Submission date: 12/04/2023
Camera ready submission date: 07/11/2023

1st round notification: 28/10/2023
Available online: 22/12/2023
Publication date: 22/12/2023

Section: regular article

Aplicando Técnicas de Recuperação da Informação para Manipulação de Dados Abertos de Compras Governamentais na Área da Saúde em Municípios Brasileiros

Applying Information Retrieval Techniques to Handle Open Government Procurement Data in the Health Area in Brazilian Municipalities

Matheus A. F. J. Araújo[1], **Raimundo C. S. Vasconcelos**[1]

[1] Instituto Federal de Brasília (IFB) - *campus* Taguatinga
Brasília - DF

{matheusjucazevedo@gmail.com, raimundo.vasconcelos@ifb.edu.br

Abstract. *Technological advances have made it possible for companies and state governments to store huge amounts of data. At the same time, public organizations are under pressure to provide political and economic transparency and therefore open repositories of government data are created. Open data, by law, must be easily accessible, free and complete, which makes them the perfect tool to monitor public spending. The objective of this work is to apply information retrieval techniques to evaluate the availability of open data on government purchases from Brazilian municipalities related to public health. Solutions are also discussed that allow the use of open data for public spending analysis, ensuring the empowerment of Brazilian society through knowledge.*

Keywords. *Data Warehouse, Open data, Public health*

Resumo. *O avanço tecnológico possibilitou empresas e Estados armazenarem grandes quantidades de dados. Ao mesmo tempo, o Estado sofre pressão para apresentar transparência política e econômica e, assim, criam-se repositórios abertos de dados governamentais. Dados abertos, por lei, devem ser de fácil acesso, gratuitos e completos, muito úteis para auditar gastos públicos. O objetivo deste trabalho é aplicar técnicas de recuperação da informação para avaliar a disponibilidade de dados abertos de compras governamentais de municípios brasileiros relacionados à saúde pública. Também são discutidas soluções que permitam o uso de dados abertos para análise de gastos públicos, apresentando um corpo de conhecimento inicial para a sociedade brasileira.*

Palavras-Chave. *Armazém de Dados, Dados Abertos, Saúde Pública*

Cite as: Araujo, M. A. F. J., & Vasconcelos, R. (2023). Applying Information Retrieval Techniques to Handle Open Government Procurement Data in the Health Area in Brazilian Municipalities (Aplicando Técnicas de Recuperação da Informação para Manipulação de Dados Abertos de Compras Governamentais na Área da Saúde em Municípios Brasileiros). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 16(1), 12:1-12:29. DOI: 10.5753/isys.2023.3089

1. Introdução

Diversas organizações internacionais apontam a grave situação da corrupção no Brasil. A Transparência Internacional, em relatório recente [Internacional 2021], afirmou que o Brasil enfrenta "sérios retrocessos no combate à corrupção". Entre 2007 e 2020, o país perdeu 22 posições no ranking da organização que mensura a percepção de corrupção nas diferentes nações, com base na opinião de representantes do setor privado e analistas. Segundo o WEF (*World Economic Forum*) [Forum 2019], o Brasil está na posição 91, de 141 países avaliados, com relação à corrupção, tendo piorado nos últimos anos. Aliado a isso, WEF aponta o Brasil na posição 111 com relação ao uso de tecnologia na gestão governamental.

Em uma entrevista com auditores fiscais [Pontes 2019], para o jornal alemão *Deutsche Welle*[1], foi relatado que diversas decisões do STF limitaram a ação da Receita Federal e COAF. Isto é especialmente perigoso devido ao fato do Brasil ser membro do Gafi/FATF (Grupo de Ação Financeira contra Lavagem de Dinheiro e Financiamento ao Terrorismo) e a falta de comprometimento do país com as ações tomadas pelo grupo pode levar a sanções econômicas e políticas afetando toda a população. Isso pode ser um dos motivadores para os baixos índices na saúde, educação e segurança do país, apesar de ser a 8ª economia mundial nos últimos anos segundo o FMI [IBGE 2017].

A percepção pela população da má qualidade dos serviços públicos prestados - saúde, educação, segurança - tem se verificado. Segundo pesquisa da Empresa Brasil de Comunicação, quase 90% dos entrevistados considera a saúde brasileira inferior ou igual a regular [Laboissière 2018]. Destes entrevistados, mais de 80% acredita que os recursos são mal administrados e, para 26% dos entrevistados, devem ser tomadas medidas para lutar contra a corrupção dentro do ambiente de saúde pública garantindo um serviço de qualidade, igualitário e gratuito. É importante ressaltar que o Brasil tem tomado diversas ações em relação ao combate à corrupção e em prol do aumento da transparência e uma dessas atitudes é a disponibilização de Dados Abertos (DA) de gastos do governo [[de Cultura Digital et al. 2011]]. O TCU publicou uma revista que informa os cinco motivos para a abertura de dados na administração pública, que são [Tribunal de Contas da União 2015]:

1. porque a sociedade exige mais transparência na gestão pública;
2. porque a própria sociedade pode contribuir com serviços inovadores ao cidadão;
3. porque ajuda a aprimorar a qualidade dos dados governamentais;
4. para viabilizar novos negócios;
5. porque é obrigatório por Lei.

Em 18 de novembro de 2011, foi sancionada a Lei de Acesso a Informação Pública (Lei 12.527/2011), que regula o acesso a dados e informações detidas pelo governo, tornando obrigação do Estado a disponibilização de Dados Abertos (DA) na *internet* de forma que todos possam acessar. O órgão responsável pela gestão e monitoramento da política é a CGU, por meio da Infraestrutura Nacional de Dados Abertos

[1] http://dw.com

[de Dados Abertos 2019]. Regular informação que afeta mais dos 200 milhões de brasileiros e de 570 municípios diferentes é algo difícil e muito pouco explorado, conforme citado no Manual de dados abertos:

> Dados abertos, especialmente os governamentais, são um ótimo recurso ainda muito pouco explorado. Muitos indivíduos e organizações coletam uma ampla gama de diferentes tipos de dados para executar suas tarefas. O governo é particularmente importante nesse contexto, tanto por causa da quantidade e da centralidade dos dados que coleta quanto pelo fato de que tais dados são públicos, um direito garantido no artigo 5º da Constituição Federal Brasileira [**?**, manual-dados-abertos, pg.7]

Portanto, torna-se essencial a criação de novas funcionalidades e *softwares* capazes de lidar com grande quantidade de informação e de trazer conhecimento para o povo brasileiro de forma que seja fácil ler e entender, permitindo aumentar a transparência do uso do dinheiro público e democratizar o conhecimento. A partir do avanço tecnológico, redução dos custos do *hardware* e disseminação do acesso a *internet* torna-se imprescindível, em vista do crescente volume de usuários, a criação de ferramentas e técnicas para lidar com uma grande quantidade de dados que pode ser quantificada, tratada, interpretada e comercializada.

Dentre tais tecnologias e ferramentas, destaca-se a Mineração de Dados, que pode ser definida como um processo analítico para uma ampla quantidade de dados de forma que seja possível gerar conhecimento. Um dos fatores de sucesso é o fato de dezenas e centenas de milhões de reais serem gastos pelas companhias na coleta dos dados no decorrer de anos. Esses dados podem ser usados por algum tipo de processo para descoberta de conhecimento. Isso é especialmente verdade quando se tem diversos dados que compõem um mesmo assunto, mas que não são dispostos de maneira similar [Camilo and da Silva 2009]. Tais dados devem ser preparados, limpos e validados antes do processo de análise acontecer através da tecnologia de armazém de dados, chamado de *Data Warehouse*, que pode ser entendido, segundo Paim ([Paim 2003] apud [Inmon 1996], p.43), como: "É uma coleção de dados orientada a assunto, integrada, não-volátil e variante no tempo em suporte a decisões gerenciais". O conceito de um armazém de dados surgiu pela necessidade de oferecer uma origem de dados única e limpa e consistente o suficiente para apoiar tomadas de decisão, servindo como um sistema de apoio crucial para transformação de dados em conhecimento.

Já existem casos de sucesso no Brasil que partem dessa mesma premissa, como a Operação Serenata de Amor [2], que tem como objetivo auxiliar na auditoria de pedidos de reembolsos de deputados, disponibilizando os dados na plataforma chamada Jarbas[3] de forma entendível e sinalizando pedidos de reembolso suspeitos.

Assim este trabalho tenta responder os seguintes questionamentos: Existe a disponibilidade dos Dados Abertos dos municípios brasileiros relacionados a saúde? De que

[2] `http://serenata.ai/`

[3] `http://jarbas.serenata.ai/dashboard/chamber_of_deputies/reimbursement/`

forma essa disponibilização pode ser melhorada?

A premissa desse trabalho é a avaliação da disponibilização de dados de compras de materiais relacionados a saúde de municípios brasileiros para auxílio na auditoria de gastos públicos, de forma que sejam analisadas categorias de produtos, tempo e o preço dos itens, fornecendo soluções para melhorar a forma como essas informações são dispostas.

Considerando isso, os passos seguintes consistem em identificar os locais de armazenamento e buscar as informações necessárias para análise e interpretação dos gastos municipais na área da saúde. Uma vez identificados, necessita-se checar se as informações estão organizadas de forma a garantir seu consumo.

Neste trabalho foram realizadas buscas em *sites* governamentais de municípios brasileiros que possuíssem informações relativas à aquisição de materiais hospitalares, através de uma pesquisa exploratória.

Este artigo está organizado da seguinte forma: na próxima seção são levantados os conceitos básicos de Dados Abertos, Compras Governamentais e Pesquisa da Informação relacionados com a proposta; os trabalhos relacionados são apresentados na seção 3; a seção 4 detalha o desenvolvimento do trabalho e a seção 5 descreve as conclusões e sugestões de trabalhos futuros.

2. Conceitos Básicos

Nesta seção serão abordados os conceitos de Dados Abertos, tipos de compras governamentais, Mineração de Dados, Recuperação da Informação, fundamentais para a compreensão desse trabalho.

2.1. Dados Abertos

Na atualidade, a *internet* tornou-se uma importante fonte de informações para a maioria dos brasileiros e, portanto, como um meio de acesso comum, tende a se tornar o palco para unificar informações pertinentes à sociedade. Dessa forma, órgãos e instituições governamentais, de todos os níveis têm tornado públicas suas ações através da *internet*. Em 18 de novembro de 2011 foi sancionada a Lei de Acesso a Informação Pública (Lei 12.527/2011), que regula o acesso a dados e informações detidas pelo governo. Essa lei constitui um marco para a democratização da informação pública e, preconiza, dentre outros requisitos técnicos, que a informação solicitada pelo cidadão deve seguir critérios tecnológicos alinhados com as três leis de dados abertos [Infraestrutura Nacional de Dados Abertos 2019].

