UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Engenharia Elétrica e de Computação

IVÁN JOSÉ MESTRE FERNÁNDEZ

L5P – UMA PLATAFORMA DE E-GOV PARA A COLETA E ANÁLISE DE DADOS PÚBLICOS DA CÂMARA E DO SENADO FEDERAL DO BRASIL E QUESTÕES LEVANTADAS PELA LGPD

CAMPINAS

2020

IVÁN JOSÉ MESTRE FERNÁNDEZ

L5P – UMA PLATAFORMA DE E-GOV PARA A COLETA E ANÁLISE DE DADOS PÚBLICOS DA CÂMARA E DO SENADO FEDERAL DO BRASIL E QUESTÕES LEVANTADAS PELA LGPD

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na área de Telecomunicações e Telemática.

Supervisor/Orientador: Leonardo de Souza Mendes

Este trabalho corresponde à versão final da dissertação defendida pelo aluno Iván José Mestre Fernández, orientada pelo Prof. Dr. Leonardo de Souza Mendes.

Assinatura do Orientador

CAMPINAS

2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Mestre Fernández, Iván José, 1988-
M464L L5P - Uma plataforma de e-Gov para a coleta e análise de dados públicos da Câmara e do Senado Federal do Brasil e questões levantadas pela LGPD / Iván José Mestre Fernández. – Campinas, SP : [s.n.], 2020.

Orientador: Leonardo de Souza Mendes.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Governo eletrônico. 2. Mineração de dados (Computação). 3. Tecnologia da informação e comunicação. 4. Gestão da informação. 5. Transparência na administração publica. I. Mendes, Leonardo de Souza, 1961-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: L5P - An e-Gov platform for the collection and analysis of public data from the Chamber and Federal Senate of Brazil and issues raised by the LGPD
Palavras-chave em inglês:
Electronic government
Information and communication technology
Information management
Transparency in public administration
Área de concentração: Telecomunicações e Telemática
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Leonardo de Souza Mendes [Orientador]
André Marcelo Panhan
Gean Davis Breda
Data de defesa: 27-11-2020
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-7575-3379
- Currículo Lattes do autor: http://lattes.cnpq.br/1959393144717793

COMISSÃO JULGADORA – DISSERTAÇÃO DE MESTRADO

Candidato: Iván José Mestre Fernández RA: 209446
Data da defesa: 27 de novembro de 2020
Título da Tese: “L5P – Uma Plataforma de e-Gov para a Coleta e Análise de Dados Públicos da Câmara e do Senado Federal do Brasil e Questões Levantadas pela LGPD”

Prof. Dr. Leonardo de Souza Mendes (Presidente)
Dr. André Marcelo Panhan
Dr. Gean Davis Breda

A Ata de Defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Aos meus pais. Seu amor, ensinamentos e sacrifícios fizeram de mim a pessoa que sou hoje.

A minha namorada Liz, por estar ao meu lado há tantos anos, me apoiando e acreditando em mim.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Agradeço,

Ao meus pais pelo apoio e amor incondicional

A minha namorada Liz pelo apoio, amor, suporte e companhia

A minha família cubana e brasileira pelo suporte e confiança

Ao meu professor e orientador Leonardo de Souza Mendes pelo apoio e acompanhamento durante a realização deste trabalho.

Aos professores da FEEC (Faculdade de Engenharia Elétrica e Computação) pelos conhecimentos adquiridos durante as disciplinas

A Universidade Estadual de Campinas pelo suporte

A todos os que de alguma forma ajudaram e colaboraram com a realização deste trabalho

RESUMO

A disseminação da corrupção no mundo é considerada um dos maiores problemas globais, que afetou quase todos os países do mundo, desenvolvidos ou subdesenvolvidos, ao longo da história. O caso específico da corrupção institucionalizada, normalmente associada à estrutura política de um país, pode ser muito difícil de expor e combater. No caso do Brasil, nos últimos anos, grandes escândalos de corrupção surgiram que abalaram completamente o ambiente político nacional. Durante séculos, a sociedade ficou indefesa diante de tais problemas; no entanto, amparado por legislação recente sobre conformidade e transparência governamental, o uso de Mineração de Dados e Análise de Dados pode oferecer ferramentas importantes para a sociedade no combate a essas práticas antigas e condenadas de corrupção. Por tudo isso, é necessário o desenvolvimento de ferramentas tecnológicas que permitam aos cidadãos obter conhecimento sobre os membros e o funcionamento do governo. Este trabalho apresenta a implementação de um sistema de recuperação de informação, utilizando técnicas de Mineração de Dados na Web, para coleta de informações de interesse público de políticos e processos políticos no Brasil, a partir de informações públicas e de acesso aberto das casas legislativas federais brasileiras; bem como uma análise da Lei Geral de Proteção de Dados Pessoais, a qual foi aprovada durante o desenvolvimento do sistema e regula o tratamento de dados pessoais de pessoas físicas.

Palavras-chave: Mineração Web, Mineração de Conteúdo, Mineração de Dados, Política.

ABSTRACT

The overspread of corruption in the world is considered one of the major global problems, one that has affected almost every country in the world, developed or underdeveloped, throughout history. The specific case of institutionalized corruption, normally associated with the political structure of a country, can be exceedingly difficult to expose and fight. In the case of Brazil, in recent years, major corruption scandals have arisen and completely shaken the national political environment. For centuries society was left defenseless in the face of such problems; however, supported by recent legislation on government compliance and transparency, the use of Data Mining and Data Analytics can offer important tools for society in the combat of these old and condemned practices of corruption. This work presents the implementation of an information retrieval system using Web Data Mining techniques, to collect information of public interest from politicians and political processes in Brazil, using public and open access information from the Brazilian federal legislative houses; as well as an in-depth analysis of the Brazilian General Personal Data Protection Law, which was approved during the development of the system and regulates the processing of personal data of individuals.

Key Words: Web Mining, Content Mining, Data Mining, Politics.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ACID – Atomicity, Consistency, Isolation, Durability – Atomicidade, Consistência, Isolamento, Durabilidade

ACL – Access Control Lists – Listas de Controle de Acesso

AJAX – Asynchronous JavaScript and XML – JavaScript e XML Assíncrono

ANPD – Autoridade Nacional de Proteção de Dados

API – Application Programming Interface – Interface de Programação de Aplicativo

CCPA – California Consumer Privacy Act – Lei de Privacidade do Consumidor da Califórnia

CEPESP – Centro de Política e Economia do Setor Público

CMS – Content Management System – Sistema de Gerenciamento de Conteúdo

CNPJ – Cadastro Nacional da Pessoa Jurídica

CORS – Cross-Origin Resource Sharing – Compartilhamento de Recursos Entre Origens

CPF – Cadastro de Pessoas Físicas

CRISP-DM – Cross-Industry Standard Process for Data Mining – Processo Padrão Entre Indústrias para Mineração de Dados

CRUD – Create, Read, Update, Delete – Criar, Ler, Atualizar, Excluir

CSRF – Cross-Site Request Forgery – Falsificação de Solicitação Entre Sites

CSV – Comma Separated Values – Valores Separados Por Virgula

DB – Database – Banco de Dados

DM – Data Mining – Mineração de Dados

DMM – Data Mining Model – Modelo de Mineração de Dados

EDA – Exploratory Data Analysis – Análise Exploratória de Dados

FGV – Fundação Getúlio Vargas

GDPR – General Data Protection Regulation – Regulamento Geral de Proteção de Dados

HTML – Hypertext Markup Language – Linguagem de Marcação de Hipertexto

HTTP – Hypertext Transfer Protocol – Protocolo de Transferência de Hipertexto

HTTPS – HyperText Transfer Protocol Secure – Protocolo Seguro de Transferência de Hipertexto

ISO – International Organization for Standardization – Organização Internacional para Padronização

JSON – JavaScript Object Notation – Notação de Objetos JavaScript

KDD – Knowledge Discovery in Databases – Descoberta de Conhecimento em Bancos de Dados

LAMP – Linux, Apache, MySQL, PHP

LGPD – Lei Geral de Proteção de Dados Pessoais

MVC – Model-View-Controller – Modelo-Visão-Controle

OM – Opinion Mining – Mineração de Opiniões

ORM – Object-Relational Mapping – Mapeamento Objeto-Relacional

PDF – Portable Document Format – Formato de Documento Portátil

PHP – PHP: Hypertext Preprocessor – PHP: Pré-processador de Hipertexto

PT – Partido dos Trabalhadores

REST – Representational State Transfer – Transferência de Estado Representacional

SA – Sentiment Analysis – Análise de Sentimentos

SQL – Structured Query Language – Linguagem Estruturado de Consulta

TM – Text Mining – Mineração de Textos

URL – Uniform Resource Locator – Localizador Uniforme de Recursos

WCM – Web Content Mining – Mineração de Conteúdo da Web

WDM – Web Data Mining – Mineração de Dados na Web

XML – Extensible Markup Language – Linguagem de Marcação Extensível

XSS – Cross-Site Scripting – Scripting Entre Sites

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

A corrupção política é um problema global que afetou quase todos os países do mundo, desenvolvidos ou subdesenvolvidos, ao longo da história. Nos últimos tempos, devido à globalização dos meios de comunicação e ao surgimento das redes sociais, os escândalos da corrupção política em várias partes do mundo tornaram-se eventos da vida cotidiana.

No caso específico do Brasil, nos últimos anos, surgiram grandes escândalos de corrupção que abalaram o ambiente político nacional. Esses escândalos deram origem a investigações gigantescas que cobriam todos os níveis da política brasileira. Exemplos desses escândalos são o Escândalo do Mensalão e a Operação Lava Jato. A corrupção tem sido mencionada como um dos fatores que provocaram os protestos ocorridos em 2013 em todo o país (CNN, 2013).

O Escândalo do Mensalão foi o nome dado ao processo judicial que investigou a compra ilegal de votos no Congresso durante o primeiro mandato do ex-presidente Luiz Inácio Lula da Silva. O escândalo surgiu em 2005 quando um congressista acusou publicamente o Partido dos Trabalhadores (PT) de ter pagado R$ 30.000 por mês aos aliados políticos desde 2003 para garantir que votaram a favor de leis e emendas propostas pelo PT. Foi considerado no momento o maior caso de corrupção na história recente do Brasil e estava prestes a colapsar o governo de Lula (MICHENER; PEREIRA, 2016).

A Operação Lava Jato é uma investigação criminal realizada pela Polícia Federal do Brasil e presidida judicialmente pelo juiz Sérgio Moro desde 17 de março de 2014 e em 2019 pelo Juiz Luiz Antônio Bonat. Seu objetivo era investigar um esquema de lavagem de dinheiro e subornos suspeitado de movimentar mais de R$ 10 bilhões e abrangendo pelo menos 10 países em América Latina. É considerado pela Polícia Federal como a maior investigação de corrupção na história do Brasil e um dos maiores escândalos de corrupção da história (BLOOMBERG, 2017) (DW, 2016).

De acordo com o Índice de Percepção de Corrupção, elaborado anualmente pela Organização Não Governamental Alemã Transparency International, cujo objetivo é combater a corrupção global e prevenir atividades criminosas que são produto de corrupção, o Brasil ocupou em 2019 a posição 106 junto com varios outros países como Albânia, Costa do Marfim e Egito (Fig. 1) (TRANSPARENCY INTERNATIONAL, 2019).

39	Serbia	91	34	Kazakhstan	113	28	Dominican Republic	137	24	Zimbabwe	158
39	Turkey	91	34	Nepal	113	28	Kenya	137	23	Eritrea	160
38	Ecuador	93	34	Philippines	113	28	Lebanon	137	22	Nicaragua	161
38	Sri Lanka	93	34	Eswatini	113	28	Liberia	137	20	Cambodia	162
38	Timor-Leste	93	34	Zambia	113	28	Mauritania	137	20	Chad	162
37	Colombia	96	33	Sierra Leone	119	28	Papua New Guinea	137	20	Iraq	162
37	Ethiopia	96	32	Moldova	120	28	Paraguay	137	19	Burundi	165
37	Gambia	96	32	Niger	120	28	Russia	137	19	Congo	165
37	Tanzania	96	32	Pakistan	120	28	Uganda	137	19	Turkmenistan	165
37	Vietnam	96	31	Bolivia	123	26	Angola	146	18	Democratic Republic of the Congo	168
36	Bosnia and Herzegovina	101	31	Gabon	123	26	Bangladesh	146	18	Guinea Bissau	168
36	Kosovo	101	31	Malawi	123	26	Guatemala	146	18	Haiti	168
36	Panama	101	30	Azerbaijan	126	26	Honduras	146	18	Libya	168
36	Peru	101	30	Djibouti	126	26	Iran	146	17	Korea, North	172
36	Thailand	101	30	Kyrgyzstan	126	26	Mozambique	146	16	Afghanistan	173
35	Albania	106	30	Ukraine	126	26	Nigeria	146	16	Equatorial Guinea	173
35	Algeria	106	29	Guinea	130	25	Cameroon	153	16	Sudan	173
35	Brazil	106	29	Laos	130	25	Central African Republic	153	16	Venezuela	173
35	Cote d'Ivoire	106	29	Maldives	130	25	Comoros	153	15	Yemen	177
35	Egypt	106	29	Mali	130	25	Tajikistan	153	13	Syria	178
35	North Macedonia	106	29	Mexico	130	25	Uzbekistan	153	12	South Sudan	179
35	Mongolia	106	29	Myanmar	130	24	Madagascar	158	9	Somalia	180
34	El Salvador	113	29	Togo	130						

Figura 1 Índice de Percepção de Corrupção 2019
Fonte: (TRANSPARENCY INTERNATIONAL, 2019)

Por tudo isso, é necessário ter ferramentas tecnológicas que permitam aos cidadãos obter conhecimento sobre os membros e o funcionamento do governo. É necessário contar com maneiras de rastrear o dinheiro das doações para as campanhas eleitorais e as relações que essas doações têm com a ação subsequente dos políticos. Além disso, as ferramentas devem ajudar as pessoas a exercer seu direito e dever como cidadãos para votar em seus representantes, mas também para acompanhar os atos desses representantes posteriormente, a fim de garantir o cumprimento de suas promessas de campanha.

Para isso, foi decidido implementar um sistema de coleta de informações, disponíveis livremente na web, sobre os políticos. Este sistema usa técnicas de Mineração de Dados na Web para encontrar as informações, essas informações são consolidadas em um único repositório para análise posterior. Durante o

desenvolvimento do sistema, foi aprovada a Lei Geral de Proteção de Dados Pessoais (LGPD), que regula o tratamento de dados pessoais de pessoas físicas. Esta lei impõe restrições ao processamento de dados pessoais, a fim de proteger os direitos que as pessoas físicas possuem sobre eles. Neste trabalho, realizamos uma análise da LGPD, identificando o impacto desta nova lei no projeto e as medidas a serem tomadas para garantir a conformidade com a lei, descrevemos a base teórica sobre a qual o sistema foi implementado, detalhamos o processo de desenvolvimento do sistema de recuperação de informações L5P e apresentamos o sistema e suas principais funcionalidades.

1.2 Objetivos

O presente trabalho tem os seguintes objetivos gerais:

- Desenvolvimento de um sistema de recuperação de informações públicas sobre os políticos brasileiros, especificamente os membros da Câmara dos Deputados e do Senado.
- Identificar o impacto da LGPD no desenvolvimento do referido sistema, bem como possíveis soluções para o cumprimento da lei.

Esses objetivos podem ser divididos nos seguintes objetivos específicos:

- Identificar as fontes de informação necessárias para coletar informações dos políticos
- Implementar técnicas de coleta de informações
- Realizar a limpeza e consolidação dos dados coletados
- Implementar medidas de segurança para a proteção do acesso aos dados
- Implementar a visualização dos dados consolidados
- Justificar a viabilidade do projeto no contexto da LGPD.
- Identificar as modificações que precisam ser feitas ao projeto para cumprir com as regulações da LGPD.

1.3 Organização

No Capítulo 2 deste trabalho é realizada uma análise profunda da LGPD e o impacto desta lei no desenvolvimento do projeto.

No Capítulo 3 é abordada a fundamentação teórica sobre a qual se baseia o desenvolvimento do sistema de recuperação de informação L5P. São analisados os processos de Mineração de Dados e Mineração de Dados na Web. Finalmente, é descrita a Mineração de Conteúdo da Web a qual é a área mais relevante da Mineração de Dados na Web para o desenvolvimento do sistema.

O Capítulo 4 descreve o processo de implementação do sistema de recuperação de informações L5P.

No Capítulo 5 é realizada uma apresentação do sistema de recuperação de informações L5P e suas principais funcionalidades.

O Capítulo 6 apresenta as conclusões do trabalho, e trabalhos futuros a serem considerados.

2 LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS

Durante o desenvolvimento deste projeto, a **Lei nº 13.709/2018**, conhecida publicamente como **Lei Geral de Proteção de Dados Pessoais** (LGPDP ou LGPD), foi aprovada e ratificada, o que tem um grande impacto e afeta diretamente a realização deste projeto. A LGPD, ratificada em 14 de agosto de 2018 e com entrada em vigor em agosto de 2020, é uma lei que tenta unificar mais de 40 estatutos diferentes que atualmente regulam o tratamento de dados pessoais no Brasil (GDPR.EU, 2020) e tem como objetivo a *"proteção dos direitos fundamentais da liberdade e da privacidade e o livre desenvolvimento da personalidade das pessoas singulares"* (BRASIL, 2018). A LGPD é aplicável ao processamento (tratamento, coleta ou uso dos dados para fornecer bens ou serviços) de **dados pessoais** de pessoas localizadas no território do Brasil por qualquer empresa ou organização, independentemente de sua localização (GDPR.EU, 2020) (BRASIL, 2018). Esta lei tem grandes semelhanças com outras leis de proteção de dados pessoais criadas em outros países; alguns exemplos são: a **General Data Protection Regulation** (GDPR), uma lei que entrou em vigor nos territórios da União Europeia em 25 de maio de 2018; e a **California Consumer Privacy Act of 2018** (CCPA), uma lei aprovada em 28 de junho de 2018 no estado da Califórnia dos Estados Unidos de América. Devido ao impacto que a entrada em vigor da LGPD representa para este projeto, decidiu-se realizar uma análise profunda desta Lei e quais artigos são relevantes para o projeto; uma análise de como a lei poderia afetar o projeto quando ela se tornar efetiva, usando o GDPR como referência; e uma descrição detalhada das mudanças e modificações que devem ser feitas para que o projeto esteja em conformidade com a LGPD. Esta análise é apresentada neste capítulo.

