CTI deste Hospital.
II – Os efeitos desta Portaria são retroativos a 01/09/2019.
DÊ-SE CIÊNCIA, REGISTRE-SE, PUBLIQUE E CUMPRA-SE.
Hospital Ophir Loyola.
Em, 30 de outubro de 2019.
JOSÉ ROBERTO LOBATO DE SOUZA
Diretor Geral do HOL

Protocolo: 496019

FUNDAÇÃO SANTA CASA
DE MISERICÓRDIA DO PARÁ

LICENÇA MATERNIDADE

PORTARIA Nº 1002/2018-GAPE/GP/FSCM

O PRESIDENTE DA FUNDAÇÃO SANTA CASA DE MISERICÓRDIA DO PARÁ, no uso de suas atribuições legais, que lhe são conferidas pelo Decreto do dia 30/04/2019, publicado no DOE nº 33.864, de 02/05/2019.
.CONSIDERANDO o que dispõe o Parágrafo único do art. 86 da Lei nº. 5810, de 24 de janeiro de 1994 e ainda a apresentação do Laudo Médico, firmado pelo médico devidamente inscrito no CRM sob o nº 4414
R E S O L V E
CONCEDER de acordo com o Art. 88 da Lei nº 5.810, de 24/01/1994, em combinação com a EC nº 44 que altera o inciso XII do Art. 31 da Constituição do Estado do Pará, 180 (cento e oitenta) dias de Licença Maternidade a servidora GISLANIA PONTE FRANCES BRITO, Id. Funcional nº 57221212/3, Servidora Estatutária Estável Concursada, cedida para a FSCMP, Médico com Especialidade, lotada na Gerencia de Tocoginecologia, no período de 24/10/2019 a 20/04/2020.
Art. 3º Esta Portaria entra em vigor na data de sua publicação, retroagindo seus efeitos a 24 de outubro de 2019.
Dê-se ciência, publique-se e cumpra-se.
Belém – PA, 07 de novembro de 2019.
BRUNO MENDES CARMONA
Presidente da FSCMP

Protocolo: 496071

Fonte: Os autores

Como pode ser visto na ilustração 10, órgãos e secretarias possuem fonte e estilo diferenciado do restante dos demais conteúdos.

A primeira etapa da raspagem realiza a leitura inteira do arquivo, identificando somente dos órgãos e secretarias contidas no arquivo HTML utilizando sua estilização, tais como: tamanho da fonte, cor, posição e tipo de caixa para identificar cada item. Após esse processo o sistema detém o conhecimento de todos os órgãos e secretarias contidas no arquivo como mostra a ilustração 11.

Ilustração 11 - Listagem de órgãos e secretarias

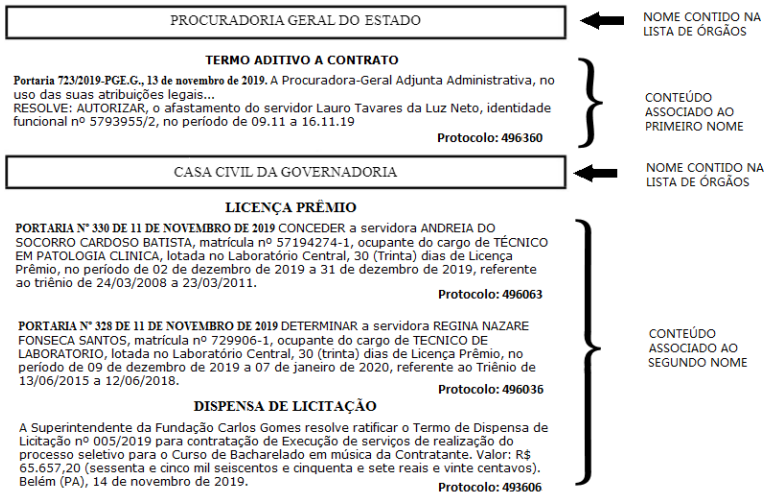
```
[
  "CASA CIVIL DA GOVERNADORIA",
  "PROCURADORIA GERAL DO ESTADO",
  "INSTITUTO DE ASSISTÊNCIA DOS SERVIDORES DO ESTADO DO PARÁ",
  "SECRETARIA DE ESTADO DA FAZENDA",
  "FUNDAÇÃO PÚBLICA ESTADUAL HOSPITAL DE CLÍNICAS GASPAR VIANNA",
  "AGÊNCIA ESTADUAL DE REGULAÇÃO E CONTROLE DE SERVIÇOS PÚBLICOS",
  "CORPO DE BOMBEIROS MILITAR DO ESTADO DO PARÁ",
  "SUPERINTENDÊNCIA DO SISTEMA PENITENCIÁRIO DO ESTADO DO PARÁ",
  "FUNDAÇÃO CARLOS GOMES",
  "SECRETARIA DE ESTADO DE EDUCAÇÃO",
  "SECRETARIA DE ESTADO DE DESENVOLVIMENTO URBANO E OBRAS PÚBLICAS",
  "COMPANHIA DE SANEAMENTO DO PARÁ",
  "SECRETARIA DE ESTADO DE TURISMO",
  "TRIBUNAL DE JUSTIÇA DO ESTADO DO PARÁ"
]
```

Fonte: Os autores

A próxima etapa consiste na associação do conteúdo ao seu respectivo órgão. Para isso, basta que, ao identificar um órgão, o sistema salve todas as linhas seguintes até que

encontre o próximo órgão. Essa etapa é factível devido ao processo de marcação descrito no tópico 3.1.2. A ilustração 12 demonstra esse processo.

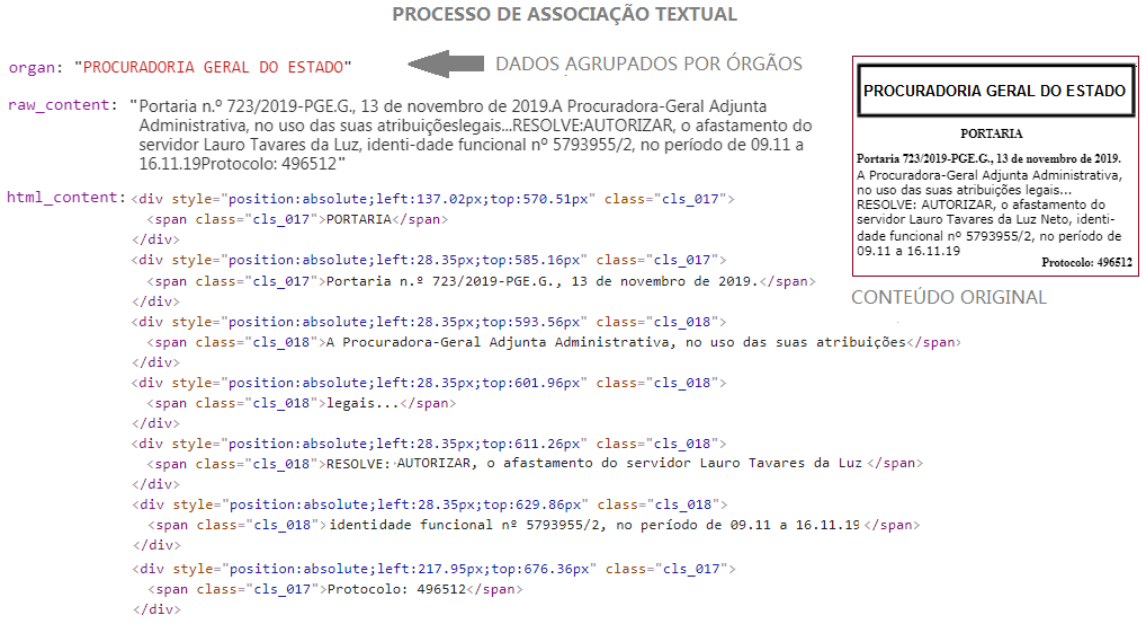
Ilustração 12 - Associação de órgãos e seu respectivo conteúdo



Fonte: Diário Oficial do Estado do Pará - Adaptado

Ao final desse processo, o sistema detém uma lista com todos os órgãos e seus respectivos conteúdos associados, contendo tanto o conteúdo textual original quanto o marcado.