As chamadas três leis de dados abertos são uma convenção feita por [Eaves 2009], ativista dos dados abertos e especialista em políticas públicas, descritas a seguir:

1. Se não pode ser indexada ou rastreada, ela não existe;
2. Se não está disponível em um formato aberto e legível de máquina, ela não pode ser interagida;
3. Se uma estrutura legal não permite ela ser republicada, então ela não empodera;

O manual dos dados abertos (DA), evidenciando que o conceito de DA pode possuir diversas formas de interpretação e consolidação, além de possuir essa concepção mais simples, possui ainda mais oito pressupostos para ampliar tal conceito, que são [[de Cultura Digital et al. 2011]]:

- Dados precisam ser completos: todos os dados públicos estão disponíveis. Dado público é o dado que não está sujeito a limitações válidas de privacidade, segurança ou controle de acesso;
- Dados precisam ser primários: todos os dados devem ser apresentados tal como colhidos da fonte, sem modificação ou granularidade;
- Dados precisam ser atuais: os dados são disponibilizados tão rapidamente quanto necessário à preservação do seu valor;
- Dados precisam ser acessíveis: os dados são disponibilizados para o maior alcance possível de usuários e para o maior conjunto possível de finalidades;
- Dados precisam estar em formato compreensível por máquinas: os dados são razoavelmente estruturados de modo a possibilitar processamento automatizado;
- Dados precisam ser não discriminatórios: os dados são disponíveis para todos, sem exigência de requerimento ou cadastro;
- Dados precisam estar disponíveis em formato não proprietário: os dados são disponíveis em formato sobre o qual nenhuma entidade detenha controle exclusivo;
- Dados precisam estar livres de licenciamento: os dados não estão sujeitos a nenhuma restrição de direito autoral, patente, propriedade intelectual ou segredo industrial. Restrições sensatas relacionadas à privacidade, segurança e privilégios de acesso são permitidas.

Apesar da consolidação da Política de DA ter sido estabelecida pelo Decreto n.º 8.777, de 2016, ainda não existe uma padronização para ditar como os dados devem estar dispostos [Infraestrutura Nacional de Dados Abertos 2019]. [Diniz 2013], descreve diversos aspectos que devem ser considerados na hora de disponibilizar os DA ao público como: ser independente de plataformas tecnológicas, basear-se em formatos padronizados, estarem dispostos de forma estruturada, possuir metadados auto-descritivos para facilitar a indexação por mecanismos de busca como o *Google* e facilitar o entendimento do negócio por parte do usuário, separar os dados da interface e utilizar um padrão de URI para cada conjunto de dados.

Tal esforço, para alcançar um estado mais maduro da disponibilidade de dados, tem seus benefícios como facilidade de integração com outros serviços e o incentivo para o desenvolvimento de sistemas utilizando as bases de dados governamentais. Já existem casos de sucesso em outros países em que esse trabalho de maturação já foi feito, como no *US Hospital Finder*[4], que permite encontrar os hospitais mais próximos, especialidades, localização e horário de funcionamento; outro caso de sucesso é o *Where Does My Money Go ?*[5] que permite, de forma mais simplificada, que cidadãos britânicos averiguarem como o governo gasta o dinheiro, e dessa forma, democratizam o conhecimento público. Neste trabalho, o objetivo é avaliar a disponibilidade dos dados de compras governamentais da área da saúde e, a seguir, são descritos os tipos de compra utilizados.

[4] `http://www.ushospitalfinder.com/`

[5] `https://app.wheredoesmymoneygo.org/`

2.2. Compras governamentais

Existem diversas modalidades de compras governamentais e apenas parte delas foram utilizadas nesse trabalho. A seguir elas estão descritas, bem como as leis que definem o processo de compras públicas.

A lei Nº 8.666 ([da República 1993]), de 21 de junho de 1993, tem como objetivo garantir a supremacia do interesse público sobre o privado, garantindo a livre concorrência de contratações para diversas atividades públicas de modo que o governo obtenha a contratação mais vantajosa e que ocorra uma troca de interesses entre particulares.

Existem alguns aspectos da lei que precisam ser analisados para realização de uma compra, por exemplo, a lei define que o menor preço é um critério de escolha que vem sendo cada vez menos utilizada, salva algumas exceções, permitindo a realização de contratações mal cumpridas pois o licitante coloca um preço exorbitantemente baixo mas incapaz da realização de todos os aspectos do contrato com qualidade e eficiência.

Os tipos de compra utilizados neste trabalho foram: Ata de registro de preço e Pregão.

Ata de registro de preço é um documento especial. O objetivo dele é servir de base para futuras compras e contratações onde se registram preços, fornecedores, órgãos participantes e condições a serem seguidas. Ela engloba também a adesão de outros órgãos que estejam interessados em aderir ao contrato sob as mesmas condições do edital, desde que enviem uma solicitação via o sistema Compras Governamentais. Portanto, a ata de registro de preços não define uma contratação em específico como nas outras modalidades de licitação e sim define um espaço registrado de garantia para futuras contratações por parte do governo com validade de um ano, podendo ocorrer as contratações futuramente ou não, garantindo uma edição das condições de cotações mínimas.

O pregão é um modelo de leilão, que é a modalidade de licitação de venda de bens que se tornaram do estado de alguma forma em função de penhora ou pagamento, para bens comuns. Ele funciona no sentido em que é feito um anúncio em sessão pública para aquisição de um bem ou serviço e é realizada uma série de propostas e lances por meio dos licitantes. Uma grande diferença está na fase de habilitação e análise de propostas onde a ordem é invertida e se verifica apenas a documentação do ofertante que tenha apresentado a melhor proposta, sendo que ainda pode haver uma negociação direta com o pregoeiro para uma diminuição maior no valor ofertado. Outro fator de definição do Pregão seria a aplicação da modalidade para qualquer valor de contratação, de modo que seja uma alternativa a todas as modalidades e, consequentemente, ocorre um crescimento maior dessa modalidade devido a englobar uma área maior de atuação.

2.3. Recuperação da Informação

Recuperação de informação (RI) é uma área da computação que lida com o armazenamento de documentos e a recuperação automática de informação associada a eles. É uma ciência de pesquisa sobre busca por informações em documentos, busca pelos documentos propriamente ditos, busca por metadados que descrevam documentos e busca em banco de dados, sejam eles relacionais e isolados ou banco de dados interligados em rede de hipermídia.

Sistemas de informações podem ser definidos como “Sistemas humanos de processamento de informação, sistemas eletrônicos de processamento de dados ou sistemas de recuperação da informação” [Araújo 1995] e, por conseguinte, Sistemas de Recuperação da Informação permitem a comunicação e o acesso aos conteúdos.

Quando os dados estão em organização e/ou formatos distintos, provenientes de fontes heterogêneas, é necessário uma ação de limpeza e organização prévios. Isso é necessário para empresas que armazenam uma grande gama de dados de clientes e necessitam se adaptar a constantes mudanças organizacionais e estruturais dos negócios. Portanto, é necessário criar vínculos entre tais dados para criação de indicadores que representem algum tipo de informação e *insight* negocial para a empresa, geralmente essa visualização é realizada por ferramentas OLAP (*Online Analytical Processing*). Além disso, é importante entender que com a evolução tecnológica os sistemas, tanto de *hardware* quando de *software*, tem se tornado cada vez mais complexos e conectados e, portanto, os dados tem ficado cada vez mais granularizados ao longo de diversas plataformas. Este é a principal finalidade de um *Data Warehouse* - DW.

Um data warehouse (DW) é um sistema de armazenamento digital que conecta e harmoniza grandes volumes de dados de várias fontes diferentes. Seu objetivo é alimentar relatórios, funções analíticas e dar suporte às exigências regulatórias para que as empresas transformem seus dados em conhecimento, facilitando a tomada de decisões.

Existem diversas definições, conceitos e ferramentas para criação e manutenção de um DW, onde são destacados os seguintes aspectos:

- Extração de dados de fontes heterogêneas;
- Transformação de dados antes de sua carga final;
- Normalmente requer máquina e suporte próprio;
- Visualização de dados em diferentes níveis;
- Utilização de ferramentas voltadas para acesso com diferentes níveis de apresentação;
- Dados são apenas inseridos.

Aliado a isso, observa-se que o grande volume de dados, após passar por este processo de DW, precisa ser transformado em conhecimento, algo que seja útil para as organizações. Não existe um concesso geral para uma definição única sobre Mineração de Dados - MD. O trabalho de [Camilo and da Silva 2009] apresenta diversas definições, de acordo com o ponto de partida. Partindo do KDD (*Knowledge Discovery in Databases*), tem-se que a MD é apenas uma etapa do processo de geração do conhecimento em que são utilizadas técnicas para conseguir realizar uma tarefa; já partindo do lado estatístico tem-se que MD é a análise de grandes conjuntos de dados com o propósito de encontrar relacionamentos inesperados e resumir os dados de forma útil; para uma perspectiva de banco de dados tem-se que MD é um campo interdisciplinar que reúne técnicas de IA, reconhecimento de padrões, estatística, banco de dados e visualização para extrair informação; além disso, a perspectiva do lado do aprendizado de máquina diz que MD é um passo na descoberta do conhecimento que tem sua base na análise de dados e aplicação de algoritmos de descoberta produzindo padrões. De forma geral, pode-se dizer que MD utiliza técnicas computacionais divididas em fases para lidar com grandes quantidades de

dados a fim de permitir análise daqueles dados, sendo o produto final ou apenas uma etapa de um processo maior para gerar conhecimento.

3. Trabalhos relacionados

Nesta seção são descritos trabalhos relacionados ao uso de MD sobre DA governamentais. Observe que são exemplos de aplicações sobre áreas distintas da saúde e nem sempre estão relacionadas com compras. Caracterizam-se por apresentarem questões comuns relativo a acesso aos dados, falta de padronização, necessidade de limpeza e organização. Da mesma forma, permitem compreender a necessidade desses dados serem abertos, pois trouxeram ganhos à sociedade.

A empresa Aquarela, que atua na área de soluções analíticas utilizando IA, possui um *case* de sucesso utilizando MD de DA na área da saúde. Apesar de não ser aplicada à área de compras, a empresa utilizou MD para diminuir o número de pacientes que faltam a consultas marcadas na cidade de Vitória, Espírito Santo e, dessa maneira, reduzir o prejuízo aos cofres públicos [Aquarela 2017].

A solução validou mais de 1.575.487 registros colhidos nos anos de 2014 e 2015 analisando diversos aspectos como: idade, sexo, dia da semana, tempo de espera, deficiência, alcoolismo, dentre outros fatores. O processo mais demorado foi a parte da limpeza de dados, Foram criados nove perfis de pacientes faltosos, dentre esses, quatro grupos representavam 75% das ocorrências, além de pontuarem os aspectos que mais influenciavam as faltas, que são, respectivamente: idade, espaço entre o dia de marcação e atendimento e, por fim, presença de algum tipo de deficiência física. Também foi possível verificar que os dias mais propensos a faltas são segundas e sextas.

Após toda essa análise e processamento das amostras o resultado foi a economia de pelo menos R$ 1,3 milhões no primeiro ano e a redução do volume de faltas de 30,14% dos agendamentos para 16% por meio de indicações de ações como: perfis mais faltosos serem associados aos dias que são mais propensos aos pacientes comparecerem, utilização da estratégia de *overbooking*, sugestão da criação de um sistema de *check-in online* e duplo *check-in* para perfis de risco. A análise também conseguiu validar que o envio de SMS para confirmação da consulta, que tem um custo de R$ 10.000 reais por mês, possui uma taxa de relevância muito baixa, podendo realocar esse dinheiro em outras áreas.