2.1 O que é a LGPD?

A **Lei Geral de Proteção de Dados Pessoais** (**Lei nº 13.709/2018**) (LGPDP ou LGPD) é uma legislação que tem como objetivo principal a criação de uma estrutura legal para a regulamentação de atividades de processamento de dados pessoais realizadas no território brasileiro ou relacionadas a pessoas físicas

localizadas neste território, independentemente da localização da entidade que realiza o processamento. Dentro do texto da LGPD, grandes semelhanças com a **General Data Protection Regulation** (GDPR) podem ser observadas. Essas semelhanças serão abordadas nas próximas seções deste capítulo e são as bases utilizadas para a análise prática realizada nos possíveis cenários em que a LGPD afetaria diretamente o projeto atual. A LGPD foi aprovada pelo Congresso Nacional em 14 de agosto de 2018 e sancionada conclusivamente pelo presidente Jair Bolsonaro em julho de 2019. Entrou em vigor em 15 de agosto de 2020, seis meses após a data inicialmente prevista de fevereiro de 2020. Esta lei visa substituir os mais de 40 regulamentos legais que atualmente dispõem sobre a proteção e a privacidade dos dados, incluindo o **Marco Civil da Internet** e o **Código de Proteção ao Consumidor**. A LGPD define a criação da **Autoridade Nacional de Proteção de Dados** (ANPD), um órgão de supervisão cujos objetivos serão criar normas, estabelecer padrões técnicos, supervisionar, auditar, educar sobre a lei, receber notificações de violações de segurança dos dados e aplicar as sanções previstas na lei. A atividade regulatória desta lei é baseada nos seguintes valores (BRASIL, 2018):

I. *o respeito à privacidade;*
II. *a autodeterminação informativa;*
III. *a liberdade de expressão, de informação, de comunicação e de opinião;*
IV. *a inviolabilidade da intimidade, da honra e da imagem;*
V. *o desenvolvimento econômico e tecnológico e a inovação;*
VI. *a livre iniciativa, a livre concorrência e a defesa do consumidor; e*
VII. *os direitos humanos, o livre desenvolvimento da personalidade, a dignidade e o exercício da cidadania pelas pessoas naturais.*

A LGPD dispõe *"sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural"* (BRASIL, 2018). Isso significa que possui uma aplicação transversal e multissetorial, uma vez que seus regulamentos se aplicam aos setores público e privado, mas também aos dados

físicos, digitais, online e offline. A LGPD também possui aplicação extraterritorial, de acordo com o Art. 3° *"Esta Lei aplica-se a qualquer operação de tratamento realizada por pessoa natural ou por pessoa jurídica de direito público ou privado, independentemente do meio, do país de sua sede ou do país onde estejam localizados os dados"* se atenderem às seguintes condições (BRASIL, 2018):

I. *a operação de tratamento seja realizada no território nacional;*
II. *a atividade de tratamento tenha por objetivo a oferta ou o fornecimento de bens ou serviços ou o tratamento de dados de indivíduos localizados no território nacional; ou*
III. *os dados pessoais objeto do tratamento tenham sido coletados no território nacional.*

Isso significa que qualquer pessoa física, independentemente de ser ou não um cidadão brasileiro, cujos dados foram coletados ou processados no território brasileiro, é protegida pela LGPD. Existem vários cenários para o tratamento de dados pessoais aos quais a LGPD não se aplica, estes são coletados no Art. 4º e são os seguintes (BRASIL, 2018):

I. *realizado por pessoa natural para fins exclusivamente particulares e não econômicos;*
II. *realizado para fins exclusivamente:*
    a) *jornalístico e artísticos; ou*
    b) *acadêmicos, aplicando-se a esta hipótese os arts. 7º e 11 desta Lei;*
III. *realizado para fins exclusivos de:*
    a) *segurança pública;*
    b) *defesa nacional;*
    c) *segurança do Estado; ou*
    d) *atividades de investigação e repressão de infrações penais; ou*
IV. *provenientes de fora do território nacional e que não sejam objeto de comunicação, uso compartilhado de dados com agentes de tratamento brasileiros ou objeto de transferência internacional de dados com outro país que não o de proveniência, desde que o país de proveniência proporcione grau de proteção de dados pessoais adequado ao previsto nesta Lei.*

De acordo com o Art. 6º da LGPD, todas as atividades de tratamento de dados pessoais *"deverão observar a boa-fé e os seguintes princípios"* (BRASIL, 2018):

I. *finalidade: realização do tratamento para propósitos legítimos, específicos, explícitos e informados ao titular, sem possibilidade de tratamento posterior de forma incompatível com essas finalidades;*
II. *adequação: compatibilidade do tratamento com as finalidades informadas ao titular, de acordo com o contexto do tratamento;*
III. *necessidade: limitação do tratamento ao mínimo necessário para a realização de suas finalidades, com abrangência dos dados pertinentes, proporcionais e não excessivos em relação às finalidades do tratamento de dados;*
IV. *livre acesso: garantia, aos titulares, de consulta facilitada e gratuita sobre a forma e a duração do tratamento, bem como sobre a integralidade de seus dados pessoais;*
V. *qualidade dos dados: garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados, de acordo com a necessidade e para o cumprimento da finalidade de seu tratamento;*
VI. *transparência: garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial;*
VII. *segurança: utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão;*
VIII. *prevenção: adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais;*
IX. *não discriminação: impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos;*
X. *responsabilização e prestação de contas: demonstração, pelo agente, da adoção de medidas eficazes e capazes de comprovar a observância*

> *e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas.*

A LGPD, especifica os **9 direitos fundamentais** que os indivíduos têm sobre seus dados pessoais, define exatamente o que se entende por **dados pessoais** e cria **10 bases legais** que as entidades que realizam o processamento podem usar para o tratamento legal de dados.

2.1.1 Direitos dos titulares

No Art. 5º, a LGPD define como **titular** a *"pessoa natural a quem se referem os dados pessoais que são objeto de tratamento"*. Essa lei também estabelece, em seu Art. 17º, que *"Toda pessoa natural tem assegurada a titularidade de seus dados pessoais e garantidos os direitos fundamentais de liberdade, de intimidade e de privacidade"* e determina, no Art. 18º, **9 direitos fundamentais** que as pessoas físicas possuem sobre seus dados pessoais, diretamente relacionados aos princípios mencionados acima. Esses direitos são (BRASIL, 2018):

I. *confirmação da existência de tratamento;*
II. *acesso aos dados;*
III. *correção de dados incompletos, inexatos ou desatualizados;*
IV. *anonimização, bloqueio ou eliminação de dados desnecessários, excessivos ou tratados em desconformidade com o disposto nesta Lei;*
V. *portabilidade dos dados a outro fornecedor de serviço ou produto, mediante requisição expressa, de acordo com a regulamentação da autoridade nacional, observados os segredos comercial e industrial;*
VI. *eliminação dos dados pessoais tratados com o consentimento do titular, exceto nas hipóteses previstas no art. 16 desta Lei;*
VII. *informação das entidades públicas e privadas com as quais o controlador realizou uso compartilhado de dados;*
VIII. *informação sobre a possibilidade de não fornecer consentimento e sobre as consequências da negativa;*
IX. *revogação do consentimento, nos termos do § 5º do art. 8º desta Lei.*

O Art. 9º também define o direito dos titulares de serem informados sobre o tratamento dos seus dados pessoais.

2.1.2 Dados pessoais

A LGPD fornece uma definição bastante ampla de **dados pessoais**. Nos termos do seu Art. 5º, os **dados pessoais** podem ser definidos como *"informação relacionada a pessoa natural identificada ou identificável"*, essa definição é bastante vaga e pode se referir a qualquer tipo de informação que, por si só ou combinada com outros tipos de informação, possa ser usada para identificar uma pessoa natural. Espera-se que uma definição mais específica seja criada pela ANPD assim que a agência for constituída. A lei também oferece uma definição mais específica de **dado pessoal sensível**, o qual pode ser definido como *"dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural"*. Outra definição importante sobre **dados pessoais**, presente no Artigo 5º, é a de **dados anonimizados**, definido como *"dado relativo a titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento"*. Com relação aos **dados anonimizados**, o Art. 12º declara que *"Os dados anonimizados não serão considerados dados pessoais para os fins desta Lei, salvo quando o processo de anonimização ao qual foram submetidos for revertido, utilizando exclusivamente meios próprios, ou quando, com esforços razoáveis, puder ser revertido"* (BRASIL, 2018).

2.1.3 Bases legais para o processamento de dados

As entidades que realizam o processamento de dados pessoais devem justificar a legalidade desse processo, usando uma das **10 bases legais** encontradas no Art. 7º da LGPD. Essas **bases legais** são (BRASIL, 2018):

I. *mediante o fornecimento de consentimento pelo titular;*
II. *para o cumprimento de obrigação legal ou regulatória pelo controlador;*
III. *pela administração pública, para o tratamento e uso compartilhado de dados necessários à execução de políticas públicas previstas em leis e regulamentos ou respaldadas em contratos, convênios ou instrumentos congêneres, observadas as disposições do Capítulo IV desta Lei;*
IV. *para a realização de estudos por órgão de pesquisa, garantida, sempre que possível, a anonimização dos dados pessoais;*

V. *quando necessário para a execução de contrato ou de procedimentos preliminares relacionados a contrato do qual seja parte o titular, a pedido do titular dos dados;*

VI. *para o exercício regular de direitos em processo judicial, administrativo ou arbitral, esse último nos termos da Lei nº 9.307, de 23 de setembro de 1996 (Lei de Arbitragem) ;*

VII. *para a proteção da vida ou da incolumidade física do titular ou de terceiro;*

VIII. *para a tutela da saúde, exclusivamente, em procedimento realizado por profissionais de saúde, serviços de saúde ou autoridade sanitária;*

IX. *quando necessário para atender aos interesses legítimos do controlador ou de terceiro, exceto no caso de prevalecerem direitos e liberdades fundamentais do titular que exijam a proteção dos dados pessoais; ou*

X. *para a proteção do crédito, inclusive quanto ao disposto na legislação pertinente.*

Essas **bases legais** são as únicas hipóteses sob as quais o processamento de dados pessoais pode ser realizado.

A LGPD também define a figura do **Responsável pelo Tratamento de Dados Pessoais**. O Responsável será nomeado pela entidade que realiza o processamento e é responsável por verificar se a entidade está em conformidade com os regulamentos presentes nesta lei, aceitando reclamações dos titulares e mantendo comunicação com a ANPD. Se houver violação da lei, a LGPD define sanções administrativas impostas pela ANPD, dentre as quais estão *"multa simples, de até 2% (dois por cento) do faturamento da pessoa jurídica de direito privado, grupo ou conglomerado no Brasil no seu último exercício, excluídos os tributos, limitada, no total, a R$ 50.000.000,00 (cinquenta milhões de reais) por infração"* e *"multa diária, observado o limite total a que se refere o inciso II"* (BRASIL, 2018).

2.2 Impacto da LGPD no projeto

Para analisar corretamente o impacto que a implementação da LGPD tem no projeto, é necessário primeiro analisar se a lei é aplicável a ele. Para esse fim, precisamos definir os principais objetivos do projeto, identificar e limitar possíveis cenários da implementação e analisar se esses cenários são contemplados no texto da LGPD.

Considerando o exposto anteriormente neste documento, podemos definir os principais objetivos do projeto como:

1. Informar os cidadãos sobre a operação do governo e as ações de seus membros
2. Auxiliar os cidadãos no processo de tomada de decisão, descobrindo informações úteis e acionáveis através do uso de análises computacionais

Tendo definido esses dois objetivos principais, é possível proceder à identificação de possíveis cenários de implementação do projeto. Esses cenários podem ser definidos da seguinte maneira:

1. **Jornalístico**: Criação de um site puramente jornalístico, com o objetivo de informar a população
2. **Acadêmico**: Uso de dados para pesquisas acadêmicas, produzindo estudos, trabalhos ou artigos que permitam uma melhor compreensão das ações dos membros do governo
3. **Comercial**: Criação de um site com o objetivo de informar a população, permitindo o acesso a ferramentas de análise para usuários utilizando seus perfis e fornecendo serviços de análise de dados a terceiros. Incluir mecanismos de monetização para financiar a atividade de pesquisa e desenvolvimento do projeto.

Uma vez identificados os possíveis cenários, é possível analisar se estes são contemplados pela LGPD, se a lei se aplica a eles e se eles são legais no âmbito da LGPD. Em caso afirmativo, é preciso determinar quais medidas ou mudanças, administrativas e técnicas, devem ser tomadas para que o projeto esteja em conformidade com o texto da Lei.

2.2.1 Jornalístico

Sobre o primeiro cenário, o **jornalístico**, a LGPD afirma em seu Art. 4º (BRASIL, 2018):

Art. 4º Esta Lei não se aplica ao tratamento de dados pessoais:

II. realizado para fins exclusivamente:

a) *jornalístico e artísticos;*

Isso significa que, neste primeiro cenário, a LGPD não se aplica, portanto, é possível realizar o processamento de dados pessoais com um objetivo puramente **jornalístico**, sem ser governado por os regulamentos da LGPD. Observe que a lei determina que o processamento deve ser realizado para *"fins exclusivamente: jornalístico"*, qualquer processamento de dados pessoais realizado para outros fins é regulamentado pela LGPD, inclusive o processamento realizado por entidades jornalísticas. Por exemplo, o processamento de dados pessoais realizado para preparar um artigo para um portal web de notícias não é regulamentado pela LGPD; mas sim é regulamentado o processamento dos dados pessoais dos usuários desse portal web para analisar seus padrões de navegação e criar perfis comportamentais.

2.2.2 Acadêmico

A seguir, analisamos o segundo cenário, o cenário **acadêmico**. Segundo o Art. 4º da LGPD (BRASIL, 2018):

Art. 4º Esta Lei não se aplica ao tratamento de dados pessoais:

II. realizado para fins exclusivamente:

b) acadêmicos, aplicando-se a esta hipótese os arts. 7º e 11 desta Lei;

Por conseguinte, a LGPD não se aplica ao processamento de dados pessoais para fins puramente **acadêmicos**, mas, ao contrário do cenário anterior, há certas restrições que devem ser levadas em consideração. Essas restrições estão detalhadas nos Arts. 7º e 11º, abaixo mencionamos as mais relevantes (BRASIL, 2018):

Art. 7º O tratamento de dados pessoais somente poderá ser realizado nas seguintes hipóteses:

> *IV. para a realização de estudos por órgão de pesquisa, garantida, sempre que possível, a anonimização dos dados pessoais;*
>
> *Art. 11. O tratamento de dados pessoais sensíveis somente poderá ocorrer nas seguintes hipóteses:*
>
> *II. sem fornecimento de consentimento do titular, nas hipóteses em que for indispensável para:*
>
> *c) realização de estudos por órgão de pesquisa, garantida, sempre que possível, a anonimização dos dados pessoais sensíveis;*

Para entender melhor o que é declarado nesses artigos, é necessário saber a que a LGPD se refere quando menciona **órgão de pesquisa**. No Art. 5º, a lei define o **órgão de pesquisa** como: *"órgão ou entidade da administração pública direta ou indireta ou pessoa jurídica de direito privado sem fins lucrativos legalmente constituída sob as leis brasileiras, com sede e foro no País, que inclua em sua missão institucional ou em seu objetivo social ou estatutário a pesquisa básica ou aplicada de caráter histórico, científico, tecnológico ou estatístico;"*. De acordo com os artigos anteriores, é necessário realizar, sempre que possível, um processo de **anonimização** dos dados pessoais utilizados nos estudos acadêmicos. O Art. 5º define a **anonimização** como a *"utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo;"*. O Art. 12º expõe outros elementos a serem considerados relacionados à anonimização e dados anonimizados, mas não é muito específico sobre os requisitos técnicos do processo de anonimização de dados e delega a responsabilidade de definir padrões e técnicas, além de verificar a segurança do processo à ANPD.

Caso a anonimização dos dados não seja possível, devem ser levados em consideração o restante dos requisitos e regulamentos incluídos no texto da LGPD, que serão explicados abaixo, ao analisar o terceiro cenário de implementação do projeto, o cenário **comercial**.

2.2.3 Comercial

Como mencionado acima, para realizar qualquer tratamento de dados pessoais, é necessário estabelecer a atividade dentro das **10 bases legais** definidas no Art. 7º, as mais relevantes neste caso são (BRASIL, 2018):

Art. 7º O tratamento de dados pessoais somente poderá ser realizado nas seguintes hipóteses:

I. mediante o fornecimento de consentimento pelo titular;

IX. quando necessário para atender aos interesses legítimos do controlador ou de terceiro, exceto no caso de prevalecerem direitos e liberdades fundamentais do titular que exijam a proteção dos dados pessoais;

Isso significa que, para realizar o processo de tratamento de dados pessoais de uma pessoa natural, é necessário obter o consentimento dela ou, se fosse necessário para servir os interesses legítimos da entidade que realiza o processamento ou de terceiros. Dado que as informações coletadas e processadas por nosso sistema são de interesse público, pode-se argumentar que o processamento de dados é realizado para atender ao interesse legítimo de terceiros, a população. O processamento de dados pessoais com base no interesse legítimo do responsável pelo tratamento ou de terceiros pode ser justificado apenas para fins legítimos e exige que uma série de medidas administrativas e técnicas sejam tomadas para garantir a necessidade dos dados utilizados e a transparência do processo, conforme definido no Art. 10º. O Art. 7º também define as seguintes exceções a serem consideradas (BRASIL, 2018):

§ 3º. O tratamento de dados pessoais cujo acesso é público deve considerar a finalidade, a boa-fé e o interesse público que justificaram sua disponibilização.

§ 4º. É dispensada a exigência do consentimento previsto no caput deste artigo para os dados tornados manifestamente públicos pelo titular, resguardados os direitos do titular e os princípios previstos nesta Lei.

§ 6º. A eventual dispensa da exigência do consentimento não desobriga os agentes de tratamento das demais obrigações previstas nesta Lei, especialmente da observância dos princípios gerais e da garantia dos direitos do titular.