Ilustração 13 - Listagem de órgãos e secretarias associados com o conteúdo

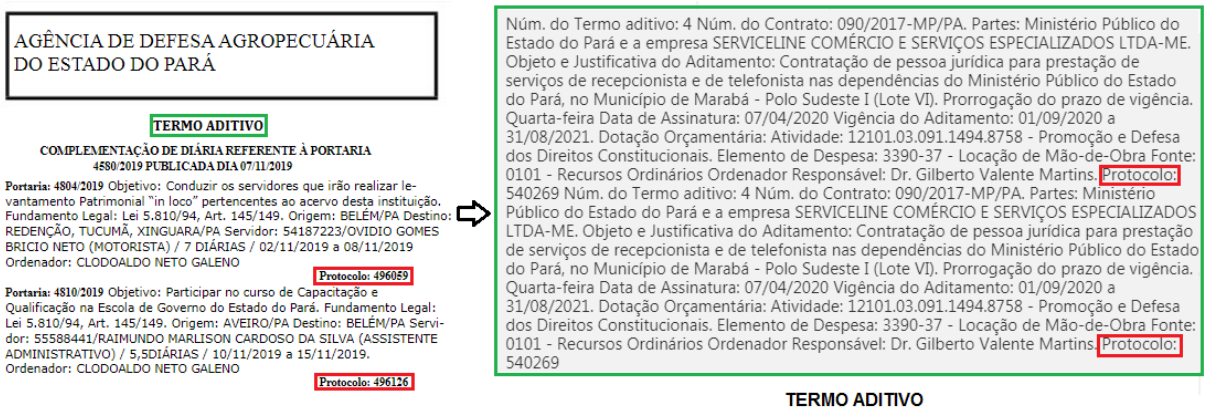


Fonte: Os autores

Após a associação dos conteúdos com seus respectivos órgãos, inicia-se a identificação e classificação do tipo de conteúdo. Inicialmente, foram catalogados 3 tipos de atos públicos, sendo eles: contratos, termos aditivos e dispensas de licitações, porém o sistema possui a capacidade de catalogar outros atos públicos, através de um dicionário de atos que é utilizado como entrada de dados para identificar os protocolos de interesse.

A identificação do tipo de protocolo utilizando o dicionário de interesse citado, agrupa todo texto processado até que encontre outro item do dicionário de interesse. O processo é semelhante ao mostrado na ilustração 12, porém ao invés da pesquisa utilizar a lista de órgãos como delimitadores dentro do texto, é utilizado o catálogo de interesse. A ilustração 14 mostra o agrupamento de termos aditivos de um órgão.

Ilustração 14 - Dicionário de interesse utilizado para associação de protocolos



Fonte: Os autores

Após este segundo agrupamento é necessário contabilizar e delimitar o início e fim de cada protocolo, para isso, foi analisado a arquitetura do Diário Oficial e sendo identificado que o termo “Protocolo: X”, onde “X” representa a numeração do protocolo, se repete no final de todos os protocolos e que, dessa forma, pode ser utilizado como um identificador textual válido tanto para a delimitação de seu conteúdo, quanto para a quantificação de protocolos de um item do catálogo de interesse. A ilustração 15 demonstra a utilização do termo “Protocolo: X” como contador, e sua utilização para separar o texto associado a um dado protocolo.

Ilustração 15 - Protocolos para um órgão

Segunda-feira, 18 DE NOVEMBRO DE 2019 DIÁRIO OFICIAL Nº 34037 □□69□

AGÊNCIA DE DEFESA AGROPECUÁRIA DO ESTADO DO PARÁ

TERMO ADITIVO

COMPLEMENTAÇÃO DE DIÁRIA REFERENTE À PORTARIA 4580/2019 PUBLICADA DIA 07/11/2019

Portaria: 4804/2019 Objetivo: Conduzir os servidores que irão realizar levantamento Patrimonial "in loco" pertencentes ao acervo desta instituição. Fundamento Legal: Lei 5.810/94, Art. 145/149. Origem: BELÉM/PA Destino: REDENÇÃO, TUCUMÃ, XINGUARA/PA Servidor: 54187223/OVIDIO GOMES BRICIO NETO (MOTORISTA) / 7 DIÁRIAS / 02/11/2019 a 08/11/2019
Ordenador: CLODOALDO NETO GALENO
Protocolo: 496059

Portaria: 4810/2019 Objetivo: Participar no curso de Capacitação e Qualificação na Escola de Governo do Estado do Pará. Fundamento Legal: Lei 5.810/94, Art. 145/149. Origem: AVEIRO/PA Destino: BELÉM/PA Servidor: 55588441/RAIMUNDO MARLISON CARDOSO DA SILVA (ASSISTENTE ADMINISTRATIVO) / 5,5DIÁRIAS / 10/11/2019 a 15/11/2019.
Ordenador: CLODOALDO NETO GALENO
Protocolo: 496126

Portaria: 4842/2019 Objetivo: Dar apoio nas ações de controle de foco de raiva. Fundamento Legal: Lei 5.810/94, Art. 145/149. Origem: IGARAPÉ Destino: MARACANÃ/PA Servidor: 541885631/KID STELIO ALMEIDA (AC DE DEFESA AGROPECUÁRIA) / 8,5 DIÁRIAS / 18/11/2019 a 26/11/2019
Ordenador: CLODOALDO NETO GALENO
Protocolo: 496453

Portaria: 4812/2019 Objetivo: Participar no curso de Capacitação e Qualificação na Escola de Governo do Estado do Pará. Fundamento Legal: Lei 5.810/94, Art. 145/149. Origem: SOURE/PA Destino: BELÉM/PA Servidor: 54193783/ELCIDES MIRANDA MORAIS (ASSISTENTE ADMINISTRATIVO) / 5,5DIÁRIAS / 10/11/2019 a 15/11/2019.
Ordenador: CLODOALDO NETO GALENO
Protocolo: 496251

Portaria: 4836/2019 Objetivo: Dar apoio nas ações de controle de foco de raiva. Fundamento Legal: Lei 5.810/94, Art. 145/149. Origem: CASTANHAL/PA Destino: MARACANÃ/PA Servidor: 55586131/PAULO ADRIANO DA SILVA (TÉCNICO AGRÍCOLA) / 8,5 DIÁRIAS / 18/11/2019 a 26/11/2019
Ordenador: CLODOALDO NETO GALENO
Protocolo: 496371

Portaria: 4837/2019 Objetivo: Dar apoio na realização de fiscalização em propriedades rurais. Fundamento Legal: Lei 5.810/94, Art. 145/149. Origem: ULIANÓPOLIS/PA Destino: IPIXUNA DO PARÁ/PA Servidor: 10294017/ZEDEQUIAS RODRIGUES DA SILVA (TÉCNICO AGRÍCOLA) / 2,5 DIÁRIAS / 04/11/2019 a 06/11/2019
Ordenador: CLODOALDO NETO GALENO
Protocolo: 496381