Outro *case* de sucesso é a Operação Serenata de Amor. O projeto tem como premissa auxiliar na auditoria pública de pedidos de reembolso, com o gasto de despesas necessárias para a realização das atividades da função de senadores e deputados, utilizando os DA brasileiros por meio de MD, criando uma ferramenta contra a corrupção e ao auxílio da transparência dos gastos públicos [Lima 2019].

O primeiro desafio foi a transformação dos dados consumidos pelo Portal da Câmara em disponíveis e acessíveis, tendo a necessidade de refinar os dados e dar sentido a eles. A partir desses dados refinados, iniciou-se o processo da criação do mecanismo de Inteligência Artificial, apelidada de “Rosie” . Para realizar a visualização dos dados, após o treinamento e validação do trabalho feito pela Rosie, foi criado o “Jarbas”, um *dashboard*, que permite a visualização simples de todos os dados integrados além de colocar a fonte de cada um deles; porém a maior divulgação de casos suspeitos é feita automatica-

mente por um perfil no *Twitter* da Rosie[6] e, assim, todos os perfis da plataforma podem visualizar a mensagem.

A operação é um sucesso, o projeto já sinalizou mais de 8.000 reembolsos suspeitos, que correspondem a mais de R$ 3,6 milhões de reais, colocando em torno de 735 deputados diferentes em alerta. A equipe denunciou 629 reembolsos suspeitos à Câmara dos Deputados, envolvendo 216 deputados diferentes e mais de R$ 378 mil.

O projeto em construção Cuidando do meu bairro[7] tem como propósito que a população exerça o controle e a fiscalização dos gastos realizados em equipamentos públicos da cidade e promova ações concretas no seu bairro.

O sistema utiliza um mapa da cidade de São Paulo e, por meio de um código de cores é possível apresentar em tempo real o status de um projeto iniciado e em qual área ele afeta, integrado ao E-Sic da Prefeitura de São Paulo. O projeto conta com um sistema de cadastro que permite os usuários acompanharem projetos específicos e, a cada alteração de estado, é enviada uma notificação. Além de contar com um sistema de perguntas e comentários de acordo com cada projeto, promovendo a discussão e a validação dos usuários com as obras que impactam diretamente seus bairros.

O projeto é desenvolvido pela equipe da *Open Knowledge Brasil* e o acompanhamento pode ser feito pelo blog[8] e já conta com mais de 1800 projetos mapeados e acompanhados em tempo real.

O trabalho de [Gomes 2015] faz utilização de técnicas de mineração de dados relacionados aos dados abertos governamentais da prova do ENEM, prova de conhecimentos gerais realizada, geralmente, por alunos na conclusão do ensino médio.

O projeto utilizou dados abertos disponibilizados pelo INEP da prova de 2014, a mais recente quando foi desenvolvido. Foi utilizado o KDD como base para o processo de mineração de dados, analisando mais de 5 gigabytes de dados em mais de 8 milhões de tuplas no formato CSV disponibilizado pelo INEP. A primeira etapa do processo foi a limpeza e separação desses dados, sendo necessária a criação de um programa em Python para fazer uma leitura separada e mais lenta desses dados, utilizando menos memória. A limpeza foi realizada retirando informações ausentes, inconsistências e valores não pertencentes ao domínio, isto é, não relacionados ao problema; houve também uma redução da dimensionalidade de dados de amostragem por meio de três passos, seleção de grupos e subgrupos representativos, cálculo de peso relativo e seleção de população, transformando os dados e facilitando o seu entendimento.

Foram selecionadas de forma aleatória 10 mil tuplas de amostragem, contudo, optou-se por selecionar dados da região nordeste para direcionar a pesquisa e os resultados para a região local do autor. Os dados então foram integrados ao *Weka*, um software escrito em Java, sob licença de uso geral público, que se trata de uma coleção de algoritmos de aprendizagem de máquina que possibilitam mineração de dados. Foram realizadas as seguintes tarefas de mineração de dados: agrupamento, classificação e associação. Foi

[6] `https://twitter.com/rosiedaserenata`
[7] `https://cuidando.vc/?/home`
[8] `https://colab.each.usp.br/blog/tag/cuidando-do-meu-bairro/`

utilizado o algoritmo *APRIORI* (construção de regras de associação), para análise de provas objetivas e o *J48* (árvores de decisão) para análise de dados referentes a redação.

O resultado das análises foram variados, desde associações de senso comum como inscritos que realizaram ensino fundamental somente em escola pública apresentam forte tendência a realizar o ensino médio em escolas públicas ou famílias da classe E apresentam forte tendência a estudarem em escolas públicas. O trabalho também reiterou algumas afirmações como relação entre notas abaixo da média nacional com baixa de renda de família e suas respectivas classes, relação de redações em branco ligadas a famílias de classe A, B e C enquanto famílias das classes D e E possuem tendências a realizar fuga ao tema; abrindo espaço para uma série de questionamentos e avaliação da educação pública brasileira.

O trabalho de [Zacarias et al. 2019] realiza mineração de dados sobre dados abertos para descoberta de conhecimento extraídos dos boletins de ocorrência de rodovias federais brasileiras do ano de 2012.

O maior problema enfrentado pela equipe foi a etapa de pré-processamento de dados, para a eliminação de ruídos e imperfeições. Pode-se destacar:

- Diversos atributos presentes não são úteis no processo de MD e para o próprio sistema BR-Brasil, sistema utilizado pela polícia rodoviária brasileira, pois eles foram descontinuados em novas versões do sistema ou alterados.
- Grande quantidade de dados faltosos, de 390.973 registros foram encontradas 181.428 ocorrência de dados faltantes, representando mais de 46% dos registros.
- Dicionário de dados incompleto e falta de padronização na sua utilização.
- Algumas tabelas, como a que identifica o modelo e a pista, não estão no conjunto de dados públicos.
- A separação de ocorrência por semestre é feita com base na data de finalização, ou seja uma ocorrência que foi aberta em 2001 mas só foi finalizada em 2012 é registrada no ano de fechamento.
- Alteração de formatação de diversos dados como data, municípios, data de fabricação do veículo e horário de acidente.

Portanto, foi dedicado muito tempo para realização de técnicas de limpeza de dados para garantir a padronização de dados ao longo de todos os registros e minimizar a ocorrência de *"?"*, símbolo utilizado para atributos faltosos, e seleção de atributos relevantes para o processo de adquirir conhecimento e constantes ao longo dos registros. Na parte de mineração de dados foram utilizados os algoritmos *J48* (árvores de decisão), *PART* (indução de regras de conhecimento) e *APRIORI* (construção de regras de associação).

As análises realizadas pelo *PART* e *J48* foram feitas utilizando a técnica de validação cruzada e estratificada e, como a classe majoritária do conjunto de dados é "Falta de atenção", com 46,3% dos exemplos, o erro majoritário utilizado foi 53,7% e a média de taxas de erro são menores ou iguais a esse valor. Foram criadas nove regras de associação pelo algoritmo *J48* e seis regras selecionadas de árvore de decisão induzida pelo algoritmo *PART*, ambos relacionando diversos atributos ao longo dos resultados utilizando métricas como tipo de pista, tipo de veículo, causa do acidente, estado físico da

Tabela 1. Resumo comparativo dos trabalhos relacionados

Trabalhos Relacionados	Fonte dos dados	Ações
[Aquarela 2017]	Secretaria de Saúde Vitória-ES	Mineração de Dados
[Lima 2019]	Câmara Federal	Limpeza de dados, múltiplas fontes
Cuidando do meu bairro	Projetos na cidade de SP	visualização e múltiplas fontes
Dados Abertos Educacionais	dados ENEM do Nordeste	limpeza de dados para cruzamento e projeções
[Zacarias et al. 2019]	estradas federais	limpeza de dados, mineração e cruzamento
Proposta	compras municipais da área da saúde	limpeza de dados, mineração e cruzamento

pessoa. Para o algoritmo *APRIORI* foram geradas oito regras de associação com taxa de confiança maior que 0,8 e duas acima de 0,9; novamente relacionando diversos tipos de atributo como tipo de veículo, tipo de pista, estado físico da pessoa, hora do dia. Pode-se observar algumas interessantes, como, por exemplo, a regra "SE Causa do Acidente = Não Guardar Distância de Segurança ENTÃO Tipo de Acidente = Colisão Traseira".

A Tabela 1 apresenta uma comparação entre os principais trabalhos relacionados com a proposta do presente artigo. Observa-se que em todos a atividade primordial foi a limpeza e reorganização dos dados para facilitar a obtenção de informações úteis. Também vale salientar que os cruzamentos e perguntas nesta etapa eram direcionadas, ou seja, já havia questões prévias a serem respondidas e a mineração estava dirigida.

Um dos questionamentos levantados pelo trabalho foi a discussão de uma das premissa dos dados abertos terem de ser disponibilizados da forma como foram coletados, o que gera uma grande dificuldade de leitura e um grande tempo na etapa de pré-processamento e limpeza dos dados pela falta de padronização e de serem difíceis de entender tanto pelo homem quanto pela máquina. Os autores sugerem que, além da forma já existente, os dados sejam disponibilizados utilizando algum tipo de padrão. No texto eles citam triplas RDF, para facilitar a leitura automatizada por máquinas. Outro ponto foi a interpretação dos dados disponibilizados não serem uma tarefa trivial para o cidadão e ressalta a importância de facilitar a leitura por máquina dos dados disponíveis para iniciar projetos com princípios de *web* semântica, isso permitiria a criação de plataformas de visualização de dados de forma mais orgânica para o cidadão e dessa forma democratizar o conhecimento.

4. Desenvolvimento

O escopo inicial do trabalho foi escolher um município para avaliar e implementar uma solução de análise dos gastos públicos com saúde. Foi escolhida a saúde como foco devido a facilidade do acesso entre os meios de consumo de dados, padronização de equipamento utilizado e ser um assunto que impacta a maioria dos brasileiros.

É importante destacar que parte das compras da área da saúde são realizadas pelo Estado e, em seguida, distribuídas entre seus municípios. Além disso, também existe um sistema federal de compras governamentais onde qualquer outra entidade (estados, municípios, órgãos) podem aderir a compras executadas por outras entidades, desde que o vendedor aceite adicionar.

O porte dos municípios brasileiros varia bastante e, a presença de *sites* com apresentação de dados e sistemas de gerenciamento e apresentação desses dados estão mais presentes em capitais e municípios com maior capacidade de arrecadação.

De qualquer forma, os municípios possuem autonomia para realizar todo o processo de compra e, por último, as compras devem estar publicizadas nos Diários Oficiais. Esta é a forma de divulgação exigida pela legislação brasileira e, por esta razão foi considerada a fonte de dados. Deve-se lembrar que o formato dessa divulgação é em PDF.