§ 7º. O tratamento posterior dos dados pessoais a que se referem os §§ 3º e 4º deste artigo poderá ser realizado para novas finalidades, desde que observados os propósitos legítimos e específicos para o novo tratamento e a preservação dos direitos do titular, assim como os fundamentos e os princípios previstos nesta Lei.

De acordo com essas exceções, é possível processar dados pessoais cujo acesso é público, desde que sejam considerados o objetivo, a boa fé e o interesse público deles. Em outras palavras, eles não devem ser usados para fins incompatíveis com o objetivo original que justificou a publicação desses dados. No contexto do presente trabalho, as informações públicas coletadas e processadas foram divulgadas por órgãos oficiais do governo, como a Câmara dos Deputados e o Senado Federal, cumprindo as leis de acesso à informação em vigor no Brasil e com o objetivo de informar à população, o mesmo objetivo que busca a realização deste projeto.

Também é declarado nessas exceções que você está isento de obter consentimento para o processamento de dados, desde que os dados utilizados tenham sido manifestamente tornados públicos pelo proprietário. Esse seria o caso de dados que os candidatos políticos tornaram públicos durante a campanha para cumprir as leis de informações e financiamento de campanhas políticas, bem como dados divulgados durante seu mandato na Câmara ou no Senado, como discursos públicos, intervenções nas sessões da Câmara ou do Senado, votações públicas, gastos públicos, pronunciamentos nas redes sociais por contas oficiais, etc. Deve-se ter em mente que a não obrigatoriedade de obter permissão não isenta a entidade que realiza o processamento de cumprir as outras obrigações estabelecidas nesta lei, especificamente os princípios gerais da lei e os direitos dos titulares. De particular importância é o direito de ser informado sobre o processamento de dados estabelecido no Art. 9º (BRASIL, 2018):

Art. 9º O titular tem direito ao acesso facilitado às informações sobre o tratamento de seus dados, que deverão ser disponibilizadas de forma clara, adequada e ostensiva acerca de, entre outras características previstas em regulamentação para o atendimento do princípio do livre acesso:

I. finalidade específica do tratamento;

II. *forma e duração do tratamento, observados os segredos comercial e industrial;*
III. *identificação do controlador;*
IV. *informações de contato do controlador;*
V. *informações acerca do uso compartilhado de dados pelo controlador e a finalidade;*
VI. *responsabilidades dos agentes que realizarão o tratamento; e*
VII. *direitos do titular, com menção explícita aos direitos contidos no art. 18 desta Lei.*

Desde que sejam cumpridas as demais obrigações presentes na lei, é possível usar o processamento de dados pessoais obtidos pelas exceções anteriores para novos fins legítimos.

2.3 Como obter conformidade?

Para cumprir com tudo estabelecido na LGPD, é necessário tomar uma série de medidas, tanto administrativas quanto técnicas, para garantir a proteção dos direitos dos titulares sobre seus dados pessoais. Ainda existe um certo nível de incerteza sobre algumas dessas medidas, porque as especificidades de cada uma devem ser definidas pela ANPD. Abaixo, detalhamos essas medidas e os artigos relevantes.

2.3.1 Direito de ser informado

O Art. 9º desta lei afirma que *"O titular tem direito ao acesso facilitado às informações sobre o tratamento de seus dados"* (BRASIL, 2018). É necessário criar um mecanismo automatizado que permita o contato com os titulares dos dados coletados e informá-los sobre o tratamento a ser realizado sobre seus dados pessoais, especificando as informações exigidas no Art. 9º. Se não for possível entrar em contato com os proprietários automaticamente, é preciso tentar contatá-los manualmente, desde que isso não implique um esforço desproporcional. Se não houver informações de contato dos titulares, é aconselhável anonimizar ou descartar tais dados.

2.3.2 Garantir a transparência do tratamento

O Art. 10º define uma série de condições para poder basear o processamento de dados pessoais na hipótese de interesse legítimo, *"O controlador deverá adotar medidas para garantir a transparência do tratamento de dados"* e *"A autoridade nacional poderá solicitar ao controlador relatório de impacto à proteção de dados pessoais"*. Estas condições são complementadas pelo Art. 37º que afirma que *"O controlador e o operador devem manter registro das operações de tratamento de dados pessoais que realizarem"* e pelo Art. 38º, que diz o seguinte: *"A autoridade nacional poderá determinar ao controlador que elabore relatório de impacto à proteção de dados pessoais, inclusive de dados sensíveis, referente a suas operações de tratamento de dados"* (BRASIL, 2018). É por isso que devem ser criados sistemas de registro de operações durante o processo automático de coleta e processamento de dados. As decisões manuais tomadas por todos os envolvidos no processo de tratamento também devem ser registradas. É imperativo criar uma metodologia para a criação de um relatório de impacto de proteção de dados. Este relatório deve ser feito mesmo sem ser exigido pela ANPD, pois pode ser útil para entender o tratamento realizado e pode ajudar a alcançar a conformidade com a LGPD mais facilmente. Este relatório deve incluir, entre outras coisas, o seguinte: os objetivos do tratamento, que tipo de dados são processados, quem dentro da entidade tem acesso a esses dados, quais terceiros têm acesso a esses dados, que medidas de segurança estão em vigor para garantir a proteção dos dados e por quanto tempo esses dados vão ser mantidos.

2.3.3 Dados anonimizados

Como foi explicado anteriormente, a LGPD apresenta uma exceção para *"realização de estudos por órgão de pesquisa, garantida, sempre que possível, a anonimização dos dados pessoais"*, e o Art. 12º afirma que *"Os dados anonimizados não serão considerados dados pessoais para os fins desta Lei"* (BRASIL, 2018). Considerando isso, é necessário realizar um processo de anonimização dos dados, sempre que possível e isso permitir que os objetivos do processamento sejam alcançados. A Anonimização de Dados pode ser definida como o processo de proteção de dados privados e sensíveis, modificando-os para ocultar informação identificadora. Segundo o Art. 10º, *"Quando o tratamento for baseado no legítimo interesse do controlador, somente os dados pessoais estritamente necessários para*

a finalidade pretendida poderão ser tratados.", portanto, recomenda-se anonimizar dados que não são considerados essenciais; um exemplo vinculado ao nosso projeto poderia ser os dados pessoais de pessoas físicas que fizeram doações para as campanhas eleitorais de candidatos políticos. O assunto de anonimização de dados é bastante amplo e complexo e é de responsabilidade da ANPD definir padrões e técnicas de anonimização, razão pela qual não é possível realizar uma análise de técnicas específicas neste trabalho. Aqui estão algumas técnicas usadas hoje na indústria de desenvolvimento (IMPERVA, 2020):

- **Mascaramento de dados**: Refere-se ao processo de ocultar informações usando técnicas de alteração dos dados, como criptografia, substituição ou reposicionamento de caracteres.
- **Pseudonimização**: É um método de gerenciamento de dados que consiste em modificar identificadores por pseudônimos ou falsos identificadores e excluir dados pessoais do banco de dados principal, mantendo outro banco seguro onde esses dados estão localizados.
- **Generalização**: Consiste na eliminação de elementos específicos dos dados para dificultar a identificação. Valores específicos podem ser substituídos por intervalos de valores que permitem reter a precisão dos dados. Essa técnica é usada pela Google como parte de suas políticas de privacidade e proteção de dados (GOOGLE, 2020).
- **Troca de dados**: Consiste em alterar os valores dos dados entre diferentes registros. Por exemplo, alterando a data de nascimento ou nome de uma pessoa para outra, em todo o repositório. Isso permite ter informações que aparentemente não permitem identificar pessoas específicas sem perda de dados.
- **Perturbação de dados**: É o processo de modificar os dados usando técnicas de arredondamento de valor ou adicionar ruído aleatório. Esta é outra técnica usada pela Google como parte de suas políticas de privacidade e proteção de dados (GOOGLE, 2020).
- **Dados sintéticos**: Criar dados sintéticos por algoritmos usando padrões encontrados nos dados originais como base. É comum usar técnicas estatísticas para criar esses dados.

2.3.4 Garantir os direitos dos titulares

O principal objetivo da LGPD é garantir os direitos que os titulares têm sobre seus dados pessoais, razão pela qual é essencial implementar mecanismos que garantam a integridade desses direitos durante o processo de tratamento de dados. O Art. 18º define os direitos fundamentais dos titulares, que são: confirmação do tratamento, acesso aos dados, correção de dados incompletos, desatualizados ou imprecisos, anonimização ou exclusão de dados não essenciais, portabilidade e exclusão de dados obtidos por meio de consentimento, informações de terceiros com os quais os dados são compartilhados, informações sobre a possibilidade de negar o consentimento e suas consequências, e revogação do consentimento. O Art. 19º indica também que a entidade deve poder fornecer, técnica e administrativamente, a confirmação do tratamento e acesso aos dados em um formato simplificado imediatamente ou em um formato mais completo dentro de um período de até 15 dias a partir da solicitação do titular (BRASIL, 2018). Um sistema automatizado deve ser implementado que permita a confirmação do tratamento e o acesso imediato e simplificado aos dados; a necessidade de verificar a identidade do titular deve ser levada em consideração ao implementar o sistema. Também é necessário criar uma metodologia para a preparação manual de informações detalhadas a serem entregues ao titular dentro do prazo de 15 dias.

O Art. 20º requer a existência de mecanismos que permitam ao titular solicitar uma revisão das decisões tomadas pelo processamento automatizado de dados (BRASIL, 2018). A entidade que realiza o processamento deve fornecer informações claras sobre os critérios usados para o tratamento automatizado de dados pessoais.

2.3.5 Medidas administrativas de segurança, boas práticas e governança

De acordo com o Art. 39º, os operadores designados pela entidade processadora devem *"realizar o tratamento segundo as instruções fornecidas pelo controlador, que verificará a observância das próprias instruções e das normas sobre a matéria"* (BRASIL, 2018). É por isso que deve existir uma metodologia bem descrita em relação ao processo de tratamento, e é responsabilidade da entidade que essas instruções sejam seguidas.

A entidade deve designar um Responsável pelo Tratamento de Dados Pessoais, de acordo com o Art. 41º, cuja identidade e informações de contato devem

ser tornadas públicas (BRASIL, 2018). O Responsável tem a responsabilidade de se comunicar com os titulares dos dados, aceitar suas reclamações e tomar medidas em relação a essas comunicações, manter contato com a ANPD e tomar as medidas necessárias, instruir os membros da entidade sobre as políticas de privacidade e proteção de dados e executar outras atribuições ou regras ditadas pela entidade ou pela ANPD.

No que diz respeito à segurança durante o tratamento de dados pessoais, o Art. 46º estabelece as seguintes (BRASIL, 2018):

Art. 46. Os agentes de tratamento devem adotar medidas de segurança, técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou qualquer forma de tratamento inadequado ou ilícito.

§ 1º. A autoridade nacional poderá dispor sobre padrões técnicos mínimos para tornar aplicável o disposto no caput deste artigo, considerados a natureza das informações tratadas, as características específicas do tratamento e o estado atual da tecnologia, especialmente no caso de dados pessoais sensíveis, assim como os princípios previstos no caput do art. 6º desta Lei.

§ 2º. As medidas de que trata o caput deste artigo deverão ser observadas desde a fase de concepção do produto ou do serviço até a sua execução.

Como os padrões mínimos para garantir a segurança do processo de tratamento ainda não foram definidos pela ANPD, não é possível analisar em profundidade os padrões, técnicas ou medidas de segurança que devem ser adotadas. Levando em conta a natureza do projeto, que está sendo desenvolvido usando tecnologias da Web, algumas ameaças conhecidas que devem ser consideradas podem ser mencionadas (MOZILLA MSDN, 2020):

- **Cross-Site Scripting (XSS)**: Refere-se ao tipo de ataques que permitem a injeção de scripts do lado do cliente nos navegadores dos usuários. As melhores defesas contra esse tipo de ataque são remover elementos da página que podem ser usados para executar código,

como *<script>*, *<link>*, etc. e limpar as informações inseridas no site pelos usuários e salvas no banco de dados.

- **Injeção SQL**: Consiste em passar o código SQL através de um campo de entrada que pode alterar a natureza da consulta original e modificar o banco de dados. Uma maneira de evitar esse ataque é limpar as informações inseridas no site pelos usuários.
- **Cross-Site Request Forgery (CSRF)**: Consiste em usar a identidade de usuários conectados ao fazer solicitações para um site. Para evitar esse ataque, é necessário incluir *tokens* em formulários nas páginas web e negar solicitações que não contenham esses *tokens*.

Todas as pessoas envolvidas durante o processo de tratamento devem garantir a segurança dos dados, mesmo após o término do processo de tratamento, conforme o Art. 47º (BRASIL, 2018). Por esse motivo, a entidade de processamento deve criar instruções de segurança claras, levando em consideração os padrões que a ANPD deve definir.

O Art. 48º exige que a ANPD seja informada de qualquer violação da segurança dos dados (BRASIL, 2018). O termo para esta comunicação ainda não foi definido pela ANPD, mas deve ser um período razoável a partir da data da violação de segurança. A entidade deve apresentar as seguintes informações quando ocorrer uma violação de segurança e deve adotar as disposições exigidas pela ANPD (BRASIL, 2018):

I. a descrição da natureza dos dados pessoais afetados;
II. as informações sobre os titulares envolvidos;
III. a indicação das medidas técnicas e de segurança utilizadas para a proteção dos dados, observados os segredos comercial e industrial;
IV. os riscos relacionados ao incidente;
V. os motivos da demora, no caso de a comunicação não ter sido imediata; e
VI. as medidas que foram ou que serão adotadas para reverter ou mitigar os efeitos do prejuízo.

Finalmente, o Art. 50º define os poderes e obrigações da entidade processadora para a criação de regras, normas e medidas de boas práticas e

governança para a proteção da privacidade e segurança dos dados. A seguir, reproduzimos o Art. 50º em sua totalidade, uma vez que define detalhadamente os requisitos que devem ser atendidos para criar essas regras (BRASIL, 2018):

Art. 50. Os controladores e operadores, no âmbito de suas competências, pelo tratamento de dados pessoais, individualmente ou por meio de associações, poderão formular regras de boas práticas e de governança que estabeleçam as condições de organização, o regime de funcionamento, os procedimentos, incluindo reclamações e petições de titulares, as normas de segurança, os padrões técnicos, as obrigações específicas para os diversos envolvidos no tratamento, as ações educativas, os mecanismos internos de supervisão e de mitigação de riscos e outros aspectos relacionados ao tratamento de dados pessoais.

§ 1º. Ao estabelecer regras de boas práticas, o controlador e o operador levarão em consideração, em relação ao tratamento e aos dados, a natureza, o escopo, a finalidade e a probabilidade e a gravidade dos riscos e dos benefícios decorrentes de tratamento de dados do titular.

§ 2º. Na aplicação dos princípios indicados nos incisos VII e VIII do caput do art. 6º desta Lei, o controlador, observados a estrutura, a escala e o volume de suas operações, bem como a sensibilidade dos dados tratados e a probabilidade e a gravidade dos danos para os titulares dos dados, poderá:

I. implementar programa de governança em privacidade que, no mínimo:

a) demonstre o comprometimento do controlador em adotar processos e políticas internas que assegurem o cumprimento, de forma abrangente, de normas e boas práticas relativas à proteção de dados pessoais;

b) seja aplicável a todo o conjunto de dados pessoais que estejam sob seu controle, independentemente do modo como se realizou sua coleta;

c) seja adaptado à estrutura, à escala e ao volume de suas operações, bem como à sensibilidade dos dados tratados;

d) estabeleça políticas e salvaguardas adequadas com base em processo de avaliação sistemática de impactos e riscos à privacidade;

e) tenha o objetivo de estabelecer relação de confiança com o titular, por meio de atuação transparente e que assegure mecanismos de participação do titular;

f) esteja integrado a sua estrutura geral de governança e estabeleça e aplique mecanismos de supervisão internos e externos;

g) conte com planos de resposta a incidentes e remediação; e

h) seja atualizado constantemente com base em informações obtidas a partir de monitoramento contínuo e avaliações periódicas;

II. demonstrar a efetividade de seu programa de governança em privacidade quando apropriado e, em especial, a pedido da autoridade nacional ou de outra entidade responsável por promover o cumprimento de boas práticas ou códigos de conduta, os quais, de forma independente, promovam o cumprimento desta Lei.

§ 3º. As regras de boas práticas e de governança deverão ser publicadas e atualizadas periodicamente e poderão ser reconhecidas e divulgadas pela autoridade nacional.

2.4 LGPD na prática, GDPR

Conforme mencionado acima, a LGPD e a GDPR têm uma grande semelhança. Essas semelhanças são as seguintes (GDPR.EU, 2020):

- **Dados protegidos**: Tanto a GDPR quanto a LGPD regulamentam o processamento de dados pessoais de pessoas físicas. A GDPR apresenta, em seu Art. 4º, uma definição bastante específica de dados pessoais, especificando alguns tipos de dados que podem ser usados para identificar uma pessoa. Esses tipos de dados são: um nome, um número de identificação, dados de localização, um identificador online ou um ou mais fatores específicos para identidade física, fisiológica, genética, mental, económica, cultural o social. Entretanto, a definição de dados pessoais da LGPD é muito mais ampla, considera-se dado

pessoal qualquer informação que por si ou em conjunto com outros dados possa identificar uma pessoa.

- **Direitos fundamentais**: A GDPR determina 8 direitos fundamentais dos titulares dos dados. Esses direitos correspondem diretamente aos 9 direitos fundamentais presentes na LGPD. O “Direito de ser informado” da GDPR, corresponde a 2 direitos definidos na LGPD: “Direito à confirmação da existência do tratamento” e “Direito à informação das entidades públicas com as quais o controlador compartilhou os dados”
- **Escopo territorial**: A GDPR e a LGPD são leis com escopo extraterritorial. Essas leis se aplicam a qualquer organização que realize um processo de tratamento de dados de pessoas localizadas na União Europeia ou no Brasil, respectivamente.
- **Responsável pelo Tratamento de Dados**: Ambas as leis exigem que as entidades que realizam o tratamento designem um Responsável pelo Tratamento de Dados. A GDPR determina que apenas as entidades que processam dados em grande escala tenham de designar o Responsável, enquanto a LGPD exige que o Responsável seja definido por qualquer entidade, independentemente da dimensão desta ou da quantidade de dados processados.
- **Bases legais**: O GDPR define 6 bases legais para realizar o processo de tratamento de dados. Essas bases legais são: consentimento explícito, para execução de contratos, para políticas públicas, para proteção da vida, para obrigações legais e no legítimo interesse. Essas bases jurídicas estão incluídas na LGPD, também são adicionadas mais 4 bases legais: realização de estudos por órgão de pesquisa, para o exercício de direitos em processos judiciais, para proteção da saúde e proteção do crédito.
- **Solicitações de acesso aos dados**: Em ambas as leis, os titulares dos dados têm o direito de solicitar, da entidade que realiza o tratamento, o acesso aos seus dados. A GDPR determina um prazo de 30 dias para responder a essas solicitações. Na LGPD o período de resposta dura apenas 15 dias.