DISPENSA DE LICITAÇÃO

COMPLEMENTAÇÃO DE DIÁRIA REFERENTE À PORTARIA 4580/2019 PUBLICADA DIA 07/11/2019

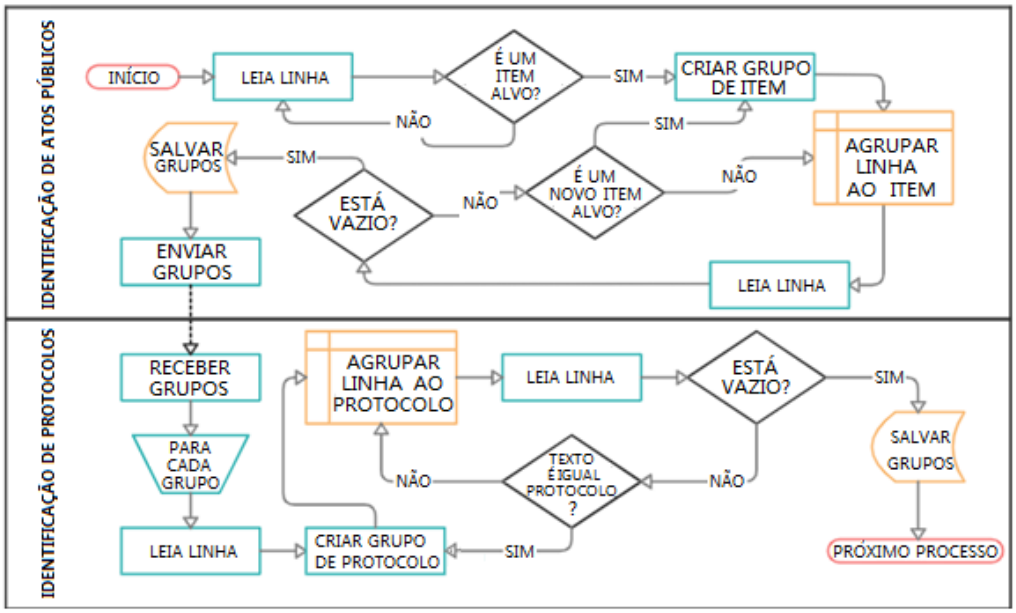
Portaria: 4832/2019 Objetivo: Realizar monitoramento de armadilhas para o levantamento de detecção da mosca da carambola. Fundamento Legal: Lei 5.810/94, Art. 145/149. Origem: BREVES/PA Destino: ANAJÁS/PA Servidor: 5939071/JAQUELINE MENDES DE MELO (AGENTE DE DEFESA AGROPECUÁRIA) / 3,5 DIÁRIAS / 06/11/2019 a 09/11/2019.
Ordenador: CLODOALDO NETO GALENO
Protocolo: 496343

Portaria: 4807/2019 Objetivo: Acompanhar as ações do GTV Açaí. Fundamento Legal: Lei 5.810/94, Art. 145/149. Origem: BELÉM/PA Destino: ABAETETUBA/PA Servidor: 5950874/ SALOMÃO ALMEIDA PEREIRA (GERENTE) / 6,5 DIÁRIAS / 10/11/2019 A 16/11/2019
Ordenador: CLODOALDO NETO GALENO
Protocolo: 496074

Fonte: Diário Oficial do Estado do Pará - Adaptado

Dessa forma, o algoritmo mostrado na ilustração 16 descreve os processos executados anteriormente (ilustração 12 e ilustração 14) com a finalidade de identificar e agrupar todo o texto de um protocolo juntamente com o tipo de ato.

Ilustração 16 - Algoritmo de identificação e agrupamento de atos públicos e protocolos



Fonte: Os autores

A aplicação de todos os processos descritos no tópico 3.1.4 já é suficiente para prover uma estrutura de dados minimamente viável para que seja possível extrair algumas informações importantes para a análise de terceiros. Tais informações como vigência, data de assinatura, valor, partes envolvidas, são de extrema importância para apoiar a fiscalização de maneira automatizada. Todavia a identificação desses termos é consideravelmente mais complexa, visto que o meio pelo qual o conteúdo dos atos públicos é inserido dentro do Diário Oficial não é automatizado, e sim manual, colocando o fator "humano" no processo, inserindo dados manualmente sem um padrão bem definido e passível de falhas como erro de escrita, abreviação de palavras, concordância e demais variações textuais.

Ilustração 17 - Diferença na organização de escrita de atos públicos

SECRETARIA DE ESTADO DA FAZENDA

TERMO ADITIVO A CONTRATO

Secretaria de Estado de Desenvolvimento Urbano e Obras Públicas–CNPJ 03.137.985/0001 90 Pará Segurança LTDA – CNPJ 04.113.174/0001-11 Objeto: Prestação de Serviços de Vigilância./Segurança Patrimonial Armada nos Turnos Diurnos e Noturnos nesta SEDOP, neste Estado. Justificativa: Repactuação de serviços nos termos da cfe. art.65 II, d, da Lei nº 8.666/93 REPACTUAÇÃO: R$ 1.208,49 Data da Assinatura: 01/07/2020 Ordenador Responsável: Benedito Ruy Santos Cabral Secretário de Estado de Desenvolvimento Urbano e Obras Pública

Protocolo: 557787

INSTITUTO DE ASSISTÊNCIA DOS SERVIDORES DO ESTADO DO PARÁ

TERMO ADITIVO A CONTRATO

PROCESSO Nº 2020/193393
Termo Aditivo:4º
Data da Assinatura: 01/04/2020
Vigência: 01/04/2020 a 01/04/2021
Justificativa: Prorrogação do prazo de vigência.
Contrato:126
Exercício:2016
Dotação Orçamentária: 8888- 0261 - 339039
Contratada: CENTRO DE DIAGNÓSTICO LABORATORIAL LTDA- ME (BIO-LAB)
CNPJ n. 03.903.998/0003-94
Endereço: Rua João Pessoa, nº.26, bairro Centro, PA, CEP:68.700-030
Ordenador: BERNARDO ALBUQUERQUE DE ALMEIDA

Protocolo: 540052

Fonte: Diário Oficial do Estado do Pará - Adaptado

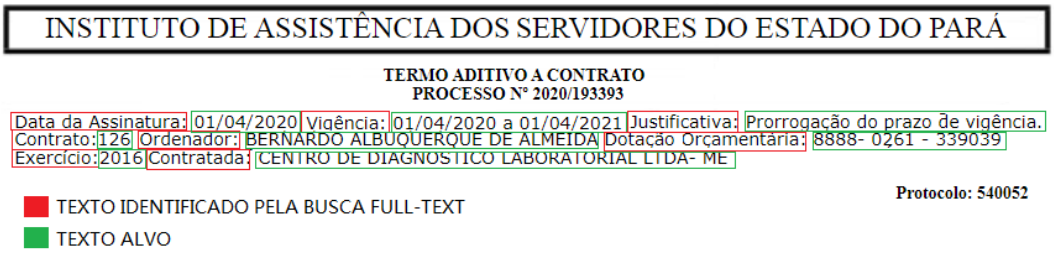
Levando em consideração as possíveis variações textuais que podem ocorrer dentro de um texto, é impossível obter 100% de precisão na identificação textual de qualquer palavra. Todavia, existem algoritmos e estratégias bem consolidadas que tentam minimizar o erro relacionado a identificação textual.

A estratégia de busca *full-text*, é comumente utilizada em sites de e-commerce com o intuito de tentar associar um determinado termo de busca com resultados relacionados ao termo. O *full-text* faz pesquisas, não apenas por uma palavra ou frase exata, mas também por suas variantes verbais (exemplo: correr, correndo, correu) ou ainda estejam estas no singular ou no plural. Outro fator levado em consideração na busca é a remoção de *stopwords*[17] que não agregam valor nenhum a uma frase.