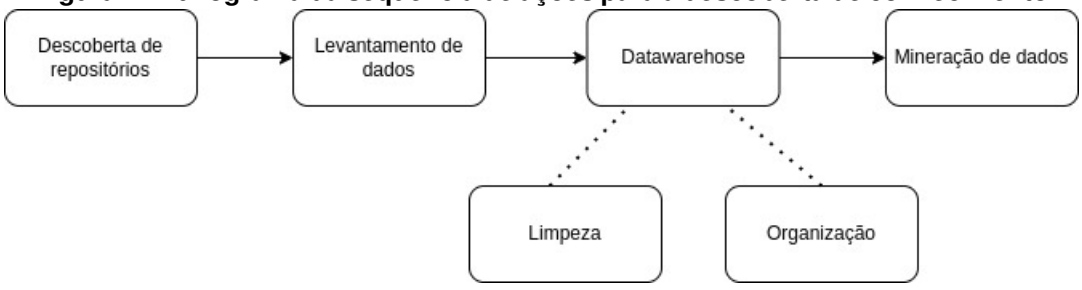
A primeira etapa, portanto, consistiu em buscar de sistemas municipais, com interface pública pela *web* e/ou *sites* que continham informações sobre gastos de saúde.

A etapa de coleta de dados foi seguida pela etapa de avaliação das informações contidas e observou-se que os dados não seguiam um padrão e, a cada licitação por exemplo, seus dados seguiam formatos aleatórios de publicação, o que praticamente inviabilizava algum mecanismo de automatização. A limpeza de dados espúrios e organização da ”informação”era imprescindível. Esta etapa permitiu a compreensão de como a falta de estrutura e organização desses dados podem prejudicar a transparência desejada.

A etapa seguinte - mineração de dados - foi realizada de forma simplificada, apenas para mostrar sua viabilidade.

O fluxograma apresentado na Fig. 1 permite uma melhor compreensão das etapas envolvidas neste processo.

Figura 1. Fluxograma da sequência de ações para a descoberta de conhecimento



Fonte: Autor

As subseções a seguir apresentam as tecnologias utilizadas, abordam o histórico do levantamento das fontes de dados, descrevem o entendimento da coleta e transformação desses dados, problemas e soluções encontradas.

4.1. Tecnologias utilizadas

Conforme descrito em [Hardy 2017], o formato PDF foi criado pela *Adobe*[9] em 1993 com a intenção de ser um formato para documentos e imagens que seria independente do sistema operacional, *hardware* ou *software* para visualização. Cada arquivo PDF possui metadados que incluem texto, fontes, gráficos baseados em vetores, imagens e demais informações necessárias para criação de uma descrição completa de um *layout* fixo capaz de ser interpretado. O PDF foi padronizado como ISO 32000 em 2008 e, portanto, não existe a necessidade de *royalties* para sua implementação. Arquivos PDF contém uma variedade de conteúdo além de apenas textos e gráficos. Versões recentes permitem imagens, arquivos de multimídia, como vídeo e áudio, além de conteúdos auxiliares como anotações, *bookmarks*, arquivos de anexos, *hiperlinks*, estruturas lógicas e metadados. O PDF passou também a ser utilizado como forma final de documentos pela sua segurança acentuada, permitindo os arquivos estarem atrelados a senhas, além de serem mais difíceis de serem capturados por *listeners* externos, pelo fato das informações de texto estarem contidas em *streams* de conteúdo que são colocados nos devidos lugares por *bitmaps* que indicam a posição de tudo baseado no *layout* fixo do arquivo.

A codificação foi realizada com o uso da linguagem de programação Python, pela facilidade na extração de dados, quanto para a sua análise, contendo uma grande comunidade com bastante material para leitura, tecnologias atuais de mercado e ser uma linguagem eficiente para a realização de tarefas de mineração de dados ([Python Software Foundation 2021]).

Para acessar dados presentes em sites e páginas HTML estáticas ou dinâmicas fez-se necessário o uso de duas tecnologias: *Web Crawling* e *Web Scraping*. Tais termos, muitas vezes usadas de forma intercambiável, estão melhor definidos a seguir.

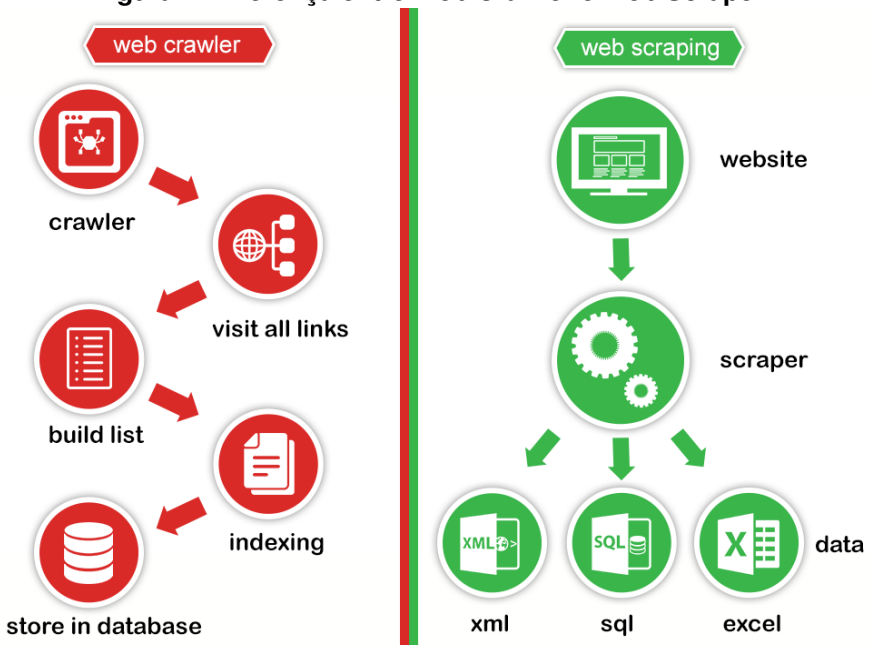
Web Scraping é definido por [Broucke and Baesens 2018] como a construção de um agente para baixar, parsear e organizar dados da *web* de forma automatizada. Portanto, pode-se entender como a substituição de um trabalho manual humano utilizando técnicas de computação, por exemplo, ao invés de o usuário realizar a cópia de informações de um *site* e colocar em uma planilha, essa ação pode ser automatizada para ser realizada de forma mais rápida, eficiente e correta do que um humano poderia realizar.

Web Crawling, segundo [Shukla and Roy 2016], pode ser definido como a criação de programas excepcionais, chamados de aranhas (*Spiders*). Tais aranhas têm como função passear por uma grande diversidade de páginas *Web* para identificar algum parâmetro específico, geralmente indexando essas páginas à procura de palavras-chaves. Pode-se citar, como exemplo de *Web crawler*: Google, Yahoo e outros *sites* de busca que indexam páginas *web*. Por exemplo, eles entram em determinada página na *web*, indexam o seu conteúdo HTML em busca de URI semelhantes e montam uma lista de páginas que a aranha deve seguir. Essa etapa ocorre de forma recursiva, indexando e armazenando o conteúdo de cada página. Dessa forma, quando uma pesquisa é feita, ele sabe indicar quais *sites* tem conteúdo relacionado àquele tema. Existem diversas técnicas para se abordar esse problema, de acordo com a necessidade.

[9] `https://www.adobe.com/`

Dessa forma, pode-se ver que todo *Web crawler* faz *Web scrapping* para conseguir indexar as páginas, mas geralmente um *Web crawler* irá em cada página de um *site*, ao invés de um conjunto específico. Além disso *Web scrapping* está focado em um tipo de dado para formar um *data set*, podendo transformar esse dado antes de salvá-lo ou não, enquanto *Web crawlers* escaneiam e salvam todo conteúdo do *site*. Um exemplo da diferença seria a criação de um *bot* para analisar o preço de um produto a cada hora, verificando promoções ou montando um histórico do preço. Um *bot* que realize essa tarefa para um produto ou uma série de produtos seria *Web scrapping*, se o *bot* passeasse pelo *site* em busca de diversos produtos além de buscar o mesmo produto em outros *sites* poderia ser considerado um *Web crawler*. A Figura 2 ajuda a compreender essa diferença.

Figura 2. Diferença entre Web Crawler e Web Scraper



Fonte: Pro Web Scraping (http://prowebscraping.com/web-scraping-vs-web-crawling/), acessado em 03/09/2020

Para a navegação de páginas e *sites*, faz-se necessária a utilização de ferramentas que automatizem essa navegação. Neste projeto foi utilizado o Selenium[10], que é um conjunto de ferramentas capaz de automatizar navegadores *web*. Esta ferramenta inclui identificação e interação com elementos da página *web* sendo capaz de realizar ações como preencher formulários, escolher opções em listas e clicar em botões. Por esse motivo o Selenium também é utilizado para *Web scraping* em situações que a retirada de informações exige a interação com o navegador, pela informação estar oculta ou depender de uma ação dinâmica ([Selenium 2018]).

A navegação entre hierarquia de páginas HTML necessita de conhecimento e manipulação de requisições HTTP, típicas na *internet*. Requests é uma biblioteca Python

[10] `https://www.selenium.dev/`

para lidar com essas requisições HTTP, sendo uma das bibliotecas mais famosas e utilizadas em Python. O objetivo da biblioteca é facilitar a utilização de métodos do padrão HTTP como POST, PUT, GET, etc; sem a necessidade de adicionar *Query strings*, lidar com *pool* de conexões e demais erros relacionados. Ao realizar uma chamada, é construído um objeto Requests que será enviado ao servidor para realizar a requisição de algum recurso; um objeto *response* é gerado e o programa recebe outro objeto Requests que encapsula esse resultado, portanto o criador da aplicação só precisa conhecer os objetos Requests ([Reitz 2011]).

4.2. Levantamento das fontes de dados

Inicialmente foi realizada uma busca simples com o modelo *"nome do município, dados abertos"* e foram encontrados os *sites* de transparência municipais das cidades do Rio de Janeiro[11], São Paulo[12] e Belo Horizonte[13]. Vale destacar que são municípios com grande população e mais recursos, comparados com os demais municípios brasileiros. Poucos municípios brasileiros possuem portal de transparência com dados disponibilizados sobre despesas e receitas. Como característica comum entre os encontrados, foi evidenciada a utilização de uma divisão por categorias, como visto na Figura 3, apesar de não haver padronização de categorias entre os municípios. Ao analisar o conteúdo da categoria de saúde, comum em todos os sistemas, foi observado que não existe uma grande quantidade de dados e os dados disponíveis se limitam a folhas de pagamento ou registros de hospitais. Não é informada a relação de gastos públicos com equipamentos, conforme evidenciado pelas Figuras 4 e 5. Em vista da baixa quantidade de dados encontrados na primeira busca foi realizada uma nova pesquisa de novos *sites* abordando palavras como licitação, câmara municipal, compra direta e registro de preço.