- **Notificações de violação de segurança de dados**: As violações de segurança de dados devem ser notificadas às Autoridades de Proteção de Dados, tanto na GDPR quanto na LGPD. A diferença está no prazo para fazê-lo. A GDPR define um prazo estrito de 72 horas, enquanto a LGPD não define nenhuma duração específica. A LGPD determina que as notificações devem ser feitas em um prazo razoável, a partir da ocorrência da violação de segurança.
- **Multas:** Ambas as leis permitem que as Autoridades de Proteção de Dados apliquem multas altas para entidades que violam os regulamentos. A GDPR define multas de até 4% dos lucros anuais da entidade ou de até €20 milhões, o que for maior. A LGPD define multas de até 2% do lucro anual da entidade ou de até R$ 50 milhões, o que for maior.

Para analisar quais são as consequências da violação das indicações da LGPD na prática, usamos a GDPR como ponto de comparação e analisamos um caso que ocorreu em 15 de março de 2019 na Polônia.

A empresa de agregação de dados Bisnode, sediada na Suécia, recebeu uma multa de aproximadamente € 220.000 por violação do Art. 14º da GDPR. O Art. 14º da GDPR é equivalente ao Art. 9º da LGPD e exige que as entidades que realizam o processamento informem os titulares dos dados sobre qualquer tratamento realizado, eles devem informar quais dados são processados, como são processados, por quanto tempo serão processados, etc. Este artigo tem algumas exceções a serem consideradas, as entidades estão isentas de informar os titulares se *"o fornecimento de tais informações for impossível ou implicaria um esforço desproporcional"* ou se essa obrigação *"provavelmente impossibilitará ou prejudicará seriamente a realização de os objetivos desse processamento"* (GERRISH LEGAL, 2019) (TECHNOLOGY LAW DISPATCH, 2019).

A Bisnode, com o objetivo de fornecer relatórios e serviços de verificação, usou bancos de dados e registros públicos que continham dados pessoais de empresários e empreendedores. Esse conjunto de dados continha aproximadamente 7,6 milhões de registros com dados pessoais, dos quais apenas 700.000 tinham e-mail, o restante apenas endereços postais e números de celular. Bisnode usou os e-

mails para informar os titulares dos dados e colocou um aviso no seu site para informar as pessoas que não foram notificadas. Segundo Bisnode, o custo de notificar o restante das pessoas usando mensagens de texto ou pelo correio teria sido desproporcional. O custo estimado para notificar os titulares usando correio foi de aproximadamente € 7,7 milhões, superior aos ganhos da Bisnode no ano anterior. O custo adicional da contratação de pessoal para realizar a notificação comprometeria a continuação das operações da Bisnode na Polônia.

Apesar dos argumentos de Bisnode, a Autoridade Polonesa de Proteção de Dados (UODO) decretou que Bisnode tinha violado o Art. 14º da GDPR ao não informar todos os titulares do processamento de dados. O UODO alega que Bisnode estava errada em sua racionalização de que informar os proprietários teria sido um esforço desproporcional e que continuar processando os dados, tendo pleno conhecimento da GDPR e de suas responsabilidades como entidade que realiza o processamento, foi um fator agravante a decisão. Muitos titulares não sabiam que seus dados estavam sendo processados e, ao não os informarem, Bisnode os privou de seus direitos e não lhes deu a oportunidade de recusar o processamento.

3 FUNDAMENTAÇÂO TEÓRICA

3.1 Mineração de Dados

Segundo Hand, Mannila e Smyth (2001), a Mineração de Dados (*Data Mining*, DM) pode ser definida como “a análise de conjuntos de dados de observação (muitas vezes grandes) para encontrar relações insuspeitadas e resumir os dados de novas maneiras que são compreensíveis e úteis para o proprietário dos dados. Os relacionamentos e resumos derivados de um exercício de Mineração de Dados são frequentemente chamados de modelos ou padrões”. Existe outro conceito intimamente relacionado ao DM, o Descobrimento de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases*, KDD): "KDD é o processo não trivial de identificar padrões válidos, novos, potencialmente úteis e, em última instância, compreensíveis nos dados." (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Os termos DM e KDD são comumente tratados como sinônimos e usados indistintamente por algumas fontes, enquanto outras fontes consideram DM como "uma etapa no processo de KDD que consiste na aplicação de algoritmos de análise e descoberta de dados que, sob limitações de eficiência computacionais aceitáveis, produzem uma enumeração particular de padrões (ou modelos) nos dados" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Para garantir que conhecimento útil seja extraído dos dados, as etapas adicionais no processo de KDD são essenciais. Durante o processo de DM, são usados algoritmos matemáticos sofisticados para deduzir e extrair os padrões, tendências e relacionamentos existentes nos dados. Esses padrões não podem ser detectados usando técnicas tradicionais de exploração e consultas de Banco de Dados, isso ocorre porque os relacionamentos entre os dados são muito complexos ou há uma quantidade extremamente grande de dados para processar.

O DM pode ser categorizado em diferentes tipos de tarefas, dependendo dos objetivos perseguidos ao realizar o processo (HAND; MANNILLA; SMYTH, 2001):

1. **Análise Exploratória de Dados** (*Exploratory Data Analysis*, EDA): O objetivo do EDA é a exploração e visualização de dados, usando

técnicas de visualização interativas, projetadas para conjuntos de dados de baixa dimensão.

2. **Modelagem Descritiva**: O objetivo da Modelagem Descritiva é descrever os padrões e relacionamentos que existem dentro dos dados.
3. **Modelagem Preditiva**: O objetivo da Modelagem Preditiva é a construção de um modelo que permita a predição de uma ou mais variáveis, usando como base os padrões existentes em um repositório de dados.
4. **Descobrir Padrões e Regras**: O objetivo é detectar padrões incomumente frequentes ou infrequentes que existem nos dados.
5. **Recuperação por Conteúdo**: O objetivo é, tendo um conjunto de padrões existentes, encontrar padrões semelhantes nos dados.

Segundo Aggarwal (2015), Bramer (2007), Han, Pei e Hamber (2011) e Jackson (2002) existem vários problemas fundamentais a serem resolvidos pelo processo de DM, esses problemas são:

1. **Mineração de Padrões Frequentes**: Os padrões frequentes são conjuntos de elementos, subsequências ou subestruturas que aparecem em um conjunto de dados com uma frequência não inferior a um limite especificado pelo usuário. Encontrar padrões frequentes desempenha um papel essencial na mineração de associações, correlações e muitos outros relacionamentos interessantes entre os dados.
2. **Classificação de Dados**: É o processo de encontrar um modelo que descreve e distingue classes de dados e conceitos. A classificação é o problema de identificar a qual conjunto de categorias um novo registro pertence, usando um conjunto de treinamento de dados que contém registros pertencentes a categorias previamente conhecidas. O objetivo é prever o valor de uma ou mais variáveis discretas.
3. **Predição Numérica** (*Regressão*): É semelhante ao processo de Classificação, a diferença é que, ao invés de prever variáveis discretas, o objetivo é prever o valor de variáveis numéricas contínuas.

4. **Agrupamento de Dados** (*Clustering*): É o processo de agrupar elementos existentes dentro de um repositório de dados de acordo com sua similaridade. O objetivo é maximizar a similaridade intraclasse e minimizar a similaridade interclasse. Isso significa que cada elemento pertencente a um grupo compartilha uma alta similaridade entre eles e uma baixa similaridade com membros de outros grupos.
5. **Detecção de Valores Atípicos** (*Outliers*): "Um valor atípico é uma observação que se desvia tanto das outras observações que levanta a suspeita de que foi gerada por um mecanismo diferente." (HAWKINS, 1980). Portanto, um valor atípico geralmente contém informações úteis sobre características anormais de sistemas e entidades que afetam o processo de geração de dados.
6. **Descrição de Classes/Conceitos**: É o processo de extrair uma descrição compacta dos dados, classes e conceitos. Podem ser extraídas resumindo uma classe específica (Caracterização de Dados) ou comparando várias classes (Discriminação de Dados).

3.1.1 Modelos de Mineração de Dados

Os padrões e relacionamentos encontrados durante o processo de DM podem ser compilados e definidos como um Modelo de Mineração de Dados (Data Mining Model, DMM). Um DMM pode ser definido como "um resumo geral de um conjunto de dados para identificar e descrever as principais características do formato da distribuição." (JACKSON, 2002). Esses modelos podem ser aplicados a vários cenários:

- **Previsão**: Estimativa de vendas, previsão de cargas do servidor ou tempo de inatividade do servidor.
- **Risco e probabilidade**: Determinar o provável ponto de equilíbrio para cenários de risco, atribuir probabilidades a diagnósticos ou outros resultados.
- **Recomendações**: Determinar quais produtos provavelmente serão vendidos juntos, gerar recomendações.

- **Encontrar sequências**: Analisar as seleções do cliente em um carrinho de compras, prever os próximos eventos prováveis.
- **Agrupamento**: Separar clientes ou eventos em grupos de elementos relacionados, analisar e prever afinidades.

Para gerar um DMM utilizamos um processo bastante completo que inclui desde a formulação do problema a ser resolvido até a aplicação do modelo gerado.

Existem vários esforços para definir uma metodologia geral para realizar o processo de geração de um DMM. Um desses esforços resultou na criação do projeto Cross-Industry Standard Process for Data Mining (CRISP-DM), um padrão não proprietário, documentado e disponível gratuitamente. O objetivo do projeto CRISP-DM é "definir e validar um modelo de processo DM neutro da indústria e ferramentas que tornariam o desenvolvimento de grandes e pequenos projetos de DM mais rápido, barato, mais confiável e mais gerenciável." (JACKSON, 2002). O processo CRISP-DM é cíclico, dinâmico e iterativo (Fig. 2), O que significa que, para gerar um modelo adequado, as etapas que compõem o processo devem ser repetidas várias vezes.

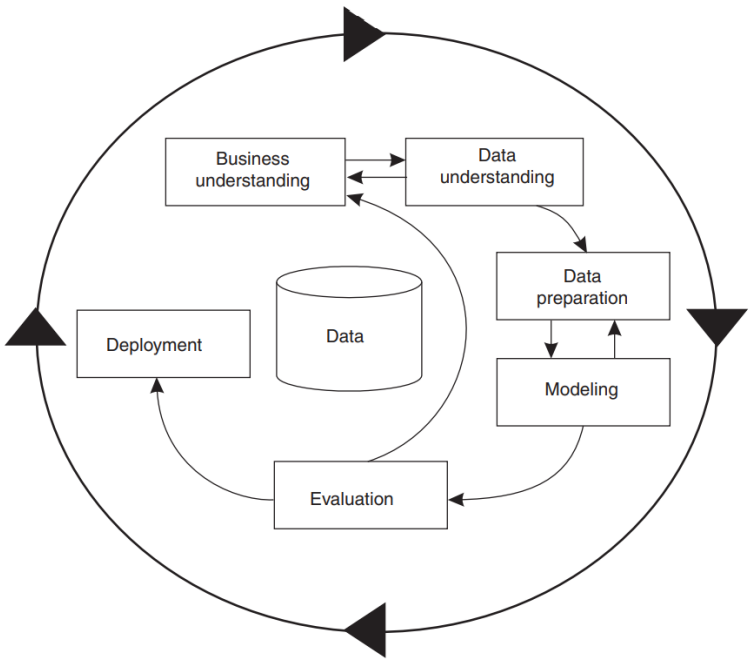

Figura 2 CRISP-DM
Fonte: (MARISCAL; MARBAN; FERNANDEZ, 2010)

O CRISP-DM consiste nas seguintes etapas ou estágios:

3.1.1.1 Compreensão do negócio/problema:

A primeira etapa no processo de geração de um DMM é definir claramente o problema a ser resolvido, bem como as formas de usar os dados disponíveis para encontrar uma solução. Inclui a definição de métricas para avaliar o modelo e os objetivos específicos perseguidos com sua geração. Para obter essas informações, é possível criar as seguintes questões:

- O que está sendo procurado? Quais são as relações que se deseja explorar?
- O modelo vai ser preditivo ou descritivo? O seja, O modelo será usado para fazer previsões ou apenas deseja encontrar associações e padrões dentro dos dados?
- Qual é o resultado ou atributo que se deseja prever?
- Que tipo de dados se tem? Que informações estão em cada coluna na tabela? Se houvesse várias tabelas, Como elas se relacionam? É necessário fazer alterações nos dados, como limpar ou adicionar novos dados?

3.1.1.2 Compreensão dos dados:

A segunda etapa do processo é a coleta inicial e exploração dos dados disponíveis.

É necessário ter um conhecimento adequado do problema e dos dados disponíveis para facilitar a tomada de decisão correta na hora de criar o DMM. Existem várias técnicas de exploração de dados, incluindo o cálculo dos valores mínimo e máximo, o cálculo da média e do desvio padrão e a análise da distribuição deles. Após a aplicação dessas técnicas, é possível determinar a qualidade dos dados e se eles representam totalmente o escopo do problema a ser resolvido. Por exemplo, se houver um desvio padrão muito grande, pode ser necessário coletar mais dados para melhorar o modelo.

3.1.1.3 Preparação dos dados:

A terceira etapa do processo é a consolidação e limpeza dos dados coletados.

A consolidação de dados consiste em agrupar todos os dados disponíveis em um único repositório. Eles devem ser armazenados no mesmo formato, para isso pode ser necessário modificar alguns dos registros: pode ser que uma das fontes tenha um atributo que representa a idade de uma pessoa, enquanto outra fonte tem a data de nascimento. É por isso que uma análise profunda das fontes utilizadas deve ser feita para se achar uma solução satisfatória para esses problemas.

Depois de consolidar os dados em um único repositório, o processo de limpeza de dados começa. Podem existir registros com atributos inválidos, incorretos ou ausentes. Nesses casos, é necessário decidir como tentar corrigir essas situações. As soluções possíveis são eliminar as entradas que apresentam esses problemas ou modificar os valores com interpolações e/ou valores médios.

3.1.1.4 Modelagem:

A quarta etapa do processo consiste em gerar o DMM utilizando os conhecimentos adquiridos nas etapas anteriores.

Nesta etapa, são especificadas as colunas que serão utilizadas como entradas para o algoritmo de DM, os atributos que serão previstos e os parâmetros que o algoritmo utilizará para processar os dados. Este processamento é denominado treinamento e consiste no processo de aplicação de um algoritmo matemático sobre os dados definidos, a fim de obter os padrões e tendências necessários para a solução do problema. Esta etapa inclui a tarefa de escolher o algoritmo que será utilizado para a obtenção do modelo, esta escolha influenciará diretamente nos padrões encontrados no processo de treinamento. Existe muitos algoritmos criados para realizar o processamento de dados em DM, cada algoritmo é projetado para uma tarefa específica, portanto esta escolha afetará diretamente o resultado a ser obtido. Os algoritmos existentes estão diretamente relacionados aos problemas que o DM aborda mencionados acima: Mineração de Padrões Frequentes, Classificação de Dados, Predição Numérica (*Regressão*), Agrupamento de Dados (*Clustering*), Detecção de Valores Atípicos (*Outliers*) e Descrição de Classes/Conceitos.

Ao adicionar novos dados, é necessário executar novamente o processo de treinamento para atualizar o modelo.

3.1.1.5 Avaliação:

A quinta etapa do processo é a avaliação dos modelos gerados para verificar seu desempenho antes de sua implantação em um ambiente de produção.

Esta etapa avalia o grau em que o modelo atende aos objetivos definidos na primeira etapa do processo e busca identificar os motivos da deficiência do modelo. Se o modelo não satisfizer corretamente os objetivos definidos, é necessário voltar às etapas anteriores do processo e ver se é necessário redefinir o problema ou aprimorar os dados usados para gerar o modelo.

3.1.1.6 Implantação:

A última etapa do processo é a implantação do DMM gerado em um ambiente de produção. Nesta etapa, é possível extrair informações acionáveis do repositório e aplicar o modelo a novos dados. É possível realizar a extração de detalhes do modelo ou a integração em aplicativos ou ferramentas de consulta e relatórios. É necessário manter o modelo atualizado adicionando novos dados ao repositório.

3.2 Mineração de Dados na Web

Não há consenso sobre a definição exata da Mineração de Dados na Web (Web Data Mining, WDM). Segundo Etzioni (1996), um dos primeiros a criar um conceito sobre o tema, o WDM pode ser definida como uma série de subtarefas:

- **Descoberta de recursos**: Encontrar documentos e serviços desconhecidos na Web.
- **Extração de informação**: Extração automática de informações específicas de recursos da web recém-descobertos.
- **Generalização**: Descobrir padrões gerais em páginas individuais e em vários sites.

Kosala e Blockeel (2000) adicionam mais uma tarefa a essa definição (Fig. 3):

- **Análise**: Validação e/ou interpretação dos padrões encontrados.

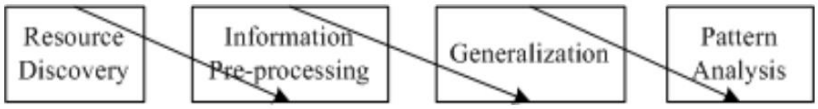

Figura 3 Processo Geral da Mineração de Dados na Web
Fonte: (BIN; ZHIJING, 2003)

Outra definição que pode ser mencionada é:

"A Mineração de Dados na Web é a aplicação de técnicas de Mineração de Dados para encontrar conhecimento interessante e potencialmente útil de dados da web. Normalmente, espera-se que a estrutura dos links da web ou os dados de registros da web ou ambos tenham sido usados no processo de mineração." (SINGH; SINGH, 2010)

Portanto, uma definição mais detalhada de WDM pode ser:

O WDM é um processo complexo que inclui a análise de uma ampla variedade de informações, como o conteúdo e a estrutura de documentos da web, arquivos de texto, links, bancos de dados, registros de acesso de usuários, registros de servidores, perfis de usuários, entre outros, a fim de encontrar informações úteis e relevantes, bem como encontrar novos conhecimentos que atendam às necessidades e a tomada de decisão dos usuários.