[17] *Stopwords*: Na computação, uma palavra vazia é uma palavra que é removida antes ou após o processamento de um texto em linguagem natural. Essas palavras geralmente não alteram o significado a frase/texto. A aplicação da técnica de remoção *stopwords* na frase: "contratação de pessoa jurídica para prestação de serviço", resulta em "contratação pessoa jurídica prestação serviço".

Após a execução da estratégia citada acima, executando a busca por termos tais como “data de vigência”, “valor”, “ordenador” e “objeto” é extraído a posição que os mesmos aparecem no texto, o que é suficiente para encontrar os termos dentro de um texto, entretanto, não é suficiente para identificar o valor associado.

Ilustração 18 - Identificação de texto identificado pelo algoritmo e o texto alvo



Fonte: Diário Oficial do Estado do Pará - Adaptado

Para a obtenção do texto alvo propriamente dito é necessário armazenar um vetor com a posição e termos na ordem de identificação. Após esse processo basta obter o texto contido entre um termo e seu subsequente e atribuí-lo ao primeiro. A ilustração 19 mostra o algoritmo descrito anteriormente implementado na linguagem Javascript.

Ilustração 19 - Implementação do algoritmo de identificação de termos com a busca *full-text* e seus conteúdos

```
> // content
  const text = `data vigência: 24/08/2020. valor: R$ 500.000,00. Ordendor: Rainaldo da Silva.`;
  // searching terms
  const searching = ["data de vigencia", "ordernador", "valor"];
  // apply fulltext for each search term
  const fulltextMatchs = searching.map((term) => fulltextSearch(text, term))
                                  .sort((a,b) => b.index - a.index ) // sort by identify order
  // output: [{ index: 0, text: "data vigência" },
  //          { index: 32, text: 'valor'},
  //          { index: 54, text: 'Ordendor'}];

  fulltextMatchs.map((match, index, array) => {
    // get substring that start with text found
    const start = text.indexOf(match.text) + match.text.length // match.text.length do closed interval ][
    // get substring that end with next text found
    const end = (index === array.length - 1) ? undefined :  text.indexOf(array[index+1].text)
    // return array that 0 is the term and 1 position is the content
    return [searching[index], text.substring(start, end)];
  });

  // output: [[ "data de vigencia", " 24/08/2020. "],
  //          [ "valor", " R$ 500.000,00. " ],
  //          [ "ordernador", " Rainaldo da Silva."]];
```

Fonte: Os Autores

A finalização da etapa de raspagem é dada após a junção do agrupamento dos órgãos e atos (ilustração 15) com a identificação textual de cada protocolo, fazendo a associação dos

valores textuais com órgão que o publicou, possibilitando a análise de terceiros, dado que a estrutura já retorna o texto original juntamente com os valores identificados catalogados para os seus respectivos órgãos. Na ilustração 20 estão listadas as entidades geradas para o processo de raspagem descrito neste tópico implementado na linguagem Typescript.

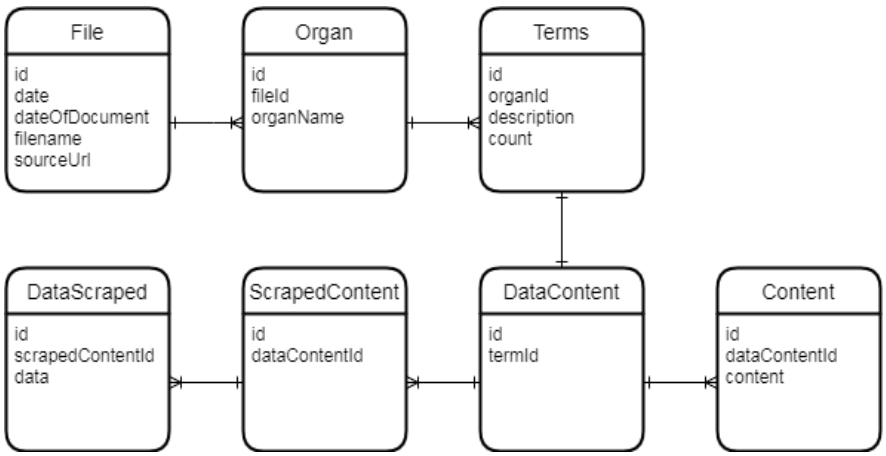
Ilustração 20 - Interfaces geradas no processo de raspagem

```
interface File {
  date: Number
  dateOfDocument: String
  fileName: String
  numberOfDocument: String
  organ: [ Organ ]
  sourceUrl: String
}
interface Organ {
  organName: String // ex: "SECRETARIA DE ESTADO DA FAZENDA"
  terms: [ Terms ]
}
interface Terms {
  description: String // ex: "TERMO ADITIVO A CONTRATO"
  count: Number // ex: 10
  dataContent: DataContent
}
interface DataContent {
  content: [ String ] // ex: "Termo:... Data da Assinatura:02/... Protocolo: 539339"
  scrapedContent: [ ScrapedContent ]
}
interface ScrapedContent {
  dataScraped: [ DataScraped ]
}
interface DataScraped {
  data: [ String ] // ex: ["data assinatura", "02/04/2020"]
}
```

Fonte: Os Autores

O Modelo Entidade Relacionamento (DER)[18], ilustração 21, abstrai e descreve os aspectos dos domínios das entidades mostradas acima.

Ilustração 21 - Diagrama Entidade Relacionamento da estruturação dos dados



Fonte: Os Autores

[18] DER é um tipo de fluxograma que ilustra como “entidades”, p. ex., pessoas, objetos ou conceitos, se relacionam entre si dentro de um sistema criado por Peter Chen na década de 1970. Atualmente, é comumente utilizado durante o projeto de software antes de codificação e desenvolvimento do banco de dados.

Dessa forma, a ilustração 22 demonstra uma parte do resultado final gerado para a estruturação dos dados extraídos do Diário Oficial do Estado do Pará no padrão JSON.

Ilustração 22 - Amostra da estrutura de dados gerada após todo o processo de raspagem


```
▼BANCO DO ESTADO DO PARÁ:
  ▼TERMO ADITIVO A CONTRATO:
    ▶content: ["TERMO ADITIVO Nº: 01 ↵ DATA DE ASSINATURA: 01.04...s da Assunção Sousa da Silva ↵ Protocolo: 540308"]
     count: 1
    ▼scrapedContent: Array(1)
      ▼0:
        ▼dataScrapped: Array(13)
          ▼0:
            ▶data: (2) ["data assinatura", " 01.04.2020"]
          ▼1:
            ▶data: (2) ["termo doacao", " nº 001/2019"]
          ▼2:
            ▶data: (2) ["partes", " banco estado s. a. obras filhas amor jesus cristo - casa menino jesus"]
          ▼3:
            ▶data: (2) ["vigencia", " 01.04.2020 31.03.2021"]
```


Fonte: Os Autores

Atualmente o JSON é padrão mais utilizado no mercado de computação na *web* no quesito de troca de dados entre sistemas (SMITH, 2015). Dessa forma a utilização deste padrão para a estruturação torna fácil a manipulação, transporte e entendimento por qualquer linguagem de programação, SGBD ou plataforma, mantendo a interoperabilidade[19] dos dados.

3.1.5 Persistência

A persistência de dados é a garantia de que um dado foi salvo e que poderá ser recuperado quando necessário no futuro, ou em outras palavras, o conceito de persistência de dados na computação existe para referenciar o ato de salvar os dados (TAKE, 2019).