Figura 3. Disposição de Categorias do *site* da transparência da prefeitura de São Paulo

Fonte: Consulta sobre o site http://dados.prefeitura.sp.gov.br/, acessado em 10/09/2019

Foi encontrado, para o município de Belo Horizonte, o *site* da câmara municipal[14] que possui uma parte destinada à transparência. São dispostos apenas gastos gerais em

[11] http://www.data.rio/
[12] http://dados.prefeitura.sp.gov.br/
[13] https://dados.pbh.gov.br/
[14] https://www.cmbh.mg.gov.br/transparencia-principal

Figura 4. Resultado da busca pela categoria de saúde do *site* da transparência da prefeitura de Belo Horizonte

Fonte: Consulta sobre o site https://dados.pbh.gov.br/dataset?q=sa%C3%BAde, acessado em 10/09/2019

Figura 5. Resultado da busca pela categoria de saúde do *site* da transparência da prefeitura de São Paulo

Fonte: Consulta sobre o site http://dados.prefeitura.sp.gov.br/group/saude, acessado em 10/09/2019

formato PDF e sem opção de pesquisa por órgão ou categoria. Não foram encontrados resultados para pesquisa em relação às licitações concluídas de itens comuns na área de saúde como gaze ou algodão. Quando pesquisado sobre o tema saúde foram encontrados somente 10 resultados com assuntos variados, desde lavanderia, fornecimento de mão-de-obra e instalação de equipamentos, que fogem do escopo do trabalho.

Há um sistema no Estado de São Paulo chamado Sigeo[15] que parecia conter as informações necessárias para o projeto. O sistema informa quanto foi gasto em determinados itens, mas não informa o valor unitário de cada compra e nem foi possível encontrar os dados de categorias ou separação exclusiva ao município, também fugindo do escopo.

A prefeitura do município do Rio de Janeiro apresentou o maior número de informações e de sistemas. O domínio E-Compras Rio[16] apresenta informações relevantes mas ainda assim não disponibiliza os dados por meio de um formato facilmente exportável, como CSV ou JSON. Todos os arquivos são disponibilizados em PDF, a separação é feita por órgão, não permitindo uma busca mais específica, os dados se limitam aos últimos dois anos com uma quantidade extremamente baixa de dados e sem exportação automática, mostrado na Figura 6.

Figura 6. Resultado da ata de registro de preços do e-compras Rio do município do Rio de Janeiro

Fonte: Consulta sobre o site E-Compras Rio, acessado em 10/09/2019

Outra plataforma de informação interessante é o *site* Rio Prefeitura [17] que disponibiliza um arquivo XLS com as despesas gastas pelo município, mas sem descrição de quais itens foram comprados, qual a quantidade e nem para qual órgão. Existe ainda o Catálogo de Itens Sco-Rio[18], que possui códigos de compras ligados ao *site* Rio Prefei-

[15] https://www.sigeo.fazenda.sp.gov.br/analytics/saw.dll?PortalGo
[16] http://ecomprasrio.rio.rj.gov.br/portal/pagina_inicial.asp
[17] http://prefeitura.rio/web/contasrio/despesa-por-acao#titulo
[18] http://www2.rio.rj.gov.br/sco/

tura, que também só disponibiliza exportação via PDF e que também não tem busca por categoria, órgão ou descrição, além de só possuir itens relacionados a construção civil.

O *site* do sistema de compras públicas do estado do Rio de Janeiro[19] foi o domínio mais próximo do desejável para criação do protótipo, pois existe uma divisão entre materiais e serviços, divisão por categorias e escolha de itens específicos, além de que, para cada item, é definida a modalidade da compra. O sistema, contudo, possui uma série de problemas, o *site* só vem recebendo dados nos últimos dois anos, além disso um registro de um item no banco de dados já qualifica como item existente na lista de produtos apesar de não ter sido feita nenhuma compra. Portanto, diversos itens, como atadura, possuem dois ou mais códigos e descrições ligeiramente diferentes mas nenhuma compra associada a eles, gerando uma grande quantidade de lixo. Foi realizado contato com os administradores do sistema para tentar ter acesso ao banco de dados para consulta, porém não foi permitido acesso às informações desejadas.

Devido à dificuldade de encontrar um sistema único municipal, houve uma alteração de estratégia para pesquisar por *sites* gerais de dados abertos. O portal brasileiro de dados abertos[20] possui um modelo interessante. Os dados estão divididos em: organizações, grupos, etiquetas, formatos e licenças. Além disso dados possuem diversas formas de exportação comportando os modelos mais comuns e disponibilizam, em alguns casos, um PDF contendo um dicionário de dados. O portal também comporta uma seção exclusiva para disponibilizar aplicativos que fazem consumo desses dados, inclusive o sistema "cuidando do meu bairro", que foi abordado na seção 3, está disponível. Para dados no grupo de saúde, havia menos de 25 conjuntos de dados, na sua grande maioria são indicadores ou localizadores de serviços relacionados ao tema como postos de saúde e farmácias populares. Existe um único arquivo que disponibiliza o preço tabelado de medicamentos, mas não disponibiliza a variação histórica e foge do escopo do trabalho.

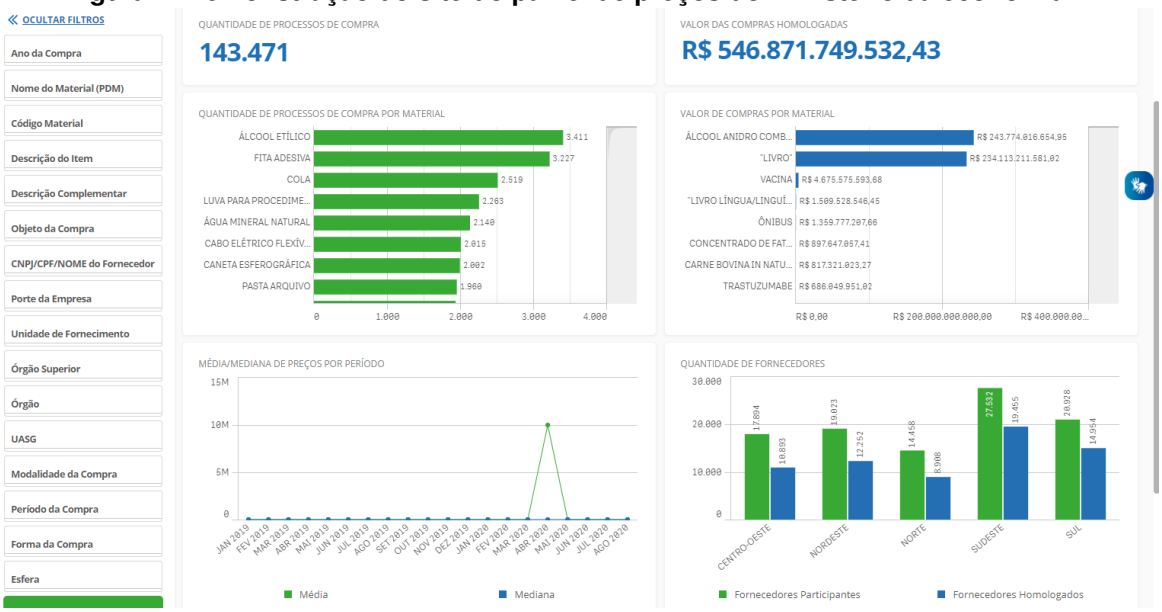
A página do painel de preços do Ministério da Economia[21] possui a melhor estrutura encontrada de um sistema dedicado a disponibilização de informações sobre compras públicas homologadas no Sistema de Compras do Governo Federal - COMPRASNET. Conforme apresentado na Figura 7 o sistema possui busca por ano de compra, material, serviço, CNPJ, Órgão expedidor, dentre outros. Quando se faz o detalhe de uma busca ele traz a variação do preço ao longo do tempo, quantidade de itens adquiridos por mês, média e mediana dos preços ao longo do tempo, compra e dívida por estados além da possibilidade de gerar um relatório detalhado nos formatos CSV e XLS. O sistema funciona como uma tentativa de auxiliar os gestores públicos em execuções de processos de compras, dando transparência em relação a todo o processo e permitindo o controle social. Porém, devido a baixa quantidade de dados, não é possível realizar uma mineração de dados utilizando os dados fornecidos por esse sistema. Além disso, os municípios podem ou não aderir a este sistema, ou seja, eles possuem autonomia para realizar as aquisições de forma independente.

[19] `https://www.compras.rj.gov.br/Portal-Siga/BancoDePrecoHistorico/listar.action`

[20] `http://dados.gov.br/`

[21] `https://paineldeprecos.planejamento.gov.br/analise-materiais`

Figura 7. Demonstração do *site* do painel de preços do ministério da economia



Fonte: Consulta sobre o site https://paineldeprecos.planejamento.gov.br/analise-materiais, acessado em 16/09/2020

Todos os dados de compras de municípios, estados e demais órgão públicos devem estar publicizados no Diário Oficial, em cada esfera pública. O município do Rio de Janeiro foi escolhido como base de pesquisa, devido a disponibilização da maior quantidade de sistemas e informações, para busca das informações de compras públicas por meio do diário oficial. Segundo a lei Nº 8.666/1993, o diário oficial deve ser o meio de divulgação pública oficial para quaisquer licitações e contratos administrativos pertinentes a obras, serviços, inclusive de publicidade, compras, alienações e locações no âmbito dos Poderes da União, dos Estados, do Distrito Federal e dos Municípios. O *site* do diário oficial do Rio de Janeiro[22] possui busca por termos, podendo ser utilizada uma busca exata do termo ou aproximada. Para a opção de *download* ele permite baixar o PDF do diário oficial completo ou apenas a página em que o termo aparece, além disso as datas de busca iniciam a partir do ano 2000.

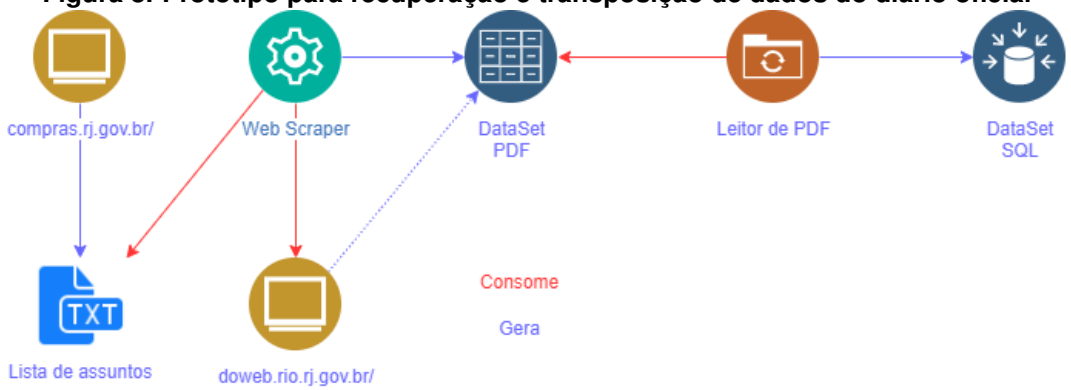
4.3. Coleta dos dados

Foi necessária a criação de um *web scraper* para recuperar as informações de forma automática do *site* do diário oficial do Rio de Janeiro. A Figura 8 mostra a diagramação de como a parte do protótipo para recuperação dos dados foi criada.