O WDM é uma subárea ou subdisciplina do DM tradicional. É um processo que, como o próprio nome sugere, tem como área de atuação os dados da World Wide Web.

Uma das principais diferenças que separa o WDM do DM tradicional é o tipo de dados nos quais o processo é aplicado. No DM tradicional, os dados estruturados são usados principalmente em um ou mais repositórios. Devido à grande heterogeneidade de dados existentes na web e ao seu enorme volume, o WDM pode ser aplicado a dados estruturados, semiestruturados e não estruturados (SINGH; SINGH, 2010). Este é um desafio muito grande que requer o uso criativo de técnicas de DM, bem como novos métodos, estratégias e algoritmos que foram desenvolvidos especialmente para WDM.

Ao longo de sua história, uma grande variedade de comunidades científicas contribuiu para o desenvolvimento do WDM. Essas comunidades surgiram de vários campos de pesquisa, como Aprendizado de Máquina, Bancos de Dados, Processamento de Linguagem Natural e Recuperação de Informação. Cada uma dessas comunidades adquiriu grande relevância pelas seguintes razões (MENDOZA, 2011):

- **Aprendizado de Máquina**: A principal tarefa do Aprendizado de Máquina é ser capaz de generalizar o comportamento aprendido a partir de padrões específicos que permitem a criação de novos conhecimentos e ações. Isso permitiu, no contexto do WDM, a identificação de padrões e tendências nos dados, o que facilitou diversas tarefas como a classificação de documentos e conteúdos.
- **Bancos de Dados**: Para organizar o grande volume de informações extraídas da web, o uso de tecnologias e ferramentas de Bancos de Dados tem se mostrado necessário e essencial.
- **Processamento de Linguagem Natural**: Muitas das informações disponíveis na web são representadas em texto. Através do uso de técnicas de Processamento de Linguagem Natural, é possível extrair informações complementares que enriquecem a descrição do conteúdo das páginas web.
- **Recuperação de Informação**: A precisão do conteúdo recomendado e das listas de documentos, criadas por sistemas de recuperação de informações, pode ser melhorada aplicando técnicas de WDM.

Kosala e Blockeel (2000) identificaram os principais problemas que afetam os usuários e são tratados através do WDM:

1. **Pesquisa de informações relevantes**: Os usuários navegam ou usam os serviços de pesquisa para encontrar informações específicas na web. As ferramentas de pesquisa atuais têm vários problemas:
    a) Baixa precisão, porque muitos dos resultados da pesquisa não são relevantes.

b) Impossibilidade de indexar todas as informações disponíveis na Web. Isso torna difícil encontrar informações relevantes que não foram indexadas.

2. **Criar conhecimentos a partir das informações disponíveis na Web**: Surge como um subproblema do anterior. Presume-se que haja uma grande coleção de dados da web disponíveis e o objetivo é extrair conhecimento potencialmente útil a partir deles.
3. **Personalização da informação**: Há uma diferença nas preferências e na apresentação do conteúdo para os diferentes usuários que interagem com a web.
4. **Aprender sobre consumidores ou usuários individuais**: Consiste em entender as ações e desejos dos usuários e consumidores. Dependendo da forma como utilizam a web, é possível descobrir informações sobre os usuários que permitem diversas ações como classificação de usuários, customização do serviço, sugestões, previsão e recomendação de ações, entre muitas outras.

O processo de WDM é muito semelhante ao processo de DM tradicional. Como mencionamos antes, ele pode ser dividido nas seguintes subtarefas, etapas ou estágios (ETZIONI, 1996) (KOSALA; BLOCKEEL, 2000):

1. **Descoberta de recursos**: Nesta etapa, todos os recursos nos quais o WDM será realizado são coletados. Esses recursos podem ser coletados de várias páginas web ou de sites inteiros. Dentro desses recursos, pode haver diferentes tipos de dados que podem ser de interesse, como documentos de texto, imagens, áudio ou vídeos. Esta etapa é conhecida como *crawling* e várias técnicas de Recuperação de Informação são usadas.
2. **Extração de Informação**: Nesta etapa, é realizado o processo de extração das informações específicas encontradas em cada recurso e descartadas as informações não relevantes ao processo de WDM. Esta etapa é conhecida como *scraping*, e as tarefas de seleção, filtragem e pré-processamento da informação são realizadas.
3. **Generalização**: Nesta etapa, os padrões de interesse, tendências e novos conhecimentos são extraídos dos repositórios de informações

identificados e coletados durante as duas primeiras etapas. Técnicas de DM supervisionadas e não supervisionadas são usadas.

4. **Análise**: Nesta etapa, é realizada a validação e interpretação dos padrões encontrados na fase anterior. Nesta etapa é possível identificar padrões relevantes, tendências e novos conhecimentos extraídos, com o objetivo de agregar valor ao entendimento dos repositórios de informações utilizados durante o processo.

Dependendo da fonte de informação utilizada e do conhecimento a ser obtido, o WDM pode ser classificado em três áreas principais de interesse (Fig. 4) (KOSALA; BLOCKEEL). Essas áreas são:

- **Mineração de Conteúdo da Web**: É o processo de extrair informações úteis do conteúdo de documentos, páginas ou sites. As técnicas de WDM são aplicadas diretamente ao conteúdo de páginas ou sites, sejam eles textos, imagens, vídeos, etc., para identificar padrões que permitem a criação de novos conhecimentos ou complementam informações existentes. Inclui tarefas de classificação, agrupamento, recuperação de informações específicas e análise de conteúdo criado por usuários para entender sua percepção de diferentes elementos, como produtos ou pessoas.
- **Mineração da Estrutura da Web**: É o processo de descobrir informações sobre a estrutura da web usando os links que conectam as diferentes páginas e sites. Ao realizar uma análise dos links entre diferentes páginas e sites, é possível extrair informações relevantes que permitem entender como vários sites se relacionam, bem como descobrir tentativas de enganar os algoritmos de posicionamento em buscadores.
- **Mineração de Uso da Web**: É o processo de recuperar informações úteis sobre os padrões de navegação e acesso dos usuários em páginas e sites usando registros de acesso, registros de servidores, dados de log, etc. Essas informações são úteis para entender as ações dos usuários dentro de um sistema web, melhorar a estrutura de um sistema, prever e recomendar ações, entre outros.

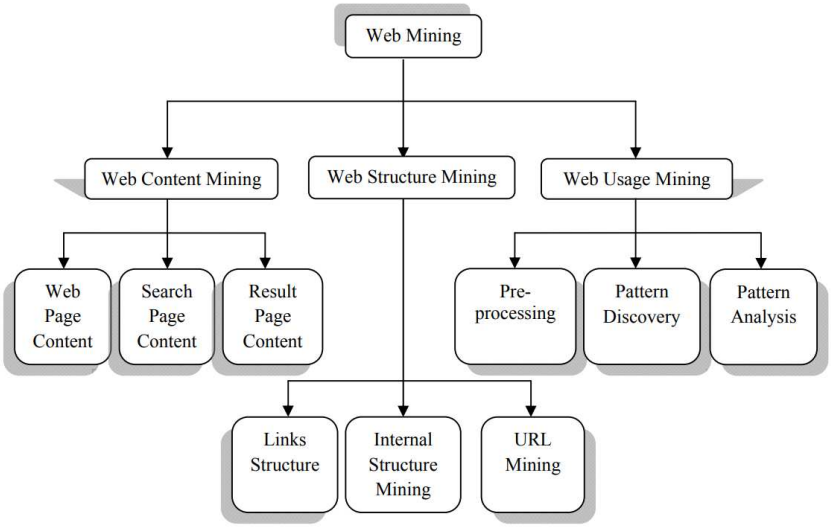

Figura 4 Taxonomia da Mineração de Dados na Web
Fonte: (SHARMA; SHRIVASTAVA; KUMAR, 2011)

3.2.1 Mineração de Conteúdo da Web

A área de interesse mais relevante do WDM para a realização deste trabalho é a Mineração de Conteúdo da Web (*Web Content Mining*, WCM). Segundo Kosala e Blockeel (2000) o WCM pode ser definido como o processo de extração de informações úteis e novos conhecimentos do conteúdo das páginas e sites disponíveis na World Wide Web. O conteúdo das páginas pode ser encontrado em diferentes formatos e tipos de dados, como texto, imagens, áudio, vídeo ou links. Muitas páginas e sites podem ser compostos por um único tipo de dados, exemplos disso são as páginas de enciclopédias, wikis ou blogs, que estão compostas predominantemente por texto. Em outros casos, as páginas podem conter vários objetos de conteúdo, cada um em um formato diferente.

O tratamento de objetos multimídia (imagens, áudio e vídeo) é uma tarefa não trivial, por isso o WCM geralmente se concentra na exploração, transformação e extração de informações de objetos de texto. As técnicas baseadas em texto são frequentemente usadas durante o processo de WCM, mesmo quando os objetos de conteúdo estão em formatos multimídia. Normalmente, os objetos multimídia são rotulados com textos, descrevendo o conteúdo desses objetos. Este processo de rotulagem pode ser realizado de forma manual, semiautomática ou automática. Depois, as técnicas de WCM são usadas sobre o texto desses rótulos. Por causa

disso, o processo de WCM é frequentemente associado à Mineração de Texto, mas essas subáreas do DM são essencialmente dois processos diferentes. A Mineração de Texto (*Text Mining*, TM) é definida como o conjunto de técnicas e métodos de DM para extrair padrões, tendências e novos conhecimentos de repositórios de documentos de texto (MENDOZA, 2011). Este processo é feito em repositórios de documentos de uso geral, enquanto o WCM se concentra exclusivamente em repositórios de documentos da Web. Apesar disso, a grande maioria das técnicas que estão presentes no TM podem ser aplicadas no processo de WCM.

O processo de WCM geralmente aborda 4 tarefas principais (MENDOZA, 2011):

3.2.1.1 Extração de informações específicas

Nesta tarefa é possível extrair informações de interesse de uma página web completa ou de qualquer um dos objetos que a compõem. No primeiro caso, é possível realizar um processo de *parsing* do código HTML da página.

Para extrair informações específicas de um objeto encontrado em uma página da Web, é realizado um processo de *scraping*, que consiste em extrair um fragmento de informação específica. Para ser eficaz, o processo de *scraping* deve ser capaz de identificar no texto da página e filtrar a descrição de um objeto. Se fosse necessário extrair informações relacionadas à uma imagem, vídeo ou áudio, as técnicas utilizadas devem ser capazes de determinar exatamente qual parte do texto completo da página está relacionada a esse objeto multimídia. (Fig. 5)

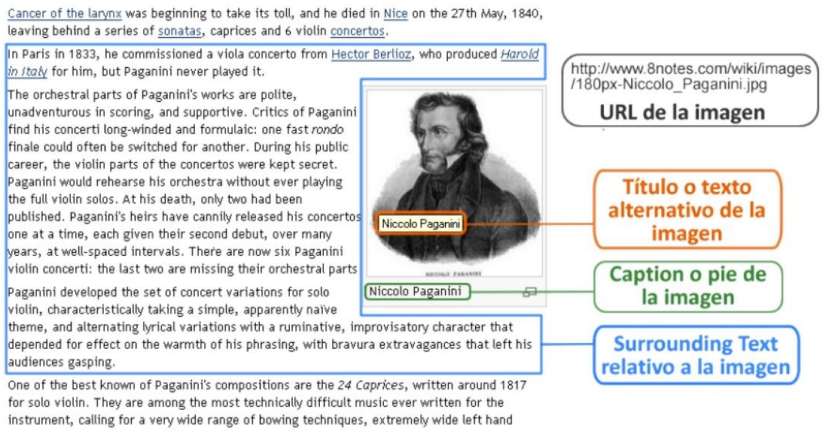

Figura 5 Extração de Informação de uma imagem na Web
Fonte: (MENDOZA, 2011)

As técnicas utilizadas no processo de *scraping* podem ser classificadas em 2 tipos (MENDOZA, 2011):

- **Técnicas de extração sem usar a estrutura da página**: Essas técnicas consistem em identificar o bloco de texto que descreve um determinado objeto, geralmente selecionando o texto ao redor do objeto. Também é comum usar atributos dos rótulos HTML do objeto para extrair informações relevantes.
- **Técnicas de extração usando a estrutura da página**: Essas técnicas geralmente são usadas para extrair informações de objetos que são representados em listas ou algum tipo de estrutura visual. São conhecidas pelo nome de *wrapper*. Para extrair informações deste tipo, é necessário criar regras que determinem o formato da estrutura e os campos que contêm informações relevantes. Essas regras são definidas manualmente por um usuário experiente, portanto, não podem ser reutilizadas em vários sites devido à grande variedade de apresentação e estruturação de recursos.

3.2.1.2 Agrupamento de documentos da web semelhantes

O agrupamento de documentos semelhantes é uma das tarefas mais comuns realizadas no WCM. Usando como base a hipótese de *clustering*, que afirma que documentos com conteúdo semelhante são relevantes para as mesmas consultas, é possível aumentar a eficiência e eficácia do processo de recuperação da informação.

Esta tarefa tem uma grande variedade de aplicações em WDM. Pode ser usada para dividir coleções de documentos em vários grupos semelhantes para distribuí-los em servidores diferentes, o que permite facilitar a organização de grandes coleções de documentos e melhora a escalabilidade dos motores de busca. Também pode ser aplicado na recomendação de grupos de artigos e documentos semelhantes para um usuário, permitindo que o usuário especialize sua consulta explorando cada um desses grupos.

3.2.1.3 Classificação de documentos da web

Outra tarefa bastante comum do WCM é a classificação de documentos em categorias previamente definidas. Essas categorias representam diferentes tópicos e os documentos devem ser rotulados para facilitar a organização e a recuperação de informações. Cada documento é analisado e classificado de acordo com seu conteúdo. Para criar um sistema de classificação de documentos, utiliza-se uma coleção de documentos de treinamento, cada documento contido nesta coleção foi previamente rotulado corretamente com a categoria correspondente. Esta coleção pode ser construída de forma manual ou semiautomática, usando algumas técnicas e estratégias de agrupamento da tarefa anterior. Um conjunto de validação e controle também é usado para poder avaliar a precisão do sistema. Os resultados do processo de classificação são geralmente validados por especialistas humanos para evitar a presença de falsos positivos.

Várias aplicações desta tarefa podem ser mencionadas no contexto do WDM. Uma dessas aplicações é a classificação de documentos para o gerenciamento automático de diretórios de informações da web, bem como a classificação de documentos para facilitar a manutenção de portais web de notícias.

3.2.1.4 Análise de sentimentos ou mineração de opiniões

O surgimento e massificação de blogs, fóruns e redes sociais tem permitido aos usuários expressar suas opiniões sobre os mais diversos temas. Análise de Sentimentos (*Sentiment Analysis*, SA) ou Mineração de Opiniões (*Opinion Mining*, OM) são os nomes dados ao processo de exploração, processamento e extração de conhecimento dessas fontes de informação (MENDOZA, 2011). O SA tem como objetivo principal a extração de padrões e conhecimento útil das opiniões dos usuários sobre temas específicos. É utilizado para determinar o grau de polaridade (positiva, negativa, neutra) das opiniões dos usuários em relação a diferentes entidades, tais como: produtos, serviços, pessoas, etc.

Embora a maioria das técnicas do WCM estejam relacionadas ao campo de Recuperação de Informação, as técnicas utilizadas em SA estão mais intimamente relacionadas ao campo de Processamento de Linguagem Natural. Isso ocorre porque para o SA certas palavras, artigos, conjunções e afirmações de polaridade (sim, não, talvez, etc.) são essenciais para determinar a intenção exata da opinião

de um usuário; já no campo Recuperação da Informação, muitas dessas palavras são descartadas porque não fornecem informações relevantes sobre as entidades presentes no texto.

Geralmente no SA, o problema a ser resolvido é abordado na perspectiva da classificação, identificando duas linhas fundamentais de estudo (MENDOZA, 2011):

- **Classificação de opinião no nível de documento**: Consiste, a partir de um documento completo, em inferir se a opinião expressa pelo autor sobre determinado assunto é neutra, positiva ou negativa.
- **Classificação de opinião no nível de frase**: Consiste em determinar se cada frase de um texto expressa uma visão neutra, positiva ou negativa de um tema específico.

Independentemente do nível de classificação usado ao realizar o SA para um determinado texto, as técnicas e ferramentas usadas para processar o texto são semelhantes. Os termos mais relevantes no texto, para SA, são as palavras usadas para expressar uma opinião sobre algo, por exemplo: ruim, bom, excelente, péssimo, gosto, odeio, etc.

4 L5P. PLATAFORMA E-GOV PARA A COLECTA E ANÁLISE DE DADOS

Para atingir os objetivos definidos anteriormente neste trabalho, optou-se por implementar um sistema de recuperação de informação que satisfaça os requisitos necessários para alimentar os processos de mineração e análise a utilizar. O sistema L5P usa várias das técnicas para coletar e analisar as informações descritas no capítulo anterior, na seção sobre Mineração de Dados na Web.

4.1 Metodologia

O processo de Desenvolvimento de Software é atualmente realizado seguindo 2 metodologias principais, o Modelo em Cascata (*Waterfall*) e o Desenvolvimento Ágil de Software (*Agile*).

Tradicionalmente, a maioria das empresas tem usado o Modelo em Cascata, que pode ser descrito como um ciclo de vida sequencial do projeto de desenvolvimento. Geralmente consiste em uma fase inicial de planejamento detalhado, nesta fase todos os aspectos do projeto são definidos, desenhados e documentados em grande detalhe. Em seguida, são determinadas as tarefas necessárias para poder terminar o projeto, e é estimado o tempo de desenvolvimento de cada tarefa, obtendo-se uma estimativa da duração total do projeto. Uma vez que os interessados (*stakeholders*) tenham aprovado o design, as tarefas a serem concluídas e a duração estimada, inicia-se o trabalho de implementação das tarefas pela equipe de desenvolvimento. Após a finalização desse trabalho, o resultado é entregue para a equipe de testes e garantia de qualidade, que valida o resultado alcançado. Finalmente, o produto é liberado para o uso por usuários finais.