Algumas técnicas de persistência de dados referem-se a qualquer forma de armazenamento de dados em um sistema, sendo o mais usual a utilização de um banco de dados. A definição de banco de dados defendida por Korth (2004) é que bancos de dados são uma “coleção de dados inter-relacionados, representando informações sobre um domínio específico”. Entretanto, ao se falar sobre banco de dados, geralmente é associado à ideia de um SGBD, todavia, nada impede que sejam utilizadas outras ferramentas, como uma planilha, ou um arquivo de texto, dependendo diretamente do propósito dos dados persistentes e de seu contexto.

[19] Interoperabilidade é a capacidade de um sistema de se comunicar de forma transparente com outro sistema.

Na arquitetura desenvolvida, umas das premissas levadas em consideração é que existirá a necessidade de executar uma densa análise em cima dos dados extraídos do Diário Oficial. Dessa forma, é interessante que os dados estejam armazenados em locais que darão suporte e maior poder de recuperação e organização dos dados, por isso a recomendação de utilizar um banco de dados.

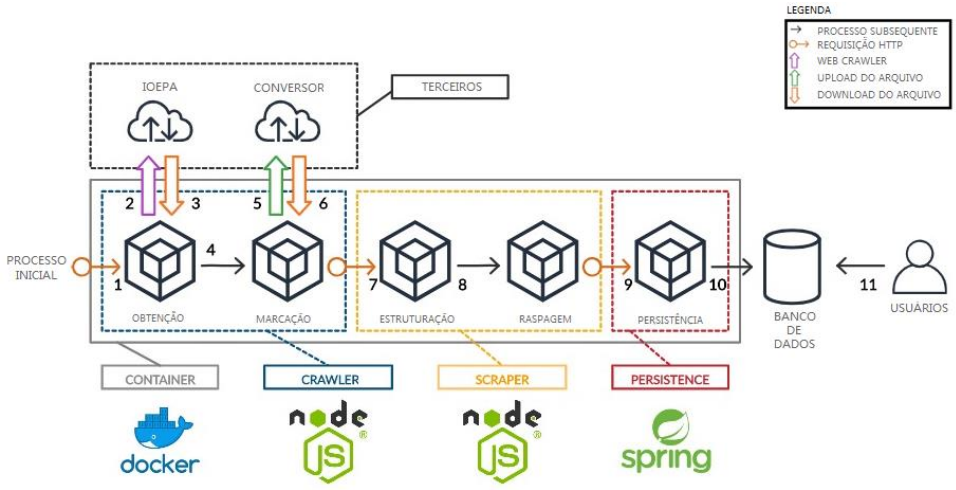
Outra premissa levantada é fator usabilidade, que é definido pela Norma ISO 9241–11 (2011), como sendo um conjunto de atributos que incidem sobre o esforço necessário para o uso do mesmo. Dessa forma, deve-se facilitar a usabilidade dos dados para o maior número de usuários possíveis. Logo, o sistema deve possibilitar ao usuário, a escolha de qual SGBD queira utilizar com a ferramenta e caso não haja preferência utilizar um banco de dados pré-selecionado. Dessa forma, é aumentada a abrangência dos usuários contemplados pela ferramenta, visto que podem customiza-la de acordo com as suas necessidades sem aumentar o custo.

Levando em consideração as duas premissas levantadas anteriormente, a etapa de persistência consiste em armazenar a estrutura de dados extraída do Diário Oficial do Estado do Pará em disco, facilitando a recuperação da informação posteriormente. Neste processo deve-se levar em considerações tanto a facilitação de futuras análises de dados, quanto a usabilidade e diminuição do custo para o acesso dessas informações por terceiros.

3.2 ARQUITETURA DE SOFTWARE

O sistema possui diversas etapas, processos e integrações para que possa ser executado corretamente atingindo os objetivos esperados. A ilustração 23 mostra um esquemático mais completo acerca dos processos do sistema bem como seus relacionamentos.

Ilustração 23 - Esquemático completo dos processos e seus relacionamentos



Fonte: Os Autores

NÚCLEO DA SOLUÇÃO: A arquitetura do *core* (equivalente ao *contêiner* da ilustração 23) planejada de modo que suas funcionalidades estejam separadas em 3 *microservices*, sendo eles: o *Crawler*; o *Scraper* e o *Persistence*.

COMUNICAÇÃO: As comunicações entre os serviços são executadas utilizando o protocolo *Hypertext Transfer Protocol* (HTTP) através de APIs de Transferência Representacional de Estado (REST) para a troca de informações.

AMBIENTE: Cada um dos *microservices* (*Crawler, Scraper* e *Persistence*), foi isolado em *containers Docker*[20]. Dessa forma, o usuário do sistema não precisa configurar ou instalar nenhum *software* para executar todo os serviços além do *Docker*.

INICIALIZAÇÃO: O sistema proposto é acionado através de um *job*[21] que executa uma requisição HTTP para o servidor *core* (demonstrado na ilustração 23) diariamente em um horário predeterminado pelo usuário para que seja possível obter o jornal do dia em questão.

CRAWLER: O *microservice Crawler* é responsável por trabalhar com os recursos da *web* (processos de Obtenção de dados e Marcação da ilustração 5), dessa forma, este sistema tem como objetivo agrupar as regras de negócio relacionada a interface *web* e executar o processo para obter os arquivos do diário oficial (informação não estruturada) e

[20] Docker é um *software* contêiner da empresa Docker, Inc, que fornece uma camada de abstração e automação para virtualização de sistema operacional.
[21] Job, para a computação, executar um processamento remotamente e podem ser iniciados a partir de uma linha de comando ou agendado para execução por um agendador de tarefas.

em seguida utilizar um sistema externo (*third party)* para fazer o processo de marcação inicial. Ao final o sistema é responsável por enviar os dados para o *Scraper*.

SCRAPER: O *microservice Scraper* tem como funcionalidade agrupar toda a regra de negócio referente a raspagem de dados (processos de Estruturação e Raspagem da ilustração 5). Dessa forma o sistema manipula a estrutura inicial (já com marcação) a fim de compilar e agrupar informações textuais, buscando compreender a estrutura na qual o Diário Oficial Do Estado do Pará foi escrito, de maneira que possa facilitar um processamento mais robusto na etapa de raspagem. Após a reestruturação o sistema aplica as técnicas de raspagem para obter os dados agrupados e bem estruturados. Finalmente o sistema envia os dados para o *Persistence.*

PERSISTENCE: O *microservice Persistence* tem como objetivo lidar com o processo de gerenciamento do armazenamento dos dados (processos de Persistência da ilustração 5). Dessa forma o sistema é responsável por identificar as preferências do usuário quanto a utilização de um banco de dados relacional[22], se conectar a ele e persistir os dados recebidos do *Scraper.* O sistema também persiste os dados em disco no formato de texto JSON, facilitando a exportação para um banco de dados não relacional[23]. O modelo proposto implementa um *Object-relational Mapping* (ORM) que abstrai toda a complexidade acerca da utilização do banco de dados por uma linguagem de programação, dessa forma, mantendo dois formatos bases para que sistemas de terceiros possam atuar sobre os dados extraídos.

ACESSO AOS DADOS GERADOS: Ao final o usuário pode acessar as fontes de dados das seguintes maneiras, podendo tanto analisar manualmente através do SGBD como também pode utilizar outras ferramentas para análises, utilizando como base os dados gerados pelo sistema proposto.

3.3 TECNOLOGIAS DA IMPLEMENTAÇÃO

Para o desenvolvimento do sistema foi adotada a linguagem de programação Javascript utilizando o *runtime* Nodejs[24] para o desenvolvimento dos *microservices Crawler* e do *Scraper*, visto que as funcionalidades dos mesmos são acessar tecnologias *web* como

[22] Banco de dados relacionais utilizam a linguagem Structured Query Language (SQL) para executar pesquisas declarativas inspirada na álgebra relacional.