Para realizar essa etapa de retirada dos dados foi criada uma lista a partir das descrições dos produtos encontrados no sistema de compras públicas do estado do Rio de Janeiro, devido ao fato de já ter registrado em seu sistema termos utilizados em compras públicas, junto das datas de início e fim de pesquisa. O programa faz a leitura de um arquivo de texto chamado *arquivo_busca_pdf.txt* onde cada linha representa um assunto a ser procurado. O objetivo do programa é funcionar para qualquer número de descrições

[22] `https://doweb.rio.rj.gov.br/`

Figura 8. Protótipo para recuperação e transposição de dados do diário oficial



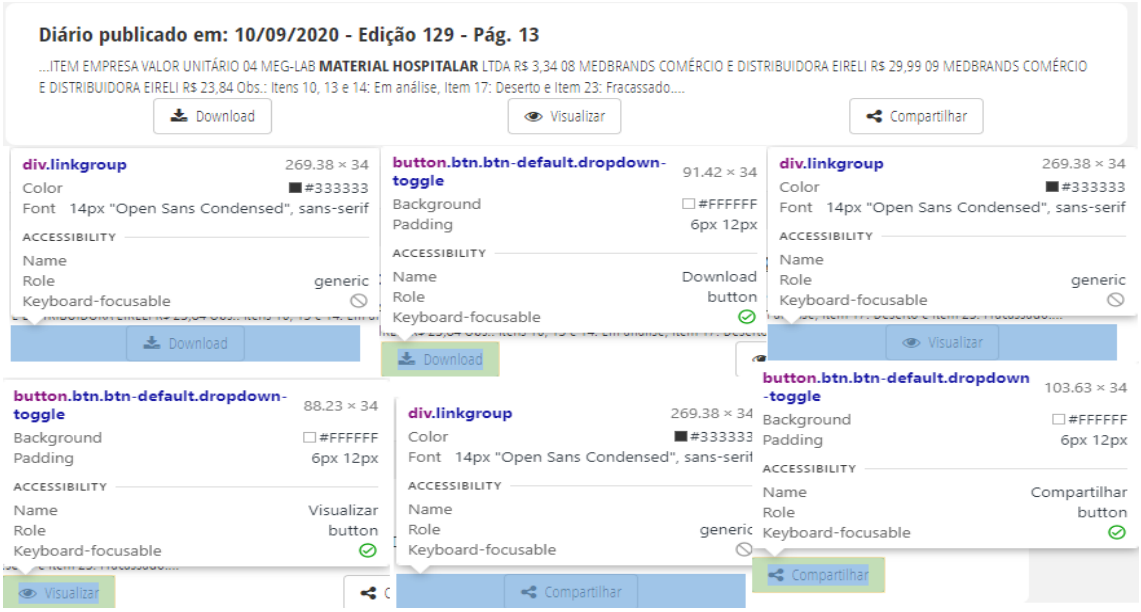
Fonte:Autor

de assuntos que venham a aparecer no diário oficial, desde que sigam o seguinte formato: *xxx, di=20100409, df=20110416*; onde xxx é o assunto a ser procurado, di e df representam, respectivamente, as datas de início e fim da procura, seguindo o formato yyyyMMdd.

Devido a necessidade de espera de carregamento de elementos e também interação com a página, foi utilizado o *Selenium* para reconhecimento e interação com o sistema. O próximo passo foi automatizar o *Download* de qualquer quantidade de PDFs encontrados.

Um dos problemas encontrados para realização dessa etapa foi o isolamento do botão de *Download*, pois foi utilizado o mesmo *alias* para os três botões assim como o mesmo nome para cada div, conforme visto na Figura 9.

Figura 9. Forma como foram construídos os elementos que compõem a página



Fonte: Autor

Por conta de não estar definido um nome único para cada componente, tornou-se

necessário passar por todos os elementos da classe da página com nome de *"dropup"* e pesquisar especificamente pelo elemento que contém o texto *" Download"*. Com isso é possível clicar no elemento através do *Selenium* e mostrar os dois botões filhos que permitem baixar o diário oficial completo ou apenas a página que possui o termo pesquisado. Optou-se por fazer o *download* apenas da página que contém o termo para poder isolar da melhor forma os dados pertinentes ao trabalho.

Ao clicar no botão é possível capturar somente a URI direta de *download* dentro da classe *link pdf-page*. Dessa forma, foi criada uma lista com todas as URI disponíveis da página. O processo é repetido até a quantidade máxima de páginas encontrada no canto superior direito como visto na Figura 10. Para isolar esse número foi buscada a *string* contendo "página 1 de", invertendo a ordem das palavras e salvando o primeiro elemento separado por espaço.

Figura 10. Demonstração de etapas para download de PDF utilizando o site do diário oficial do Rio de Janeiro

Fonte: Autor

Por fim o arquivo é salvo utilizando *Requests* que captura todo o conteúdo no momento que se acessa o endereço. Para evitar sobrecarregar os servidores e o computador, é baixado apenas um arquivo por vez em blocos de 1000 bytes.

A próxima etapa, conforme mostrado na Figura 8, seria processar os dados dos arquivos em PDF capturados pelo *Web Scrapper*. Para realizar essa etapa primeiro foi necessário transformar o conteúdo contido no PDF em um formato que fosse facilmente trabalhado e manipulado por meios computacionais, já que o conteúdo textual de um PDF está dentro de uma *stream* de conteúdo e, dessa forma, informações de tabelas e imagens estão salvas em *bitmaps* de informação, possuindo uma variação grande devido a abrangência permitida pelo PDF.

Dessa forma, é preciso reconhecer algum tipo de padrão para saber quais partes das informações disponíveis no PDF devem ser isoladas, a fim de facilitar o processo, em vista que a natureza de um diário oficial é retratar diversos temas, além de compras licitatórias, e os assuntos variam das mais diversas formas, desde nomeação de concursos, exonerações de cargos, erratas, dentre outros. Portanto, diversas informações, que não são relevantes para a extração, estão misturadas.

Após uma análise realizada em uma amostra utilizando como tema o termo "gaze", em arquivos PDF dos meses de 2020 até 2005, foi possível reconhecer limitações

e padrões dentro dos diários oficiais. Dentre os diversos tipos de licitações que existem, os que possuem a informação mais completa seriam as atas de registro de preço.

Os leilões apenas informavam valor total, empresa vencedora e qual licitação, do anúncio inicial, a qual aquela informação está relacionada. Portanto, a informação completa estaria separada em uma grande quantidade de arquivos, baseado no tempo em que o leilão ficou aberto, além da necessidade da descrição do produto conter a palavra chave utilizada no momento de pesquisa. As demais modalidades não possuíam informações completas como preço unitário de cada produto, valor total adjudicado, quantidade, ou seja, sempre faltava alguma dimensão de informação necessária. Isto é, quando a página do PDF possuía alguma informação relevante.

O *site*, que disponibiliza o *download* dos diários oficiais, funciona com um sistema de aproximação, ou seja, palavras parcialmente completas e apenas uma palavra em casos de palavras compostas. Isso pode ocasionar que uma série de dados baixados sem relevância. Por exemplo, na amostragem que foi feita o termo era "gaze", mas o diário oficial também trazia resultados relacionados a palavra "Gaz". A cada menção relacionada a alguém com nome próximo a "Gaz" acabava poluindo o pacote de PDF com dados não relevantes. Este problema pode evoluir de diversas maneiras, uma palavra como "algodão", que é um material importante e com compras recorrentes no sistema de saúde, também é utilizado para descrever o material de produtos têxteis.

Dentre os dados encontrados dentro da ata de registro de preço é possível encontrar um padrão que contém: Objeto, Órgão gestor, número do processo, modalidade de compra, empresa vencedora, itens a serem comprados, CNPJ da empresa, valor adjudicado e código do produto. Tais dados já servem como base para criação de um esquema estrela para criação do armazém de dados porque delimitam as dimensões pertinentes àquele fato, além de delimitarem as informações disponíveis relacionadas ao fato.

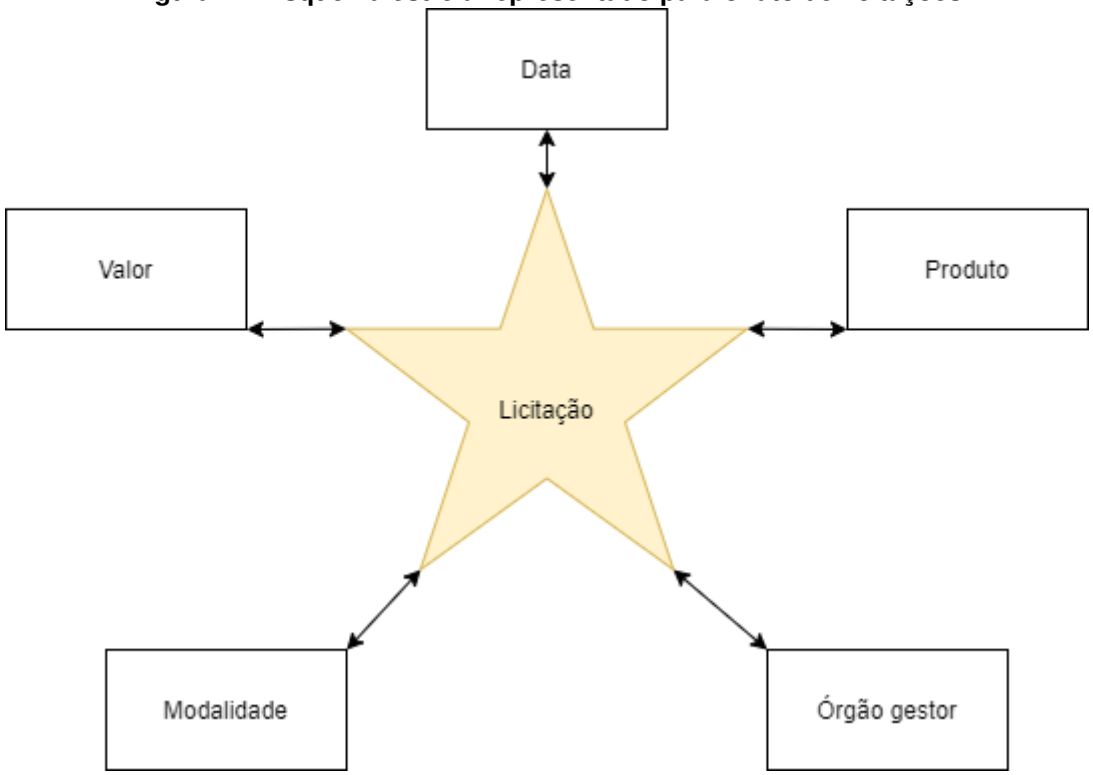
Com essas informações torna-se possível visualizar as etapas do processo de extração de dados. É necessário encontrar uma forma de extrair os dados provindos dos textos do PDF baseada em suas palavras chaves que se repetem ao longo de diversos processos licitatórios; é necessária e extração dos dados contidos em tabelas, assim como sua vinculação com o processo licitatório apropriado e, por fim, a criação de um objeto que possa ser gravado em alguma base de dados seguindo o modelo descrito na Figura 11.