Esta abordagem de trabalho tem certas fraquezas. O Modelo em Cascata é uma abordagem linear, embora existam variantes que incorporam a iteratividade ao processo, com uma separação clara entre suas diferentes etapas. Quando uma etapa termina, o resultado é passado para a próxima, cada etapa depende do trabalho realizado na anterior. Isso significa que se ocorrer alguma mudança durante

o desenvolvimento do projeto, isso pode gerar obstáculos e interrupções nas fases posteriores. Outra grande fraqueza é que a retroalimentação do usuário é obtida muito tarde no processo. Os clientes fornecem retroalimentação somente após a fase de teste, uma vez que o produto foi totalmente concluído. Se o cliente decidir que os requisitos vão mudar durante o curso de um projeto, os ajustes associados podem ser caros e demorados. Como o escopo do projeto, os requisitos, funcionalidades e tarefas a cumprir são definidos na primeira fase, podem surgir problemas posteriormente ao tentar adicionar novas funcionalidades que não sejam compatíveis com o design ou arquitetura inicialmente definida.

Devido a essas fraquezas, nos últimos anos tem havido um aumento no uso de metodologias de desenvolvimento de software iterativas, incrementais e evolutivas. Essas metodologias são conhecidas como Desenvolvimento Ágil de Software. Elas se concentram em dividir o trabalho de desenvolvimento em pequenos incrementos que minimizam a quantidade de planejamento e design iniciais. O principal elemento das metodologias ágeis é a iteração, cada iteração envolve uma equipe multifuncional trabalhando em todas as funções: planejamento, análise, design, programação e teste. O objetivo é ter uma versão entregável disponível, representativa de uma funcionalidade específica, ao final de cada iteração. Cada entrega é apresentada aos usuários finais, a retroalimentação obtida pode ser incorporada na próxima iteração.

4.1.1 Scrum

Este projeto foi desenvolvido seguindo uma metodologia de desenvolvimento ágil, utilizando o *framework Scrum*. *Scrum* é um dos *frameworks* ágeis mais populares e usados. Ele é usado por grandes e pequenas empresas, incluindo Yahoo!, Microsoft, Google, Lockheed Martin, Motorola, SAP, Cisco, GE, Capital One e a Reserva Federal dos EUA (SUTHERLAND; SCHWABER, 2007).

Scrum é um *framework* iterativo e incremental projetado para projetos de desenvolvimento de produtos e aplicativos. Como todas as metodologias ágeis, ele divide a estrutura do processo de desenvolvimento em iterações que chama de *Sprints*. Os *Sprints* são iterações de curto prazo (2 a 4 semanas) que ocorrem consecutivamente e terminam em uma determinada data, independentemente de o trabalho definido para aquela iteração ter sido concluído ou não. Cada iteração

encapsula todas as fases do ciclo de desenvolvimento de software: Requisitos, Planejamento, Design, Desenvolvimento, Teste e Implantação (Fig. 6).

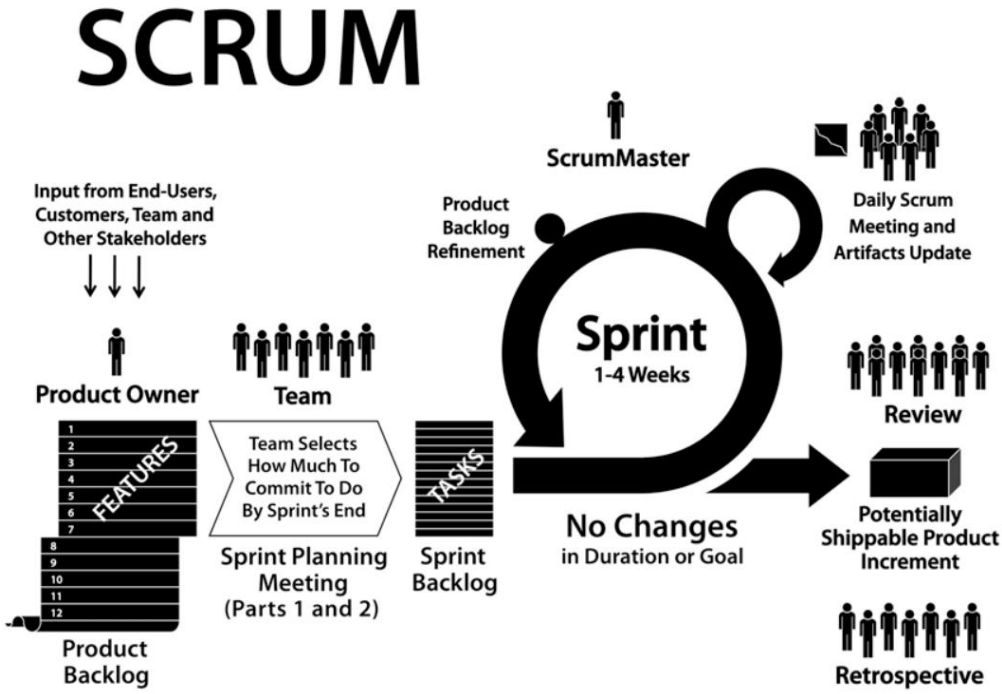

Figura 6 Processo de Scrum
Fonte: (SUTHERLAND; SCHWABER, 2007)

O processo determinado pelo *Scrum* define 3 funções principais, que compõem o *Equipe Scrum* (SUTHERLAND; SCHWABER, 2007):

- ***Product Owner*** (Proprietário(a) do produto): É responsável por identificar as funcionalidades e características do produto, criando uma lista priorizada a partir das funcionalidades identificadas e mantendo essa lista atualizada através das iterações. No caso de desenvolvimento de aplicativos internos, o *Product Owner* e o cliente são a mesma pessoa, em outros casos o *Product Owner* cumpre a função de um Gerente de Produto tradicional, com a diferença que, no *Scrum*, o *Product Owner* interage diretamente com membros da equipe de desenvolvimento e não depende de um Gerente de Projeto.
- **Equipe**: É responsável por desenvolver as funcionalidades definidas pelo *Product Owner*. É um grupo de pessoas com elevado nível de autonomia e responsabilidade, inclui todas as competências

necessárias para entregar um produto acabado em cada *Sprint* (multifuncional). A Equipe é auto-organizada, cada membro da equipe decide quais tarefas, dentro das definidas para cada iteração, vai realizar e qual é a melhor forma de fazê-las. Geralmente, a equipe é pequena, entre 5 e 10 pessoas e inclui diferentes perfis de habilidade: análise, programação, teste, design, documentação, etc. Grupos de trabalho maiores são divididos em várias Equipes de *Scrum* focados em diferentes funcionalidades do projeto.

- ***ScrumMaster*** (Mestre de *Scrum*): É responsável pela correta aplicação dos métodos e princípios *Scrum* durante o processo de desenvolvimento. Tem a responsabilidade de auxiliar o *Product Owner* e a Equipe de todas as maneiras possíveis para atingir o objetivo do projeto. O *ScrumMaster* não cumpre a tarefa de um Gerente de Projeto ou Gerente de Equipe, essas figuras não existem no Scrum, sua tarefa é apoiar a Equipe e certificar-se de que todas as práticas do Scrum sejam seguidas. O *ScrumMaster* tem como uma de suas tarefas a proteção da equipe de desenvolvimento e podem surgir situações em que as ações do *Product Owner* entrem em conflito com os interesses da Equipe, por exemplo, tentar adicionar mais tarefas a uma iteração em andamento. Portanto, o *ScrumMaster* e o *Product Owner* não podem ser a mesma pessoa.

4.1.1.1 Processo *Scrum*

A primeira etapa do processo Scrum é definir o *Backlog do Produto*. O *Backlog do Produto* pode ser definido como uma lista refinada e priorizada das funcionalidades e características do produto que representam a visão do *Product Owner* e as expectativas e interesses dos *stakeholders*. No caso de um novo produto, um refinamento inicial do *Backlog do Produto* deve ser executado antes do primeiro *Sprint*. Este processo pode demorar vários dias ou uma semana e envolve uma análise detalhada dos requisitos e uma estimativa de todos os elementos identificados para o primeiro lançamento do produto.

O *Backlog* pode incluir uma grande variedade de elementos como: funcionalidades para usuários, objetivos de engenharia, trabalhos de pesquisa, defeitos ou erros encontrados, etc. Devem ser definidos de forma clara e compreensível para todos os membros da equipe. Cada versão ou lançamento do produto precisará de um *Backlog de Lançamento*, que será um subconjunto do *Backlog do Produto* contendo apenas os elementos necessários para cada versão. O *Backlog do Produto* deve ser constantemente atualizado durante o processo de desenvolvimento do produto para refletir novas ideias e requisitos do cliente. Cada elemento tem uma estimativa de tempo, valor e riscos; com esses elementos o *Product Owner* pode priorizá-los de forma adequada.

4.1.1.2 Planejamento do *Sprint*

No início de cada iteração, são realizadas duas reuniões de Planejamento de Sprint. Na primeira reunião, os elementos de maior prioridade são analisados dentro do *Backlog do Produto* e os objetivos de cada um desses elementos são discutidos. Em seguida, é determinada a definição de "Concluído", que é o padrão que cada elemento deve atender para ser classificado como Concluído.

Na segunda reunião, a Equipe decide quais das tarefas, começando com a de maior prioridade, serão incluídas no *Sprint*, isso é feito levando em consideração a estimativa de cada tarefa, o número de membros da equipe e a duração do *Sprint*. Cada elemento selecionado do *Backlog* é dividido em tarefas individuais e adicionado a um *Backlog do Sprint*, essas são as tarefas que serão implementadas pela Equipe durante o *Sprint*. Uma vez que as tarefas do *Sprint* são definidas, nenhuma nova tarefa ou mudança pode ser adicionada, é preciso esperar pelo próximo *Sprint*. O *Sprint* pode ser encerrado antecipadamente se uma mudança de prioridades for necessária devido a circunstâncias externas. Nesse caso, novas reuniões de Planejamento devem ser realizadas e um novo *Sprint* iniciado.

4.1.1.3 *Scrum* diário

Durante o Sprint, são realizadas reuniões diárias conhecidas como o *Scrum Diário*. Essas reuniões incluem todos os membros da Equipe e devem ser de curta duração, inferior a 15 minutos. Cada membro da equipe deve responder a 3 perguntas:

1. O que você fez desde o último *Scrum Diário*?
2. O que você planeja terminar até o próximo *Scrum Diário*?
3. Existe algum obstáculo que esteja impedindo a conclusão da tarefa?

Os obstáculos são anotados e o *ScrumMaster* tem a tarefa de ajudar a equipe a resolvê-los.

4.1.1.4 Atualização do *Sprint*

Cada membro da Equipe deve atualizar, diariamente, o tempo estimado que falta para finalizar a tarefa que está realizando. Em seguida, a estimativa total da equipe é calculada e o *Gráfico de Evolução do Sprint* é atualizado (Fig. 7). Este gráfico mostra o tempo estimado necessário para concluir todas as tarefas ao longo do *Sprint*.

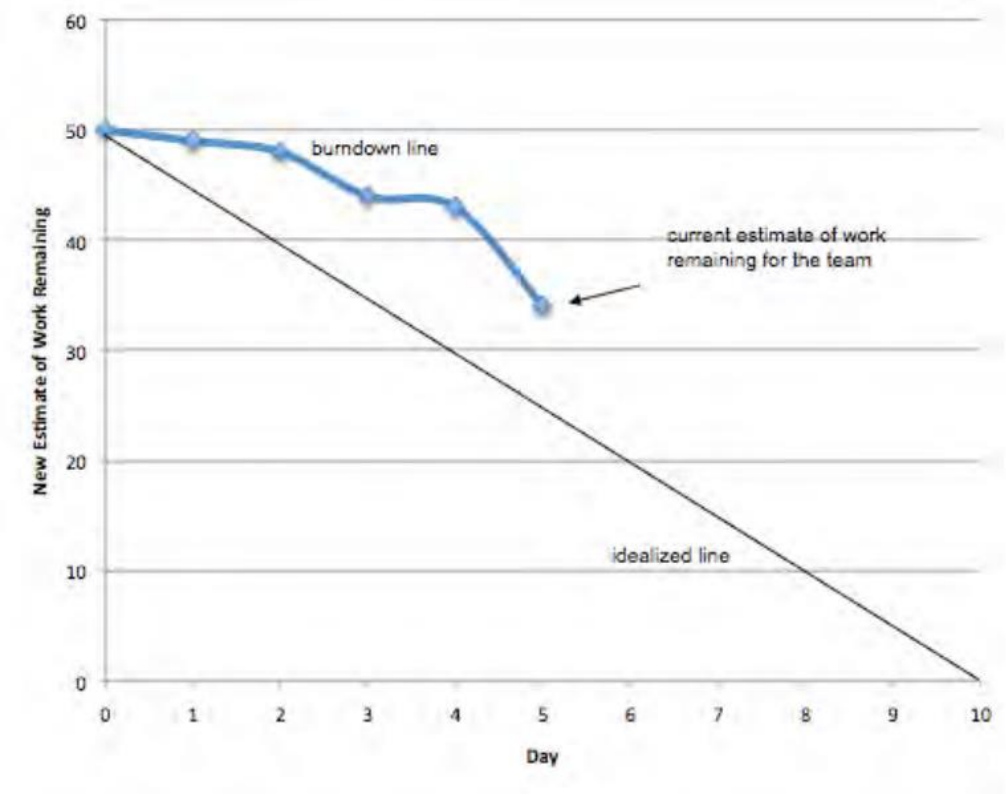

Figura 7 Gráfico de Evolução do Sprint
Fonte: (SUTHERLAND; SCHWABER, 2007)

Próximo ao final do *Sprint*, deve haver um refinamento de itens futuros no *Backlog do Produto*. Este refinamento consiste em identificar requisitos detalhados,

dividir elementos em elementos menores, estimar novos elementos ou reestimar elementos existentes.

4.1.1.5 Final do *Sprint*

Conforme mencionado acima, a duração do *Sprint* é fixa e não pode ser estendida. Quando o prazo limite chegar, o *Sprint* termina independentemente do status das tarefas definidas para aquele *Sprint*. Isso ajuda a melhorar a estimativa e o planejamento de cada *Sprint*.

4.1.1.6 Revisão do *Sprint*

Após o término do *Sprint*, uma reunião de *Revisão do Sprint* é realizada. Nesta reunião, os resultados da Equipe durante a iteração são apresentados e a retroalimentação do *Product Owner* e dos usuários é obtida. Em seguida, é realizada uma discussão entre os membros da Equipe para avaliar o processo de desenvolvimento e propor mudanças que possam beneficiar a produtividade.

4.1.1.7 Atualização do Projeto

Neste momento, a situação do projeto é atualizada, as tarefas concluídas são eliminadas do *Backlog de Lançamento* e do *Backlog do Produto*, novas tarefas são adicionadas e as estimativas são atualizadas. Em seguida, o Gráfico de Evolução de Lançamento é atualizado, é um gráfico semelhante ao *Gráfico de Evolução do Sprint*, mas que mostra o tempo restante para a conclusão da versão correspondente do produto.

4.1.1.8 Novo *Sprint*

Após a Revisão do Sprint, todos os membros do processo *Scrum* iniciam um novo *Sprint*. Geralmente, o novo *Sprint* começa no próximo dia útil após a *Revisão do Sprint*.

4.1.1.9 *Sprint* de Lançamento

Segundo *Scrum*, ao final de um *Sprint*, deve ter sido obtido um produto que pode ser potencialmente lançado. Isso significa que, teoricamente, um produto pode ser lançado após o término de uma *Revisão do Sprint*, mas isso nem sempre acontece na prática. Portanto, pode ser necessário que exista um *Sprint de Lançamento* para concluir o trabalho necessário para terminar o produto.

4.2 Seleção de Tecnologias

Antes de iniciar a implementação de ferramentas que permitam coletar informações de fontes encontradas na Web ou outros repositórios remotos, foi necessário decidir as tecnologias a serem utilizadas para o desenvolvimento do sistema. O L5P foi desenhado como um sistema web, para facilitar o acesso remoto e utilizar tecnologias de desenvolvimento consolidadas ao longo dos anos. Foi decidido usar tecnologias de código aberto para aproveitar as vantagens do grande número de soluções existentes e das grandes comunidades que essas tecnologias reúnem. Graças a isso foi possível aumentar a eficiência e eficácia dos desenvolvedores vinculados ao projeto.

Para armazenar os dados coletados pelo sistema, foi utilizado o sistema de gerenciamento de Banco de Dados (*Database*, DB) **PostgreSQL**. É um sistema de gerenciamento de DB multiplataforma, gratuito e de código aberto, com alta confiabilidade, integridade de dados, funcionalidades e desempenho. É uma opção altamente recomendada para aplicativos que usam técnicas de DM ou ferramentas de relatórios. **PostgreSQL** cumpre integralmente o padrão ACID, que garante a integridade dos dados nas mais diversas situações adversas. Também tem alta, mas não completa, conformidade com o padrão SQL, que define as especificações da linguagem de gerenciamento de DB SQL (POSTGRESQL, 2020).

O servidor **Apache HTTP Server** foi usado como servidor web para hospedar o sistema. É um servidor web de código aberto e multiplataforma, desenvolvido e mantido pela *Apache Software Foundation*. Suporta uma ampla variedade de linguagens de programação do lado do servidor, como Perl, Python e PHP. É possível hospedar vários sites na mesma instalação do Apache, usando a funcionalidade de Hosts Virtuais. É um dos servidores web mais usados, de acordo com a Netcraft (2020), em agosto de 2020, 26,62% dos sites ativos usavam o **Apache** como servidor web. É bem conhecido por sua confiabilidade, desempenho e grande comunidade de desenvolvedores.

A linguagem de programação do lado do servidor **PHP** foi usada para desenvolver o sistema. É uma linguagem de *scripting* de propósito geral especialmente adequada para o desenvolvimento de servidores web. Pode ser executado na maioria dos servidores web, como Nginx e Apache, e funciona em sistemas operacionais Windows e sistemas baseados em Unix, como Linux e Mac OS. Muitos provedores de hospedagem na web oferecem suporte à linguagem **PHP** por meio da arquitetura LAMP (Linux, Apache, MySQL e **PHP**). É uma linguagem de programação de código aberto, compatível com quase todos os sistemas de administração de DB e possui uma grande comunidade de desenvolvedores; segundo o GitHub (2020), o **PHP** é a quarta linguagem mais usada nesta plataforma. **PHP** é usado nos conhecidos e populares Sistemas de Gerenciamento de Conteúdo (*Content Management Systems*, CMS) Wordpress, Joomla, Drupal, Moodle e muitos outros. Em novembro de 2020, 78,9% de todos os sites conhecidos usavam **PHP** como linguagem de programação do servidor (W3TECHSa, 2020).