[23] Banco de dados não relacionais são baseados em formas diferentes das relações tabulares dos bancos de dados relacionais. Atualmente, o termo NoSQL (*Not Only* SQL) é utilizado para representar essa classe de banco de dados.

[24] Nodejs é um interpretador de JavaScript assíncrono com código aberto orientado a eventos, criado por Ryan Dahl em 2009, focado em migrar a programação do Javascript do cliente para os servidores.

browser ou manipular as estruturas HTML, dessa forma, o Javascript é uma linguagem naturalmente *web* e lida mais facilmente com estas estruturas. Além disso utilizou-se o *framework Express*[25] para a criação dos servidores em Nodejs, facilitando e agilizando o desenvolvimento.

No desenvolvimento do microservice *Persistence* levou-se em consideração que as principais funcionalidades do mesmo são a persistência de dados e a facilidade em manipular os dados, dessa forma, optou-se pela utilização da linguagem Java utilizando o framework Spring Boot[26] para o desenvolvimento do servidor e do *framework Hibernate*[27] como ORM para manipulação do acesso a diversos tipos de banco de dados.

Outra tecnologia utilizada para a centralização dos aplicativos desenvolvidos e abstração da complexidade de instalação dos sistemas foi o *Docker*, fornecendo uma camada de abstração e automação. Dessa forma, para instalar o *software*, basta que o usuário tenha o cliente *Docker* para que a própria ferramenta se encarregue de configurar o ambiente para inicializar o *software*.

[25] O Express.js, ou Express, é uma estrutura de aplicativo da *Web* para o Nodejs, projetado para criar aplicativos *web* e APIs.

[26] O Spring é um framework Open Source para a plataforma Java criado por Rod Johnson e descrito em seu livro "Expert One-on-One: JEE Design e Development". Trata-se de um framework não intrusivo, baseado nos padrões de projeto inversão de controle (IoC) e injeção de dependência.

[27] O Hibernate é um *framework* para o mapeamento objeto-relacional escrito na linguagem Java.

Capítulo 4

Análise de Resultados

Este capítulo apresenta os resultados obtidos a partir do desenvolvimento da ferramenta proposta, bem como também destaca os principais processos utilizados na formulação da mesma.

4 ANALISE DOS RESULTADOS

O modelo desenvolvido possui uma arquitetura complexa e robusta para atender os requisitos levantados na situação problema descritos no tópico 1.1. A arquitetura do modelo (ilustração 4) é capaz de suprir as necessidades de extração de informação em fontes de dados não estruturada utilizando os seguintes processos para a estruturação: obtenção de dados; marcação; estruturação; raspagem e persistência de dados.

Os processos desenvolvidos pela ferramenta são efetuados de maneira automática, sendo possível a execução programada diariamente através de um *job* de um agendador de tarefas por exemplo. Ao final a ferramenta converte os dados do Diário Oficial do Estado do Pará em informação estruturada, possibilitando a manipulação de dados por linguagens SQL, disponibilizadas através de uma interface (SGBD) para a análise de um usuário final. A partir desse ponto, é viabilizado através da ferramenta, a possibilidade de obter *insights* acerca de dados não só de um único arquivo, mas também, de todos os registros anteriormente armazenados no banco de dados, proporcionando facilidade ao acesso dos dados estruturados, o que pode contribuir para uma fiscalização mais célere e eficiente.

A possibilidade de extrair mais informações, ou até mesmo mais conhecimento oriundos de dados dos órgãos públicos também viabiliza o controle social através da disponibilização dos mesmos, em outras palavras, a disponibilização de informação relevante e concisas dos atos públicos também sustenta uma sociedade democrática e possibilita uma atuação mais presente por parte da população.

Capítulo 5

Conclusões

Este capítulo apresenta as considerações finais sobre este trabalho, abordando um resumo do trabalho realizado, as principais contribuições e a indicação de pesquisas futuras.

5 CONSIDERAÇÕES FINAIS

Inicialmente, foi constatado que o Diário Oficial do Estado Pará é produzido para o consumo humano e que a divulgação neste formato não facilita a manipulação das informações por sistemas computacionais, dificultando a execução de análises automatizadas por parte do mesmo. Dito isso, este cenário torna-se suscetível a fraudes ou qualquer outro ato ilícito prejudicial à administração pública. A fim de solucionar esse problema, identificou-se a necessidade de se obter dados estruturados a partir do Diário Oficial do Estado, com o intuito de apoiar a fiscalização e facilitar o compartilhamento e utilização das informações por outros sistemas.

Dessa forma, foram estabelecidos como objetivos específicos os processos necessários para sanar tal problema. Foram levantados como requisitos os objetivos específicos descritos no tópico 1.2.1. O primeiro objetivo específico tem como foco identificação dos tipos de dados disponibilizados pelo Diário Oficial do Estado do Pará, onde foi observado que é disponibilizado um arquivo para download no formato de arquivo PDF. Já no segundo objetivo específico, foram estudados os meios de obter os arquivos do portal da Imprensa Oficial do Estado do Pará. No terceiro objetivo específico, foi desenvolvido um *Web Crawler* para acessar o portal, simulando o acesso por usuário humano e realizando o download do arquivo PDF. O quarto objetivo específico teve como essência o pré-processamento dos arquivos coletados, a fim de tokenizar[28] os dados para facilitar a manipulação em processos subsequentes e viabilizar a utilização da estilização como parte da análise. O quinto objetivo se baseou na estruturação dos dados pré-processados, tendo em vista a disposição lógica dos mesmos, de tal modo que a reorganização mantenha um fluxo textual conciso, unificando colunas e páginas. Para o sexto objetivo específico, fez-se a extração das informações dos dados estruturados, onde este processo consistiu em encontrar padrões na arquitetura da informação e aplicar algoritmos para extraí-los. Por último, o sétimo objetivo específico baseou-se no projeto de um modelo capaz de persistir as informações de forma a permitir análises futuras dos dados.

Ao final desses processos foi gerado um mecanismo de software para a obtenção de dados não estruturados, extração de informação e disponibilização dos dados do Diário Oficial, contemplando o objetivo geral estipulado.

[28] Processo de inserção de *tags* ou *tokens* para delimitar sessões em um trecho de texto.

Com isso, deduziu-se que a hipótese disposta nesta proposta de pesquisa partiu da criação de um *software* capaz de estruturar esses dados para então analisá-los sistematicamente com o intuito de obter informações relevantes para a fiscalização de recursos públicos. Durante o trabalho aplicou-se técnicas para elaboração dos objetivos propostos e então foi realizada a implementação e validação da hipótese, e ao final de todas as etapas, a mesma foi confirmada devido a disponibilização dos dados de forma estruturada, permitindo que um analista realize análises sistemáticas auxiliadas por ferramentas computacionais. Logo, conclui-se que, o sistema proposto contribuí para uma maior agilidade na análise e fiscalização das informações do diário oficial do estado do Pará, disponibilizando novos meios para que os analistas atuem e possibilita integração com outros sistemas.

No entanto, foram identificadas as seguintes limitações e dificuldade durante o desenvolvimento deste trabalho: a) o sistema necessita que as informações extraídas sejam previamente mapeadas; b) o conteúdo textual é inserido manualmente (fator humano) e posteriormente é salvo em um formato que dificulta a manipulação dos dados; c) o sistema depende de uma ferramenta externa para fazer a conversão do arquivo, neste caso foi utilizado o *site* http://www.pdftohtml.com devido ao custo de desenvolvimento de uma ferramenta de conversão não ser factível em um tempo hábil. Outras bibliotecas nas linguagens Python, e Javascript foram estudadas, entretanto nenhuma ofereceu uma qualidade minimamente viável; d) dificuldade no Processamento de Linguagem Natural devido às variações textual, erros ortográficos e interpretação; e) o sistema desenvolvido não reconhece tabelas dispostas no Diário Oficial do Estado do Pará.