Para a extração dos dados que estão contidos nas *stream* de conteúdo, é necessário transformar o texto do PDF em texto que possa ser manipulado. Ao utilizar *Python* existem diversas bibliotecas que auxiliam nesse processo com resultados variantes assim como formatos de *output* diferentes.

Na construção do protótipo para leitura dos dados, foram analisadas bibliotecas para leitura de texto: *PyPDF2, PyPDF4, Pdfreader e Pdfquery*; e para leitura de tabelas: *Camelot, Pandas e Tabula*.

De modo geral, todas as bibliotecas de texto possuem dificuldades na conversão de PDF porque o texto dentro de um arquivo PDF não é apenas informação de texto e sim *streams* de conteúdo que são alocadas em posições específicas de um *bitmap* ditando o formato e *layout* do arquivo. O PDF nunca foi criado com a intenção de ser um arquivo

Figura 11. Esquema estrela representado para o fato de licitações



Fonte: Autor

de armazenamento e sim como um arquivo final. Portanto, na conversão, frases sofrem quebras de linha de forma aleatória, algumas palavras se perdem no meio da transição e, graças ao formato em que o documento foi redigido utilizando duas ou mais colunas, a ordem das frases também pode ser perdida. Dentre as bibliotecas testadas, o pacote *Pdfreader* teve o melhor desempenho, conseguindo trazer o texto dentro do PDF para um ambiente de desenvolvimento, salvo que o texto ainda possui quebras de linhas irregulares.

Já as bibliotecas de tabela possuem suas próprias peculiaridades e problemas. Um deles, que não foi citado nas bibliotecas de texto, é o reconhecimento parcial ou nulo dos textos contidos dentro de tabelas. As tabelas estão dentro de uma estrutura própria que foge a convenção comum que a maioria das bibliotecas de texto foram programadas a reconhecer e, portanto, necessitam de bibliotecas especiais para realizar a extração dos dados. Sendo assim parte do conteúdo das bibliotecas de texto que é perdido, são os contidos em tabelas.

O maior problema relacionado aos dados em tabelas é que não existe uma diretriz comum para o diário oficial seguir em relação a formatação das tabelas ou palavras chaves que devem ser utilizadas. Por exemplo: em uma tabela, que é definida a quantidade de produtos comprados em determinada licitação, é usado o termo “QTD” , enquanto em outra é “quantidade” e em uma terceira é “qtde”. Outro exemplo de problema de formatação seria a implementação de uma linha vazia anterior ao cabeçalho para fins

estéticos, que acaba por quebrar a forma como a leitura poderia ser feita utilizando os nomes das colunas como determinante. Dentre as bibliotecas vistas, a que obteve melhor resultado foi a biblioteca *Tabula*, que utiliza o *Pandas* como extensão para estrutura dos dados recolhidos e para manipulação.

Com a escolha da biblioteca, é importante pensar em uma solução que funcione com uma certa margem de adaptabilidade, devido a sempre haver uma incerteza na forma como o texto será extraído pela biblioteca, como descrito acima. Como o texto segue um padrão de palavras chaves é possível capturá-lo utilizando *Regex*, a fim de ignorar todo texto que não for relevante ao fato da licitação.

Conforme dito por [Hope 2020], *Regex* é uma forma simplificada de informar uma expressão regular (*regular expression*) e pode ser definida como uma *string* que ajuda a criar padrões que podem localizar, isolar e organizar texto. Por meio de uma sequência de caracteres específicos é possível "programar" uma maneira de encontrar padrões com uma grande flexibilização de possibilidades.

O *Regex*, apreentado na Figura 12, foi especificado a fim de capturar certos atributos que satisfazem os dados do modelo.

Figura 12. Esquema Regex para captura

```
objetoBox = re.compile
    (r'(?P<objeto>Objeto(\s)*:(\s)*(\s)([^\.](\s)?)*\.)'
    r'(?P<Orgao_gestor>Orgao(\s)*Gestor(\s)*:(\s)*(.*)\.)'
    r'(?P<Num_Processo>Processo(\s)*:(\s)*(\s)\d{2}\/\d{3}\.\d{3}\/\d{4})'
    r'(?P<Modalidade>Modalidade(\s)*:(\s)*(\s)([^\.](\s)?)*\.)'
    r'(?P<EmpresaVencedoraAndItens>Empresa(\s)*Vencedora(\s)*\-(\s)*Ite(m|
   ns)([^\:](\s)?)*:(\s)*(.*?(\s)?)?CNPJ)'
    r'(?P<cnpj>CNPJ(\s)*:(\s)*(\d){2}\.(\d){3}\.(\d){3}\/(\d){4}\-(\d){2})'
    r'(?P<ValorAdjudicado>Valor(\s)*Total(\s)*Adjudicado(\s)*:(\s)*R\$(\s)
   .*,(\d){2})'
    r'(?P<codigo>(\d){4}\.(\d){2}.(\d){3}\-(\d){2})')
```

Fonte: Autor

Com o *Regex* é possível isolar cada parte do texto em pedaços que serão atribuídos a um objeto licitação. Primeiramente é feita uma validação de que a informação capturada é relevante para a criação do objeto da licitação. As palavras seguem uma ordem lógica pelo que a amostra indicava e, então, para cada texto capturado é validado se ele segue essa ordem, já que as palavras-chave não são exclusivas à licitação. Verifica-se se a primeira é "Objeto", a segunda "Processo" e a terceira é "Modalidade" sendo que essa deve ser definida como pregão, ata de registro ou licitação; com isso, é possível ver que a sequência

é de fato de uma compra governamental. Através do código, que está contido no texto, é realizada uma ponte entre o texto capturado e a informação contida nas tabelas em que a tupla, que contêm o código, é transformada em uma lista de elementos que então é separada e utilizada para a finalização do objeto, pronto para ser gravado em uma base de dados.

Com os dados coletados pode-se realizar tarefas de mineração de dados: agrupamento, classificação e associação. Observou-se uma grande variação de preços por unidade de itens simples como máscaras, luvas cirúrgicas, gazes, entre outros. Valores que ultrapassam em mais de 500%, mesmo considerando os índices de inflação do mesmo período e a variação da cotação real-dólar. Tais valores foram coletados antes da crise de oferta causada pela pandemia do Covid-19.

A literatura existente sobre o uso de mineração de dados, relativa a gastos de saúde, está relacionada ao quanto cada usuário custa para o sistema de saúde. Cruzamentos envolvendo faixa etária, sexo, região geográfica, enfermidade e outros são utilizados para a descoberta de padrões e possíveis "*outlyers*". Estudo da evolução de preços de itens de consumo não são considerados. O objetivo inicial deste trabalho consistia em usar a metodologia CRISP-DM, sigla do inglês *Cross-Industry Standard Process for Data Mining*. O uso da metodologia CRISP permite atacar o problema de extração do conhecimento de forma progressiva e organizada, partindo de uma análise de alto nível, que busca a compreensão do negócio, e indo em direção à definição e implantação de modelos que permitam efetivamente atingir os objetivos da mineração. Tal metodologia não se fez necessária, uma vez que mesmo sem sua utilização foi possível identificar uma grande variação dos valores unitários de itens, podendo indicar cartel de fornecedores, má gestão dos recursos e a necessidade de maior transparência.

4.4. Lições aprendidas

Esta seção tem por objetivo enfatizar os problemas encontrados em cada etapa do processo de criação do protótipo para, em seguida, descrever soluções.

O principal problema encontrado durante essa etapa do processo foi encontrar os dados, ou verificar a inexistência deles. Por se tratar de licitações e compras governamentais, era de se esperar que o acesso a esses dados seria fácil para a população, afinal por lei os dados abertos sobre toda e qualquer compra governamental deveriam estar disponibilizados, bem como indicadores deveriam ser disponibilizados para fiscalização por parte da população.

A adoção ou expansão do sistema de compras do Governo Federal, *painel de preços*, para um nível estadual e municipal seria uma solução para os problemas encontrados. Uma divulgação das bases de dados consultadas pelo sistema também ofereceria uma oportunidade para criação de novas ferramentas de estudo, validação e monitoramento pelas compras feitas pelo governo, características que são asseguradas pelas definições de dados abertos visto em 2.1.

Por fim, a solução encontrada foi a retirada dos dados diretamente do diário oficial, que por lei federal, tem o dever de servir de local para divulgação, tanto do anúncio de licitações, quanto para os vencedores que concorreram, apesar dessa solução trazer

consigo outro conjunto de problemas.

Os problemas citados até agora não possuem uma solução simples ou exigem diversas rotas alternativas.

É possível seguir na mesma linha de desenvolvimento do protótipo utilizando *Python* e ferramentas de *web scrapping* e expandi-lo, agrupando os diários oficiais que compartilhem *layouts* semelhantes e utilizar a mesma solução para diversos arquivos. Teriam de ser criadas diversas versões com detalhes específicos para cada *layout*, o que não é uma solução eficiente, mas é factível para recolhimento dos dados e criação de um banco para ser trabalhado. Os protótipos utilizados nesse trabalho podem ser encontrados em repositórios no *Github*[23]: os exemplos de leitores de PDF estão no repositório LeitorPDFDORJ[24]; já o *web scrapper* se encontra no AnaliseDeDadosAbertosRJPy[25] .

Outra solução seria utilizar técnicas de reconhecimento de imagem, juntamente com técnicas de reconhecimento ótico de caracteres, para treinar um robô para conseguir recuperar as informações de forma automatizada, ou recortar os arquivos e produzir outros que sejam mais fáceis de se trabalhar. Contudo isso ainda não iria mudar a falta de padronização existente nos documentos, quebras de tabelas entre colunas, ou os diversos arquivos que são baixados pelo *web scrapper* que não são relevantes para o trabalho e que são provenientes da forma como o sistema do diário oficial foi construído.

É importante ressaltar que já existem ferramentas no mercado que fazem uma solução semelhante, o *ChronoScan* é um pacote completo de digitalização de documentos e entrada de dados, que permite a retirada e preenchimento de campos de documentos com um treinamento utilizando técnicas de *machine learning* e reconhecimento ótico de caracteres. O sistema funciona com configurações iniciais feitas pelo usuário, demonstrando ao programa quais colunas devem ser escaneadas baseadas na posição delas, o que deve ser esperado e como deve ser salvo dentro de um outro arquivo mais fácil de se trabalhar, como o CSV. O sistema, contudo, demanda que os diversos arquivos sigam o mesmo padrão de *layout*, então é eficiente para relatórios que possuem um padrão bem definido, mas não se encaixa para a variabilidade do trabalho ([ChronoScan 2021]).

Grande parte dos problemas encontrados são enraizados no fato de os dados não serem facilmente acessíveis ou não terem sido disponibilizados de modo geral. É evidente a necessidade de alguma alteração na forma como os dados são disponibilizados pelo governo municipal, pois já existem plataformas federais, como no caso do *painel de preços*, que exercem exatamente a função de validação e disponibilização dos dados abertos de compras governamentais.