Também foi usado o *framework* de desenvolvimento de aplicações web **CakePHP**. Este *framework* foi selecionado porque a equipe de desenvolvimento possui experiência prévia e grande familiaridade com ele, o que aumenta a produtividade. É um *framework* que usa conceitos de engenharia de software e padrões de projeto de software bem conhecidos, como o padrão de projeto Modelo-Visão-Controle. (*Model-View-Controller*, MVC). Possui diversas funcionalidades que facilitam o rápido desenvolvimento de protótipos e funcionalidades, como (CAKEPHP, 2020):

- **CakePHP Bake**: Uma ferramenta de geração de código que permite gerar layouts de interface, e código de Criação, Leitura, Atualização e Exclusão (*Create, Read, Update, Delete*; CRUD) a partir de uma estrutura de DB previamente definida.
- **CakePHP ORM**: Um sistema de abstração para comunicação com sistemas de gerenciamento de DB. Permite a execução de consultas SQL por meio de código de alto nível que é traduzido para cada sistema de DB específico.
- **Redirecionamento avançado**: Permite redirecionamento HTTP complexo, é possível mapear requisições de entrada para funções específicas em Controles ou Visões.

- **Paginação Automática**: Facilita a apresentação de listagens de dados paginados com mínimo esforço.
- **Plugins**: Possui um sistema de plugins que permite a reutilização e distribuição de funcionalidades através da combinação de Modelos, Controles e Visões.

Para facilitar o desenvolvimento de interfaces de usuário, optou-se por usar a biblioteca de JavaScript **JQuery**. É uma biblioteca de código aberto que fornece funcionalidades de modificação da árvore do DOM HTML e a manipulação de eventos. É uma das bibliotecas JavaScript mais usadas. Em dezembro de 2020, **JQuery** foi usado em 77,1% de todos os sites conhecidos (W3TECHSb, 2020). **JQuery** é uma biblioteca leve que oferece ótimo desempenho, oferece suporte para uma ampla variedade de navegadores da web, o que facilita a compatibilidade de código em diferentes navegadores e é facilmente extensível com diferentes plugins. Sua principal utilidade é a considerável simplificação do uso do JavaScript, várias tarefas que requerem muitas linhas de código podem ser executadas com **JQuery** em uma única linha.

4.3 Desenvolvimento

O objetivo do processo de desenvolvimento do sistema L5P foi criar um sistema de recuperação de informação usando tecnologias de desenvolvimento web. Foi projetado para realizar as duas primeiras tarefas descritas por Etzioni (1996) e Kosala e Blockeel (2000) como parte da definição do WDM: **Descoberta de recursos** e **Extração da informação**. Portanto, as principais funcionalidades que foram implementadas no sistema são: a coleta de informações de DB remotos, arquivos remotos, páginas e sites, a consolidação dos dados coletados em um único repositório, o processo de limpeza de dados e a visualização das informações coletadas. Também foram implementadas funcionalidades de autenticação, autorização e gerenciamento de usuários.

4.3.1 Seleção de fontes

O primeiro passo para iniciar o desenvolvimento do sistema foi definir as fontes de informação que alimentariam o sistema. Uma extensa investigação das fontes disponíveis na Internet foi realizada. As fontes foram priorizadas de acordo com vários parâmetros, por exemplo: se é uma fonte oficial fornecida por uma instituição governamental, se fornece serviços de consulta na web, arquivos ou a informação está no conteúdo das páginas, o formato dos serviços web ou arquivos, se o conteúdo das páginas for estruturado, semiestruturado ou não estruturado, etc. Finalmente, as seguintes fontes de informação foram selecionadas:

- **O site da Câmara dos Deputados** (https://www.camara.leg.br/): É o site oficial da Câmara dos Deputados do Brasil. Oferece uma ampla variedade de informações sobre a atuação dos Deputados Federais. É possível acessar a agenda das votações, o texto das propostas em tramitação, as leis aprovadas, a programação dos eventos, a lista de presenças dos deputados e o resultado das votações. Todos os discursos dos deputados são gravados e publicados. Também é possível ter acesso a outras informações públicas, como contratos de compra, editais de licitação, obras em andamento, dados abertos, provisão de contas de despesas parlamentares, entre outras. Uma API REST de acesso está disponível para consultas, também há muita informação no conteúdo das páginas do site. Esta fonte foi utilizada para coletar os seguintes dados: dados gerais do deputado (nome, profissão, data de nascimento, legislatura, histórico parlamentar, partido, formação, telefone e e-mail parlamentar, etc.), discursos e despesas.
- **O site do Senado Federal** (https://www12.senado.leg.br/): É o site oficial do Senado Federal do Brasil. Oferece informações semelhantes às publicadas pelo site da Câmara dos Deputados. Fornece uma API REST para consultas, muitas informações estão disponíveis no conteúdo das páginas do site, também fornece arquivos CSV com informações sobre os gastos parlamentares. Esta fonte foi utilizada para coletar os seguintes dados: dados gerais do senador (nome, profissão, data de nascimento, legislatura, histórico parlamentar,

partido, formação, telefone e e-mail parlamentar, etc.), discursos e despesas.

- **O site de Divulgação de Candidaturas e Contas do Tribunal Superior Eleitoral** (https://divulgacandcontas.tse.jus.br/): É o site responsável por publicar as informações divulgadas pelos candidatos às eleições em todos os níveis da política brasileira. Os dados dos candidatos, como dados pessoais, bens declarados e informações sobre eleições anteriores em que tenham participado, são publicados neste site. Também é possível acessar os dados das campanhas eleitorais, como doações, despesas durante a campanha e extratos bancários. As informações estão disponíveis no conteúdo das páginas do site, arquivos CSV também estão disponíveis, distribuídos em arquivos ZIP compactados. Esta fonte foi usada para coletar os seguintes dados: doações de campanha, despesas de campanha, ativos declarados durante a campanha e histórico eleitoral.
- **O Portal Brasileiro de Dados Abertos** (https://dados.gov.br/): É uma ferramenta disponibilizada pelo governo para o acesso das pessoas às informações públicas. O portal tem como objetivo fornecer dados sobre os mais diversos temas da administração pública. Contém apenas dados abertos; dados que contêm alguma restrição de acesso, como aqueles derivados de sigilo ou privacidade, não estão disponíveis. O portal fornece uma coleção de todos os dados publicados por órgãos governamentais, permitindo pesquisas dentro da coleção. Eles são publicados de acordo com cronogramas previamente definidos por cada instituição pública. Eles são fornecidos em diferentes formatos, como CSV, JSON e XML. Foi utilizado como complemento aos dados coletados em outras fontes.
- **Atlas Político** (http://atlaspolitico.com.br/): É uma ferramenta online, de acesso gratuito à população. É um site privado que tem como objetivo acompanhar o desempenho dos políticos brasileiros. É inspirado por iniciativas de sucesso, como http://www.opencongress.org/ dos Estados Unidos e https://www.votewatch.eu/ da União Europeia. Este site usa várias fontes de informações oficiais do governo para exibir informações consolidadas sobre os políticos brasileiros. Esta fonte foi

utilizada para coletar os seguintes dados: processos judiciais. Todos os dados coletados desta fonte são validados após serem coletados e os registros incorretos são removidos.

- **CepespData** (https://cepespdata.io/): É uma plataforma de acesso aos dados eleitorais brasileiros. Foi desenvolvido pelo Centro de Política e Economia do Setor Público (CEPESP) da Fundação Getúlio Vargas (FGV). Esta plataforma disponibiliza na íntegra os dados divulgados pelo Tribunal Superior Eleitoral, consolidando-os em um único repositório. A consistência e integridade dos dados são mantidas, nenhuma informação é alterada ou modificada. Fornece vários métodos para obter as informações: um serviço web de consulta REST e APIs para as linguagens de programação R e Python. Esta fonte foi utilizada para coletar os seguintes dados: doações de campanha e despesas durante a campanha. Foi utilizado como complemento aos dados coletados no site de Divulgação de Candidaturas e Contas do Tribunal Superior Eleitoral.

Ao finalizar o processo de identificação das fontes de informação, foram implementadas as funcionalidades que permitem os processos de coleta e limpeza da informação e consolidação dos dados

4.3.2 Coleta de informações

Durante o processo de coleta de informações, todas as fontes foram analisadas e várias estratégias foram desenvolvidas para coletar os dados necessários dependendo dos formatos em que foram publicados.

4.3.2.1 Serviços Web

Conforme mencionado acima, várias das fontes de informação selecionadas disponibilizam serviços web de consulta em formato JSON ou XML. É o caso do site da Câmara dos Deputados, do site do Senado Federal, do Portal Brasileiro de Dados Abertos e do CepespData. Cada um desses serviços retorna os dados em uma estrutura específica, portanto não é possível criar uma função genérica para

consultá-los. Funções personalizadas foram implementadas para consultar esses serviços, apenas os dados relevantes foram armazenados.

O processo de implementação inicia-se com a consulta da documentação do serviço. Na documentação são definidas a URL do serviço web, as funções disponíveis para consulta, a estrutura dos dados retornados, bem como as restrições existentes para a utilização do serviço, tais como, o limite de registros retornados, o limite de solicitações por unidade de tempo, etc. Imediatamente, começa a codificação das funções de coleta de dados. Primeiramente é feita uma requisição HTTP ou HTTPS na URL do serviço, especificando a função a ser acessada, em seguida, são selecionados os campos relevantes do registro retornado pelo serviço, e são armazenados em objetos de programação temporários para sua posterior limpeza e consolidação no DB. A figura 8 mostra o código para coletar os dados gerais de um deputado por meio da API do site da Câmara dos Deputados.

```
//Requisição http à url do serviço (https://dadosabertos.camara.leg.br/api/v2/), especificando
a função (deputados) e os parâmetros ($congressMan['id'])
$curlDeputado2 = curl_init();
curl_setopt($curlDeputado2, CURLOPT_URL,
'https://dadosabertos.camara.leg.br/api/v2/deputados/' . $congressMan['id']);
curl_setopt($curlDeputado2, CURLOPT_CONNECTTIMEOUT, 5);
curl_setopt($curlDeputado2, CURLOPT_RETURNTRANSFER, 1);
curl_setopt($curlDeputado2, CURLOPT_FAILONERROR, true);
curl_setopt($curlDeputado2, CURLOPT_USERAGENT, 'L5');
$queryDeputado2 = curl_exec($curlDeputado2);
$httpcode = curl_getinfo($curlDeputado2, CURLINFO_HTTP_CODE);

//Validação do resultado da consulta
if ($queryDeputado2 !== false && $httpcode === 200) {

    //Seleção de dados relevantes
    $respDeputado2 = json_decode($queryDeputado2);
    $congressMan['name'] = $respDeputado2->dados->nomeCivil;
    $congressMan['parliamentaryName'] = $respDeputado2->dados->ultimoStatus->nome;
    $congressMan['legislature'] = $respDeputado2->dados->ultimoStatus->idLegislatura;
    $congressMan['condition'] = $respDeputado2->dados->ultimoStatus->condicaoEleitoral;
    $congressMan['telephone'] = $respDeputado2->dados->ultimoStatus->gabinete->telefone;
    $congressMan['email'] = $respDeputado2->dados->ultimoStatus->gabinete->email;
    $congressMan['sex'] = $respDeputado2->dados->sexo;
    $congressMan['cabinet'] = $respDeputado2->dados->ultimoStatus->gabinete->nome;
    $congressMan['annex'] = $respDeputado2->dados->ultimoStatus->gabinete->predio;
    $congressMan['birthState'] = $respDeputado2->dados->ufNascimento;
    $congressMan['birthMunicipality'] = $respDeputado2->dados->municipioNascimento;
    $congressMan['cpf'] = $respDeputado2->dados->cpf;
    $congressMan['education'] = $respDeputado2->dados->escolaridade;
    $congressMan['website'] = $respDeputado2->dados->urlWebsite;
    $congressMan['socialNetworks'] = $respDeputado2->dados->redeSocial;
    $congressMan['birthday'] = $respDeputado2->dados->dataNascimento;
    $congressMan['deathday'] = $respDeputado2->dados->dataFalecimento;
} else {
    //Validação de erros
    $url = curl_getinfo($curlDeputado2)['url'];
    if ($errno = curl_errno($curlDeputado2)) {
        $error message = curl_strerror($errno);
        Log::write('debug', "cURL error ({$errno}):\n {$error_message} :\n {$url}");
    }
    Log::write('debug', "Error:\n {$url}");
    $error = true;
}
curl_close($curlDeputado2);
```

Figura 8 Coleta de dados gerais de um deputado por meio da API do site da Câmara dos Deputados
Fonte: Elaborado pelo autor

4.3.2.2 Arquivos CSV

Algumas das fontes selecionadas fornecem os dados ao público em planilhas no formato CSV. É o caso do Portal Brasileiro de Dados Abertos e do site do Senado Federal. Assim como os serviços web mencionados acima, cada planilha possui uma estrutura específica, por isso foi necessário criar funções personalizadas para obter as planilhas e posteriormente processá-las.

O processo de implementação inicia-se com a análise de cada planilha. Para isso, são baixados os arquivos CSV e identificada a estrutura da planilha e os dados relevantes a serem extraídos. Imediatamente, começa o processo de codificação das funções de coleta de dados. Primeiramente, é feita uma requisição HTTP ou HTTPS para a URL onde a planilha é publicada e o arquivo é salvo em uma pasta, em seguida, um processo de leitura da planilha é iniciado linha por linha. Os campos relevantes do registro são selecionados e armazenados em objetos de programação temporários para sua posterior limpeza e consolidação no DB. A figura 9 mostra o código para coletar as despesas de um senador por meio das planilhas publicadas no site do Senado Federal.

```
<?php
//Loop para pegar as informações de todos os anos desde 2008
$year = (int) date("Y");
for($i = 2008; $i <= $year; $i++){
    $atoName = "Despesas do Senado " . $i;
    $filename = $atoName . '.csv';

    //Link da planilha do ano especificado
    $link = 'http://www.senado.leg.br/transparencia/LAI/verba/' . $i . '.csv';
    $loggedIn = $this->Auth->user();
    $targetFolder = WWW_ROOT . "arquivos" . DS . $loggedIn['client_id'] . DS .
    $expensesAto->idato . DS;
    $targetFile = WWW_ROOT . "arquivos" . DS . $loggedIn['client_id'] . DS .
    $expensesAto->idato . DS . $filename;
    if (!file_exists($targetFile) || filesize($targetFile) != $this-
>curl_get_file_size($link)) {
        //Requisição a url da planilha
        $curlExpenses = curl_init();
        curl_setopt($curlExpenses, CURLOPT_URL, $link);
        curl_setopt($curlExpenses, CURLOPT_CONNECTTIMEOUT, 5);
        curl_setopt($curlExpenses, CURLOPT_RETURNTRANSFER, 1);
        curl_setopt($curlExpenses, CURLOPT_FAILONERROR, true);
        curl_setopt($curlExpenses, CURLOPT_USERAGENT, 'L5');
        $queryExpenses = curl_exec($curlExpenses);
        $httpcode = curl_getinfo($curlExpenses, CURLINFO_HTTP_CODE);

        //Validação do arquivo retornado
        if ($queryExpenses !== false && $httpcode == 200) {
            if (!file_exists($targetFolder)) {
                mkdir($targetFolder, 0777, true);
            }
```

```
            //Arquivo salvo
            $fp = fopen($targetFile, 'w');
            fwrite($fp, $queryExpenses);
            fclose($fp);
        } else {
            //Validação de erros
            $url = curl_getinfo($curlExpenses)['url'];
            if ($errno = curl_errno($curlExpenses)) {
                $error message = curl_strerror($errno);
                Log::write('debug', "cURL error ({$errno}):\n {$error_message} :\n {$url}");
            }
            Log::write('debug', "Error:\n {$url}");
            $error = true;
        }
        curl_close($curlExpenses);
    }
    if(file_exists($targetFile)){
        //Processo de leitura do arquivo
        if (($handle = fopen($targetFile, "r")) !== FALSE) {
            $j = 0;
            while (($data = fgetcsv($handle, 0, ";")) !== FALSE) {
                if ($j > 1) {
                    //Seleção dos campos relevantes
                    if (isset($data[2])) {
                        if (mb_strtolower(utf8_encode($data[2])) ===
                        mb_strtolower($senator['parliamentaryName'])) {
                            $expense['year'] = utf8_encode($data[0]);
                            $expense['month'] = utf8_encode($data[1]);
                            $expense['senator'] = utf8_encode($data[2]);
                            $expense['resourceType'] = utf8_encode($data[3]);
                            $expense['cpf/cnpj'] = utf8_encode($data[4]);
                            $expense['provider'] = utf8_encode($data[5]);
                            $expense['document'] = utf8_encode($data[6]);
                            $expense['date'] = utf8_encode($data[7]);
                            $expense['details'] = utf8_encode($data[8]);
                            $expense['value'] = floatval(str_replace(',', '.',
                            str_replace('.', '', utf8_encode($data[9]))));
                        }
                    }
                }
                $j++;
            }
            fclose($handle);
        }
    }
}
```

Figura 9 Coleta das despesas de um senador por meio das planilhas publicadas no site do Senado Federal
Fonte: Elaborado pelo autor

4.3.2.3 Conteúdo da página

Algumas das fontes de informação não fornecem métodos de consulta de dados estruturados ou algumas informações específicas não fazem parte dos serviços web ou planilhas. Nesse caso, é necessário obter os dados diretamente do conteúdo da página. O site de Divulgação de Candidaturas e Contas do Tribunal Superior Eleitoral e Atlas Político não possuem métodos estruturados de consulta, e o site da Câmara dos Deputados tem informações que só estão disponíveis no conteúdo das páginas. Como cada site contém dados diferentes e são apresentados em estruturas diferentes, *wrappers* personalizados foram implementados para extrair as informações do conteúdo. Esses *wrappers* têm a função de identificar

informações relevantes dentro do conteúdo HTML das páginas analisadas e extrair as informações que podem ser úteis para análises posteriores.