Futuramente, como melhoria a longo prazo pode-se acoplar mais funcionalidades dentro do *Scraper* para aprimorar o reconhecimento dos protocolos dos atos públicos através de algoritmos de processamento de linguagem natural combinados com redes neurais artificiais, algoritmos genéticos, árvores de decisão, mapas autoconfiguráveis, regressão linear ou Naiver Bayer, por exemplo. A curto prazo, a inclusão de novas coleções de dados na etapa de raspagem, descrita no tópico 3.1.4, já viabiliza encontrar novos padrões e consequentemente aprimorar a extração de informação do Diário Oficial do Estado do Pará.

Referências Bibliográficas

ALECRIM, E. **Arquivos PDF**. 23 setembro 2007. Disponível em: https://www.infowester.com/arquivospdf.php. Acesso em: 18 maio 2020.

ARTHUR, Charles. **What's a zettabyte? By 2015, the internet will know, says Cisco.** [S. l.], 29 jun. 2011. Disponível em: https://www.theguardian.com/technology/blog/2011/jun/29/zettabyte-data-internet-cisco. Acesso em: 2 jan. 2020.

ASKITAS, Nikolaos; ZIMMERMANN, Klaus F. The internet as a data source for advancement in social sciences. **International Journal of Manpower**, 2015.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 9241-11: Requisitos ergonômicos para o trabalho com dispositivos de interação visual Parte 11: Orientações sobre usabilidade**. Rio de Janeiro. 2011.

BEAL, Vangie; **SDK - software development** kit. [S. l.], 6 maio 2016. Disponível em: https://techterms.com/definition/sdk. Acesso em: 19 maio 2020.

BRASIL. Constituição da República Federativa do Brasil de 1988. **Preâmbulo**, Brasília, DF, maio 2020.

BRASIL. Lei Complementar Nº 131, de 27 de maio de 2009. **Normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências**, Brasília, DF, maio 2020.

CABENA, P; HADJINIAN, P; STADLER, R; JAAPVERHEES; ZANASI, A. **Discovering Data Mining: From Concept to Implementation**. Prentice Hall, 1998.

CAMILO, Cássio Oliveira et al. **Um estudo sobre a interação entre Mineração de Dados e Ontologias**. [S. l.], 2009. Disponível em:

http://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_002-09.pdf. Acesso em: 20 jul. 2020.

CANALTECH**. O QUE é Big Data?**. [S. l.], 10 jan. 2015. Disponível em: https://canaltech.com.br/big-data/o-que-e-big-data/. Acesso em: 2 jan. 2020.

CARVALHO, Rommel & PAIVA, Eduardo & ROCHA, Henrique & MENDES, Gilson. **Methodology for Creating the Brazilian Government Reference Price Database**. 2013. Disponível em: http://www.lbd.dcc.ufmg.br/colecoes/eniac/2013/0033.pdf.

CARVALHO, Rommel & PAIVA, Eduardo & ROCHA, Henrique & MENDES, Gilson. **Using Clustering and Text Mining to Create a Reference Price Database. Learning and NonLinear Models**. 2014; 12:38–52.

CASA CIVIL. **Lei nº 101, de 4 de maio de 2000**. LEI COMPLEMENTAR. [*S. l.*], 4 maio 2000. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp101.htm. Acesso em: 28 maio 2020.

CONCEIÇÃO, ANTONIO CESAR LIMA DA. **Controle Social da Administração Pública: Informação & Conhecimento – interação necessária para a efetiva participação popular nos orçamentos públicos.** Orientador: Profª. Drª. Rita de Cássia Leal Fonseca dos Santos. 2010. 36 f. Trabalho de Conclusão de Curso (Especialização em Orçamentos Públicos) - Instituto Serzedello Corrêa – ISC/DF, [S. l.], 2010. Disponível em: https://portal.tcu.gov.br/lumis/portal/file/fileDownload.jsp?fileId=8A8182A24F0A728E014F0ADED2B42F79. Acesso em: 29 maio 2020.

CONSELHO NACIONAL DE JUSTIÇA (Brasil). Poder Judiciário do Brasil. **Entenda os conceitos de improbidade administrativa, crimes contra a administração pública e corrupção**. [S. l.], 9 fev. 2015. Disponível em: https://www.cnj.jus.br/entenda-os-conceitos-de-improbidade-administrativa-crimes-contra-a-administracao-publica-e-corrupcao/. Acesso em: 24 maio 2020.

CONTROLADORIA GERAL DA UNIÃO. **O vereador e a fiscalização dos recursos públicos municipais**. Presidência da República, Controladoria Geral da União, Brasília,

CGU, Disponível em: http://www.contag.org.br/imagens/fa-fiscalizacao-dos-recursos-publicos---cartilha-do-vereador.pdf, 2009.

DAVENPORT, T. H.; PRUSAK, L. **Ecologia da informação: por que só a tecnologia não basta para o sucesso na era da informação**. São Paulo: Futura, 1998a. 316p.

DAVENPORT, Thomas H.; PRUSAK, Laurence. **Conhecimento empresarial: como as organizações gerenciam o seu capital intelectual**. 4. ed. Tradução de Lenke Peres. Rio de Janeiro: Campus, 1998. 237 p.

DEAN, Jeffrey; GHEMAWAT, Sanjay. **MapReduce: Simplified Data Processing on Large Clusters**. OSDI '04: 6th Symposium on Operating Systems Design and Implementation, [S. l.], p. 1-7, 15 jan. 2004. DOI 10.1145. Disponível em: https://static.usenix.org/publications/library/proceedings/osdi04/tech/full_papers/dean/dean.pdf. Acesso em: 29 maio 2020.

DROPA, Romualdo Flávio. **Transparência e fiscalização na administração pública.** [S. l.], 30 maio 2004. Disponível em: https://ambitojuridico.com.br/cadernos/direito-administrativo/transparencia-e-fiscalizacao-na-administracao-publica/. Acesso em: 6 maio 2020.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

FERREIRA, Tonico. **As várias faces da corrupção têm um custo alto que é pago pela sociedade: Desfalques nos cofres da União, dos estados e dos municípios reduzem a capacidade dos governos de prestar serviços essenciais**. Jornal Nacional, [S. l.], p. 1, 13 abr. 2017.

FIORO, V. **O QUE é Big Data e para que serve?**. [S. l.], 16 nov. 2014. Disponível em: https://industriahoje.com.br/o-que-e-big-data-e-para-que-serve. Acesso em: 2 jan. 2020.

FRAGA et. al., 2011; **Uma API Java para Acesso a Banco de Dados Relacional usando o OGSA-DAÍ**, Universidade Estadual do Mato Grosso do Sul, monografia de graduação, 2011.

FREITAS, Marcos. WEB Scraping com Nodejs. [S. l.], 9 maio 2020. Disponível em: https://medium.com/@marcosfreitas/web-scraping-com-nodejs-a4b53946c76. Acesso em: 4 jul. 2020.

GRANATYR, Jones. **IA Forte x IA Fraca**. [S. l.], 17 jan. 2017. Disponível em: https://iaexpert.academy/2017/01/17/ia-forte-x-ia-fraca/. Acesso em: 20 jul. 2020.