Existem diversas formas de facilitar o processo de extração de dados e disponibilização dos mesmos. Um deles seria inserir metadados dentro do PDF do diário oficial contendo informações sobre compras licitatórias, o que inclusive facilitaria a indexação dessa informação por mecanismos de busca. Conforme visto na Seção 4.1, por natureza os arquivos já possuem a capacidade de inserir tais metadados sem alteração

[23] `https://github.com/matheusjucaraujo/`
[24] `https://github.com/matheusjucaraujo/LeitorPDFDORJ`
[25] `https://github.com/matheusjucaraujo/AnaliseDeDadosAbertosRJPy`

da informação do arquivo em si. Uma adoção de padronização que permita a criação de um objeto seguindo a Figura 11 seria o suficiente para popular um DW. Os metadados poderiam ser inseridos na primeira página do diário oficial a fim de facilitar o processo de busca de tais dados em vista que o *site* atual do Rio de Janeiro disponibiliza a opção de baixar apenas a página que contenha a palavra chave de pesquisa ou a opção de baixar todo o diário oficial que contém aquela página específica.

Outra opção seria a adoção, por todos os municípios, de um portal centralizador de informação semelhante ao que é realizado ao *painel de preços*. Uma vez homologada a licitação dentro do diário oficial, seria carregada toda a informação de busca relacionada aquela licitação. O governo estadual, onde cada município reside, seria responsável pela fiscalização da inserção dos dados e seria fundamental a disponibilização dos dados abertos em formatos que são mais fáceis de trabalhar.

Por fim, poderia ocorrer um termo de ajuste de conduta governamental para a criação de um modelo padrão de divulgação de compras governamentais, desde que adotado por todas as esferas públicas. Ao definir um padrão de divulgação a criação de ferramentas se torna mais fácil, pois a forma como os dados são esperados seria conhecida e, seria possível, adoção de ferramentas que já existem no mercado, permitindo assim, um ambiente unificador de informação deixando de existir a necessidade de procura por vários endereços nos quais as informações estão espalhadas. A forma como isso ocorreria poderia ser tanto na divulgação das licitações dentro do próprio diário oficial, como a adoção de um anexo no fim do diário, contendo as informações de licitações de forma tabelada e padronizada, que facilitaria o processo de retirada de dados.

5. Conclusão

O objetivo desta pesquisa foi a avaliação dos dados abertos disponibilizados pelos municípios brasileiros relacionados a saúde, através do uso de técnicas de recuperação de informações, apontando problemas e soluções para que esses dados se transformassem em conhecimento para a sociedade.

Com o desenvolver da pesquisa percebeu-se que os dados disponibilizados pelos governos municipais eram escassos, incompletos, separados ou simplesmente não eram disponibilizados. Isso fere as leis definidas para dados abertos e exemplifica a necessidade de adoção de novas estruturas para disponibilização desses dados. Optou-se por diminuir o escopo do trabalho devido a dificuldade de encontrar dados de todos os municípios e o Rio de Janeiro foi escolhido como principal objeto de pesquisa devido a ser o município com a maior quantidade de sistemas relacionados a compras. Contudo a escolha do Rio de Janeiro, devido a sua abundância de sistemas, se mostrou irrelevante, em vista que apenas um dos sistemas disponíveis continha informações vitais ao trabalho e só havia dados a partir de 2019, além de ser um sistema estadual e não ser possível o acesso ao banco de dados diretamente.

Portanto, a solução para encontrar os dados seria a extração dos mesmos pelo meio que são obrigados por lei a serem divulgados - o diário oficial. Foi necessária a criação de um *web scrapper* para automatizar a busca aos diários oficiais que continham informações relevantes ao escopo do projeto e, depois disso, uma forma de extrair a informação de

cada um dos arquivos. Os arquivos estão disponibilizados em PDF o que é uma boa prática para divulgação de informações, contudo não é o modelo ideal para se trabalhar computacionalmente, devido a dificuldade de se acessar o seu conteúdo.

O leitor de PDF se tornou um grande desafio, devido a parte da informação necessária se encontrar em texto plano no diário oficial e parte dele estar contido em tabelas, exigindo a utilização de diversas bibliotecas para gerar um objeto com as informações mínimas para criação de um fato de licitação e que possua dimensões relevantes para análise. Devido ao diário oficial conter diversas outras informações além da licitação e, por não haver uma padronização na forma como esses dados são divulgados, tornou-se necessário estudar cada caso do arquivo para verificar a necessidade de alterações do sistema para criação de objetos.

Apesar da solução criada ser factível, ela não se mostrou eficiente e nem ser possível a criação de um leitor de PDF genérico o suficiente para comportar qualquer tipo de arquivo, portanto foram sugeridas outras soluções, além do protótipo disponível, que poderiam ser adotadas para facilitar a criação do DW. As soluções variam desde a criação de outros protótipos utilizando tecnologias diferentes, com técnicas de *machine learning* e reconhecimento ótico de caracteres, criação de um sistema centralizador de informações de compras governamentais; posturas novas que devem ser adotadas pelo próprio governo para cumprimento das leis de divulgação de dados abertos; adoção de um padrão de divulgação que deveria ser seguido por todas esferas públicas, a fim de facilitar o processo de retirada de informações dos arquivos.

Até que alguma dessas alternativas seja adotada, a criação de um *data warehouse*, extraindo dados diretamente dos arquivos em PDF disponíveis pelo diário oficial, é um trabalho hercúleo e ineficiente. A falta de outros arquivos, em formatos mais fáceis de se trabalhar, que contenham as informações necessárias não permitiu a exploração de soluções mais eficientes para extração de dados.

Considerando o foco do trabalho, a disponibilidade dos dados abertos dos municípios brasileiros relacionados à saúde é falha, incompleta e insuficiente, tornando seu uso inviável. Essa disponibilização precisa ser padronizada e publicizada.

Referências

[Aquarela 2017] Aquarela (2017). Otimizando agendamentos médicos com inteligência artificial: Case. Vitória-es. *https://www.aquare.la/otimizando-agendamentos-medicos-com-inteligencia-artificial/*, acessado em 01/11/2019.

[Araújo 1995] Araújo, V. M. R. H. d. (1995). Sistemas de informação: nova abordagem teórico-conceitual. *Ciência da Informação*, 24(1).

[Broucke and Baesens 2018] Broucke, S. and Baesens, B. (2018). *Pratical Web Scraping for Data Science*. Apress, EUA.

[Camilo and da Silva 2009] Camilo, C. O. and da Silva, J. C. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas.

[ChronoScan 2021] ChronoScan (2021). About Chronoscan. *https://chronoscan.org/*, acessado em 10/03/2021.

[da República 1993] da República, P. (1993). Lei nº 8.666, de 21 de junho de 1993. .

[de Cultura Digital et al. 2011] de Cultura Digital, L. B., de Informação e Coordenação do Ponto BR, N., and W3C (2011). *Manual dos Dados Abertos*. Governo brasileiro.

[de Dados Abertos 2019] de Dados Abertos, P. B. (2019). Política de dados abertos do poder executivo federal. *http://wiki.dados.gov.br/Politica-de-Dados-Abertos.ashx*, acessado em 10/09/2019.

[Diniz 2013] Diniz, V. (2013). Como conseguir dados governamentais abertos. In *III Congresso Consad de Gestão Pública*, page 19.

[Eaves 2009] Eaves, D. (2009). The three laws of open government data. *https://eaves.ca/2009/09/30/three-law-of-open-government-data/*, acessado em 05/10/2019.

[Forum 2019] Forum, W. E. (2019). The Global Competitiveness Report 2019. *https://www3.weforum.org/docs/WEF_TheGlobalCompetitivenessReport2019.pdf*, acessado em 21/10/2021.

[Gomes 2015] Gomes, T. C. S. (2015). Descoberta de conhecimento utilizando mineração de dados educacionais abertos. *Universidade Federal Rural de Pernambuco Departamento de Estatística e Informática*.

[Hardy 2017] Hardy, M. (2017). The application/pdf media type. *https://tools.ietf.org/html/rfc8118*, acessado em 28/02/2021.

[Hope 2020] Hope, C. (2020). Regex. *https://www.computerhope.com/jargon/r/regex.htm*, acessado em 28/02/2021.

[IBGE 2017] IBGE, A. (2017). Pib avança 1,0% em 2017 e fecha ano em r$ 6,6 trilhões. *https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/20166-pib-avanca-1-0-em-2017-e-fecha-ano-em-r-6-6-trilhoes*, acessado em 01/11/2019.

[Infraestrutura Nacional de Dados Abertos 2019] Infraestrutura Nacional de Dados Abertos (2019). Sobre o dados.gov.br. *http://dados.gov.br/pagina/sobre*, acessado em 10/09/2019.

[Inmon 1996] Inmon, W. H. (1996). *Building the Data Warehouse*. Wiley.

[Internacional 2021] Internacional, T. (2021). Índice de percepção da corrupção 2020. *https://comunidade.transparenciainternacional.org.br/ipc-indice-de-percepcao-da-corrupcao-2020*, acessado em 21/10/2021.

[Laboissière 2018] Laboissière, P. (2018). Quase 90% dos brasileiros consideram saúde péssima, ruim ou regular. *http://agenciabrasil.ebc.com.br/saude/noticia/2018-06/para-89-dos-brasileiros-saude-e-considerada-pessima-ruim-ou-regular*, acessado em 20/10/2019.

[Lima 2019] Lima, W. C. (2019). Dados abertos governamentais no contexto da ciência cidadã: O caso da operação serenata de amor.

[Paim 2003] Paim, F. R. S. (2003). Uma metodologia para definição de requisitos em sistemas data warehouse. page 198.

[Pontes 2019] Pontes, N. (2019). Auditores fiscais denunciam retrocesso no combate à corrupção. *https://www.dw.com/pt-br/auditores-fiscais-denunciam-retrocesso-no-combate-%C3%A0-corrup%C3%A7%C3%A3o/a-50816153*, acessado em 20/10/2019.

[Python Software Foundation 2021] Python Software Foundation (2021). About Python. *https://www.python.org/about/*, acessado em 31/01/2021.

[Reitz 2011] Reitz, K. (2011). Requests web site. *https://requests.readthedocs.io/en/master/*, acessado em 01/09/2020.

[Selenium 2018] Selenium (2018). Selenium web site. *https://www.selenium.dev/*, acessado em 01/09/2020.

[Shukla and Roy 2016] Shukla, V. and Roy, D. (2016). Web crawlers and web crawling algorithms – a review. *Computing Canada*, 2(2):259.

[Tribunal de Contas da União 2015] Tribunal de Contas da União (2015). *Cinco Motivos para a Abertura de Dados na Administração Pública*. Governo brasileiro. *https://portal.tcu.gov.br/biblioteca-digital/cinco-motivos-para-a-abertura-de-dados-na-administracao-publica.htm*, acessado em 22/12/2023.

[Zacarias et al. 2019] Zacarias, R. O., de Matos Abreu, L., da Hora, H. R. M., and de Vasconcelos, A. P. V. (2019). Comportamento da nota de corte do prouni e a sua relação com a política de acesso ao ensino superior: um estudo com mineração de dados. *Educação Online*, 14(31):62–81.