GLOBALAD. **O que é Crawler?**. [S. l.], 2017. Disponível em: https://globalad.com.br/blog/o-que-e-crawler/. Acesso em: 4 jul. 2020.

INTERNATIONAL ORGANIZATION OF NORMALIZAÇÃO. **ISO/IEC 21778:2017 Information technology — The JSON data interchange syntax**, novembro 2017.

HAND, D; MANNILA, H; SMYTH, P. **Principles of Data Mining**. MIT Press, 2001.

KITCHIN, R. **The data revolution: big data, open data, data infrastructures and their consequences.** [S.l.]: SAGE, 2014.

KORTH, H.F. e SILBERSCHATZ, A.; **Sistemas de Bancos de Dados**, Makron Books, 2a. edição revisada, 1994.

LEMLEY, Linda. **Discovering Computers: "Chapter 6: Output"**. University of West Florida. 3 Jun 2012. Disponível em: https://web.archive.org/web/20120614152622/http://uwf.edu/clemley/cgs1570w/notes/Concepts-6.htm. Acesso em: 04 de fev. 2019.

MANYIKA J, Chui M, Bughin J, Dobbs R, Bisson P, Marrs A. McKinsey **Global Institute D**. 2013.

MARTINS, Sergio de Castro. et al. **Gestão da informação: Estudo comparativo de modelos sob a ótica integrativa dos recursos de informação**. Dissertação (Dissertação em Ciência da Informação) – Niterói, p. 14. 2014.

MEDEIROS, Luciano Frontino de. **Inteligência artificial aplicada**: Uma abordagem introdutória. 1. ed. [S. l.]: InterSaberes, 2018. 265 p. ISBN 8559728007.

MELLO, Ronaldo dos Santos; DORNELES, Carina Friedrich; KADE, Adrovane; BRAGANHOLO, Vanessa de Paula; HEUSER, Carlos Alberto**. Dados Semi-Estruturados**. Disponível em: https://www.ime.usp.br/~jef/semi-estruturado.pdf. Acesso em: 04 de fev. 2019.

MITCHELL, Ryan. **Web Scraping com Python: coleta de dados na web moderna**. [S. l.]: Novatec, 2015. 288 p. v. 1. ISBN 9788575224472.

MITRA, Akash; **Classifying data for successful modeling**, 2011.

MONTEIRO, Leandro Pinho. **Dados Estruturados e Não Estruturados**. In: Https://universidadedatecnologia.com.br/dados-estruturados-e-nao-estruturados/. [S. l.], 4 fev. 2019. Disponível em: https://universidadedatecnologia.com.br/dados-estruturados-e-nao-estruturados/. Acesso em: 3 jul. 2020.

MOREIRA, Daniel Alexandre et al. **Teoria e prática em gestão do conhecimento: pesquisa exploratória sobre consultoria em gestão do conhecimento no Brasil**. 2005. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal de Minas Gerais, [S. l.], 2005. Disponível em: https://repositorio.ufmg.br/bitstream/1843/VALA-6K5NGG/1/mestrado___daniel_alexandre_moreira.pdf. Acesso em: 28 maio 2020.

MULLER, Martin U.; ROSENBACH, Marcel; SCHULZ, Thomas. **Big Data Knows What Your Future Holds**. [S. l.], 7 mar. 2013. Disponível em: https://www.spiegel.de/international/business/big-data-enables-companies-and-researchers-to-look-into-the-future-a-899964.html. Acesso em: 2 jan. 2020.

NASCIMENTO, Mariana. **A importância da participação popular no processo de gestão pública**. [S. l.], 21 set. 2018. Disponível em: https://www.e-gestaopublica.com.br/a-importancia-da-participacao-popular-no-processo-de-gestao-publica/. Acesso em: 21 set. 2018.

OECD. **Glossary of Statistical Terms**. [S. l.: s. n.], 2008. 605 p. ISBN 978-92-64-025561. Disponível em: https://www.oecd-ilibrary.org/docserver/9789264055087-en.pdf?expires=1590649476&id=id&accname=guest&checksum=D81B093CAAA28656EF4B2207E894249E. Acesso em: 28 maio 2020.

OLIVEIRA, ALCIONE DE P.; MACIEL, VINÍCIUS V. **JAVA NA PRÁTICA**. [S. l.: s. n.], 2002. v. 1. Disponível em: http://www.dpi.ufv.br/~vladimir/java/Java%20na%20Pratica_vol1.pdf. Acesso em: 28 maio 2020.

PADEIRO, Maione. **Os males da corrupção na economia**. [S. l.], 29 maio 2017. Disponível em: https://www.dm.jor.br/opiniao/2017/05/os-males-da-corrupcao-na-economia/. Acesso em: 23 maio 2020.

PAIVA, Eduardo de; REVOREDO Kate. **Big Data e Transparência: Utilizando Funções de Mapreduce para incrementar a transparência dos Gastos Públicos,** XII Brazilian Symposium on Information Systems, Florianópolis, SC, May 17-20, 2016

PATIL, Yugandhara; PATIL, Sonal. **Review of Web Crawlers with Specification and Working**. International Journal of Advanced Research in Computer and Communication Engineering. 1 Jan. 2016. Disponível em: https://www.ijarcce.com/upload/2016/january-16/IJARCCE%2052.pdf. Acesso em: 04 de fev. 2019.

PROVOT, Foster; FAWCETT, Tom. **Data Science para Negócios**. [S. l.]: Alta Books, 2016. ISBN 9788576089728.

REZENDE, Solange Oliveira; **Sistemas Inteligentes**, Manole Barueri São Paulo, 2003.

RUSSELL, Stuart; NORVIG, Peter (2003). **Artificial Intelligence**. A Modern Approach. 2ª ed. Upper Saddle River, New Jersey: Prentice Hall. p. 1-2. 1081 p. ISBN 0137903952

SANTOS, A.R. (org.) **Gestão do conhecimento: uma experiência para o sucesso empresarial**. Curitiba: Champagnat, 2001.

SCHOENHERR, Steven E. **The Digital Revolution.** 10. mai. 2004. Disponível em: https://web.archive.org/web/20070103161029/http://history.sandiego.edu/gen/recording/digitalrev.html. Acesso em: 3 dez. 2019.

SMITH, BEN. **JSON Básico: Conheça o formato de dados preferido da web**. [S. l.: s. n.], 2015. ISBN 8575224360.

SOUTO, Mario. **O que é front-end e back-end?**. Alura, 25 set. 2019. Disponível em: https://www.alura.com.br/artigos/o-que-e-front-end-e-back-end. Acesso em: 20 jul. 2020.

SUSANNE. **What Is Scraping? The Basics For Everyone**. [S. l.], 7 maio 2015. Disponível em: http://myhelpster.com/what-is-scraping-the-basics-for-everyone/. Acesso em: 2 jan. 2020.

TAKE. **Persistência de dados: tudo que você precisa saber sobre conceito, tipos e técnicas**. [S. l.], 16 ago. 2019. Disponível em: https://take.net/blog/devs/persistencia-de-dados#:~:text=Ou%20seja%2C%20podemos%20entender%20o,computador%20e%20o%20c%C3%A9rebro%20humano. Acesso em: 4 jul. 2020.

TAURION C. **Big data. Brasport**; 2013.

TECHOPEDIA. **Definition - What does Web Scraping mean?**. [S. l.], 4 jul. 2020. Disponível em: https://www.techopedia.com/definition/5212/web-scraping. Acesso em: 4 jul. 2020.

WORLD WIDE WEB CONSORTIUM. **Extensible Markup Language (XML)**. 5. ed. [S. l.], 26 nov. 2008. Disponível em: https://www.w3.org/TR/REC-xml/. Acesso em: 20 jul. 2020.