O processo de implementação começa com a análise de cada página. Para isso é utilizado um navegador web e a função de inspecionar o código, em seguida são identificados os rótulos HTML onde se encontram as informações a serem coletadas. É necessário encontrar um elemento próximo aos dados que contenha um atributo único (identificador, classe, nome, etc.) para poder acessar os dados relevantes sem ter que recorrer todo o conteúdo da página (Fig. 10). Assim que este elemento for encontrado, é iniciado o processo de codificação das funções de coleta de dados. Primeiro, e feita uma requisição HTTP ou HTTPS para a URL da página, essa solicitação retornará o código HTML. Em seguida, uma busca é realizada até que o elemento com um atributo único mencionado acima seja encontrado e sua estrutura é percorrida até que a informação a ser coletada seja alcançada. Finalmente, os dados são armazenados em objetos de programação temporários para posterior limpeza e consolidação no DB. A figura 11 mostra o código para coletar os discursos de um deputado a partir da estrutura das páginas do site da Câmara dos Deputados.

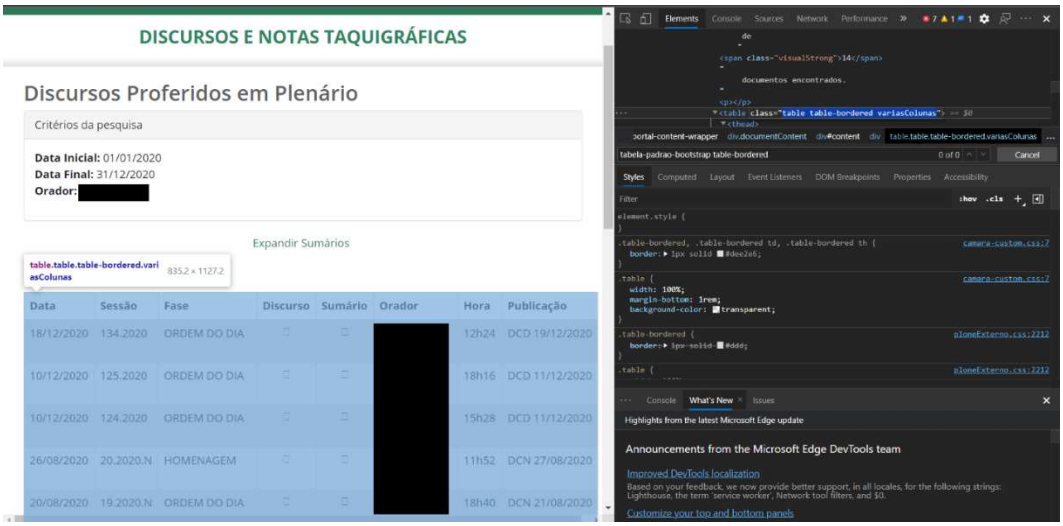

Figura 10 Identificando elemento com atributo único (classe)
Fonte: Elaborado pelo autor

```php
<?php
//Requisição http à url da página dos Discursos de um Deputado
$curlSpeeches = curl_init();
curl_setopt($curlSpeeches, CURLOPT_URL,
'https://www.camara.leg.br/internet/sitaqweb/DiscursosDeputado.asp?txOrador=' .
str_replace(" ", "+", $congressMan['parliamentaryName']) .
'&Campoordenacao=dtSessao&tipoordenacao=DESC&Pagesize=1000&txUF=' . $congressMan['state']);
```

```php
curl_setopt($curlSpeeches, CURLOPT_CONNECTTIMEOUT, 5);
curl_setopt($curlSpeeches, CURLOPT_RETURNTRANSFER, 1);
curl_setopt($curlSpeeches, CURLOPT_FAILONERROR, true);
curl_setopt($curlSpeeches, CURLOPT_USERAGENT, 'L5');
$querySpeeches = curl_exec($curlSpeeches);
$httpcode = curl_getinfo($curlSpeeches, CURLINFO_HTTP_CODE);

//Validação do resultado
if ($querySpeeches !== false && $httpcode === 200) {
    $congressMan['speeches'] = [];
    $dom = new DOMDocument();
    $dom->validateOnParse = true;
    libxml_use_internal_errors(true);
    $dom->loadHTML($querySpeeches);
    libxml_use_internal_errors(false);

    $xpath = new DOMXPath($dom);

    //Pesquisa do elemento com atributo único
    $speechesTable = $xpath->query("//*[@class= table table-bordered variasColunas']")[0];


    if ($speechesTable !== null) {

        //Recorrendo a estrutura do elemento
        $speechesLinks = $xpath->query(".//a", $speechesTable);
        foreach ($speechesLinks as $speechLink) {
            //Seleção da informação
            $speech['link'] = 'https://www.camara.leg.br/internet/sitaqweb/' .
            $speechLink->getAttribute('href');
            $speech['name'] = $congressMan['name'] . ' na sessão ' .
            trim($speechLink->parentNode->parentNode->childNodes[2]->nodeValue) . ' ' .
            $speechLink->parentNode->parentNode->childNodes[0]->nodeValue;
            $speech['date'] = $speechLink->parentNode->parentNode->childNodes[0]->nodeValue;
            $this->saveSpeechData($speech, $pesfis, $congressMan);
        }
    }
} else {
    //Validação de erros
    $url = curl_getinfo($curlSpeeches)['url'];
    if ($errno = curl_errno($curlSpeeches)) {
        $error_message = curl_strerror($errno);
        Log::write('debug', "cURL error ({$errno}):\n {$error_message} :\n {$url}");
    }
    Log::write('debug', "Error:\n {$url}");
    $error = true;
}
curl_close($curlSpeeches);
```

Figura 11 Coleta dos discursos de um deputado a partir da estrutura das páginas do site da Câmara dos Deputados
Fonte: Elaborado pelo autor

4.3.3 Limpeza e consolidação dos dados

Como a grande maioria das fontes de informação utilizadas são fontes oficiais do governo, e existem padrões e formatos definidos para o armazenamento de dados, não foi necessário realizar um processo de limpeza muito profundo. A maioria dos dados coletados está disponível no mesmo formato, com algumas exceções. Essas exceções são as datas e o CPF e CNPJ.

No caso das datas, estão disponíveis em vários formatos. Algumas fontes armazenam a data no formato *d/m/Y*, outros armazenam a data no formato ISO *Y-m-d H:i:s*, enquanto outros armazenam os campos de dia, mês e ano separadamente.

No L5P, todas as datas foram armazenadas no formato ISO *Y-m-d H:i:s*, que é o formato padrão do PostgreSQL. Para a exibição das datas, foi utilizado o formato *d/m/Y*.

No caso do CPF e CNPJ, algumas fontes armazenam esses dados apenas com caracteres numéricos, 11 caracteres no caso do CPF e 14 caracteres no caso do CNPJ. Enquanto outras fontes os armazenam com os caracteres separadores, por exemplo 000.000.000-0 para CPF e 00.000.000/0001-00 para CNPJ. No L5P, todos os CPF e CNPJ são armazenados sem os caracteres separadores, apenas os caracteres numéricos são armazenados. Caracteres de separação são adicionados ao visualizar os dados.

Após a conclusão do processo de limpeza de dados, todas as informações coletadas foram estruturadas e consolidadas em um único repositório. Os dados foram estruturados nas seguintes entidades, que representam tabelas no DB:

- **Pessoa Física**: Esta entidade representa uma pessoa física. Possui os seguintes atributos: nome, rótulo, sexo, data de nascimento e CPF. Nesta entidade foram armazenados os dados dos Deputados, Senadores e doadores que são pessoas físicas.
- **Pessoa Jurídica**: Esta entidade representa uma pessoa jurídica. Possui os seguintes atributos: nome, rótulo, lugar, CNPJ e data de abertura. Nesta entidade foram armazenados os dados dos provedores das despesas e dos doadores que são pessoas jurídicas.
- **Fato Social**: Esta entidade representa um evento ou fato. Possui os seguintes atributos: nome, rótulo, data e descrição. Nesta entidade foram armazenados os dados das campanhas eleitorais, processos judiciais, entre outros.
- **Ato**: Esta entidade representa um documento. Possui os seguintes atributos: nome, rótulo, data, link do documento e arquivo do documento. Nesta entidade foram armazenados os dados dos discursos e as planilhas de despesas e doações.
- **Lugar**: Esta entidade representa um lugar. Possui os seguintes atributos: nome, rótulo, latitude e longitude. Nesta entidade foram

armazenados os dados dos Estados, Cidades e lugares presentes nos dados.

- **Medida**: Esta entidade representa uma medida. Possui os seguintes atributos: nome, rótulo, valor, data e lugar. Nesta entidade foram armazenados os dados de doações, bens e despesas.
- **Coisa**: Esta entidade representa uma coisa. Possui os seguintes atributos: nome e rótulo. Nesta entidade foram armazenados todos os dados que não podem ser categorizados nas outras entidades.

Com essas entidades e seus relacionamentos, é possível representar qualquer entidade conhecida. Por isso é uma solução genérica que permite a adaptação a qualquer tipo de dado que precise ser salvo. Cada uma dessas entidades tem um relacionamento muitos-para-muitos com qualquer uma das outras entidades, bem como consigo mesma (Fig. 12).

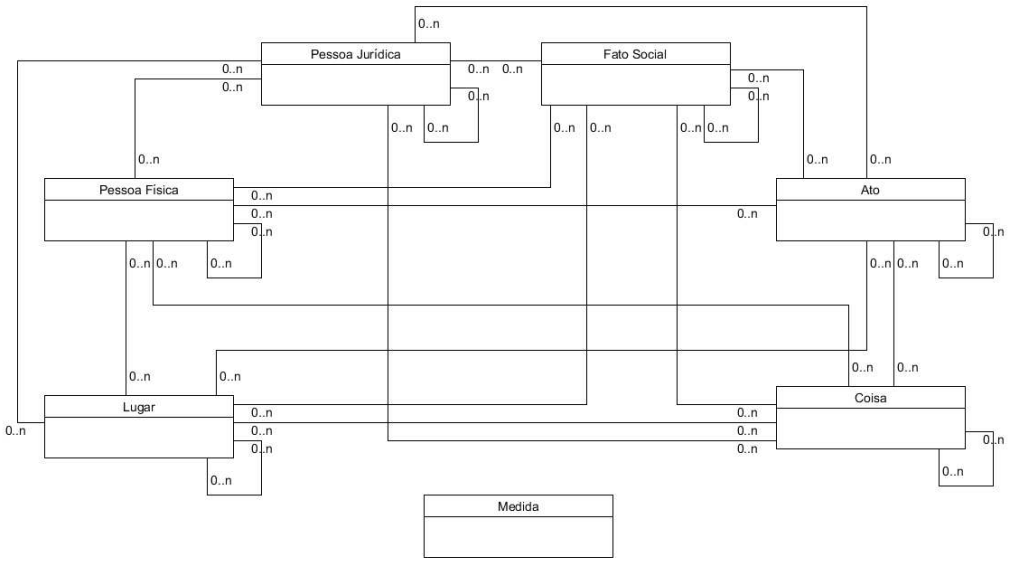

Figura 12 Entidades do Sistema
Fonte: Elaborado pelo autor

4.3.4 Segurança

Para garantir que pessoas não autorizadas não possam acessar os dados armazenados, várias medidas de segurança foram implementadas.

4.3.4.1 Autenticação

Um sistema de autenticação foi implementado para acessar o sistema. O usuário pode acessar usando um nome de usuário ou e-mail e senha. No momento não há restrições para a criação da senha, mas está prevista a adição de restrições de comprimento (mínimo 8 caracteres), bem como tipos de caracteres obrigatórios (minúsculas, maiúsculas, números e símbolos). Todas as senhas são criptografadas usando o algoritmo de criptografia de senha padrão do *CakePHP*, o algoritmo *Blowfish* que usa uma chave de 128 bits.

4.3.4.2 Autorização

Também foi implementado um sistema de autorização de acesso às funcionalidades. Cada usuário pertence a uma função dentro do sistema e cada função tem acesso a diferentes funcionalidades. Existem 3 funções, Usuário, Operador e Administrador. O Usuário só tem acesso à visualização de um subconjunto dos dados. O Operador tem acesso à visualização do resto dos dados, exceto os dados dos usuários, e às ferramentas para coletar os dados. O Administrador tem acesso para visualizar e editar todos os dados, também pode criar, modificar ou excluir usuários. Esses níveis de acesso são estáticos, as permissões das funções não podem ser facilmente modificadas sem modificar o código do aplicativo. A adição de um sistema de Listas de Controle de Acesso (*Access Control Lists,* ACL) está planejada, para aumentar a granularidade das permissões e permitir a modificação de permissões por meio de uma interface visual.

4.3.4.3 Ameaças comuns

O *framework CakePHP* implementa diversas soluções para os tipos mais comuns de ameaças da web, entre essas ameaças podemos citar:

- **Cross-Origin Resource Sharing (CORS)**: É um mecanismo baseado em cabeçalhos HTTP que permite a um servidor indicar todas as origens (domínio, esquema ou porta) a partir das quais um navegador deve permitir o carregamento de recursos. O CakePHP suporta cabeçalhos CORS, por padrão apenas as solicitações feitas da mesma origem podem acessar os serviços web do servidor.
- **Cross-Site scripting (XSS)**: É uma falha de segurança que permite que um invasor injete um código malicioso do lado do cliente em um site. Este código é executado pelas vítimas e permite que os invasores

ignorem os controles de acesso e se façam passar por usuários. No CakePHP existe uma função, *h()*, que impede a execução de código externo que já foi salvo no DB. Para evitar salvar código externo no DB que foi enviado através dos formulários de entrada, é possível usar funções nativas do PHP, como *htmlspecialchars()* e *strip_tags()* antes de salvar a informação.

- **Cross-Site Request Forgery (CSRF)**: É um tipo de ataque relacionado a XSS. O invasor faz com que o navegador do usuário faça uma requisição ao servidor do site sem o consentimento ou conhecimento do usuário. O CakePHP possui um *middleware* para lidar com esses tipos de ameaças. Este *middleware* inclui *tokens* CSRF em todos os formulários do aplicativo, também é possível gerar *tokens* para requisições AJAX. As requisições feitas que não contêm este *token* são negadas pelo servidor.
- **Injeção de SQL**: A injeção de SQL ocorre quando os aplicativos web não validam a entrada do usuário. É possível passar comandos SQL maliciosamente através da aplicação web para serem executados no servidor. Usando a injeção de SQL, é possível obter acesso não autorizado a um DB ou recuperar informações diretamente do DB. O *CakePHP* limpa automaticamente os campos enviados por meio de formulários de entrada, desde que os próprios métodos do *CakePHP* sejam usados para salvar os dados. Se for necessário executar uma consulta SQL diretamente sem usar funções do *CakePHP*, é necessário separar as consultas SQL dos parâmetros usando a função PHP nativa *prepare()*.

5 RESULTADOS

Após finalizado o processo de desenvolvimento, é possível fazer uma apresentação da visualização dos dados e das principais funcionalidades do sistema.

Ao acessar o sistema L5P, uma tela de autenticação é mostrada ao usuário (Fig. 13). Nesta tela o usuário pode se autenticar usando um nome de usuário ou e-mail e senha. Neste momento não há funcionalidade para recuperar a senha, portanto o usuário deve entrar em contato com o administrador do sistema para que uma nova senha seja gerada. As principais funcionalidades às quais o usuário terá acesso são mostradas na extremidade esquerda da tela. Essas funcionalidades são: gestão de entidades, gestão das relações entre entidades, acesso ao coletor de informações e acesso a informações de deputados e senadores.

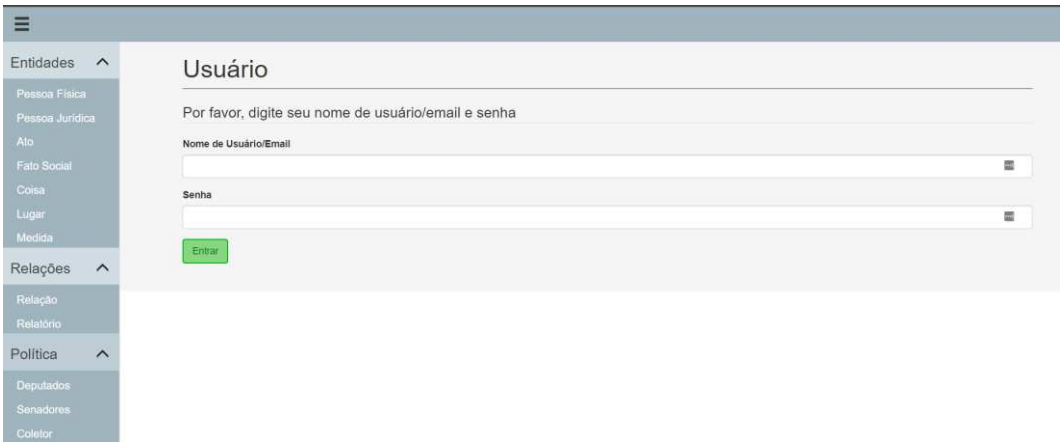

Figura 13 Tela de Autenticação
Fonte: Elaborado pelo autor

Após a autenticação, o usuário tem acesso à tela do coletor de informações (Fig. 14). O campo de seleção “Posição” permite que o usuário especifique se vai coletar dados de um deputado ou senador. O campo “Políticos” permite que o usuário selecione um político específico. Este campo é preenchido dependendo da seleção do campo “Posição” e incluirá todos os políticos que estão cumprindo um mandato naquele momento. Caso esteja selecionada a opção “Coletar tudo”, o sistema irá ignorar a seleção no campo “Políticos” e coletará as informações de todos os deputados ou senadores da atual legislatura. Se o campo “Salvar PDF” for

selecionado, o sistema salvará os discursos de políticos e processos judiciais em arquivos PDF. O resto dos campos são usados para selecionar os símbolos ou ícones que serão visualizados nos gráficos implementados.

Figura 14 Tela do coletor de dados
Fonte: Elaborado pelo autor

Quando o usuário clica no botão “Coletar”, o sistema inicia o processo de coleta de informações do deputado ou senador selecionado. Concluído esse processo, é exibida a tela com a relação dos deputados e senadores que foram coletados pelo sistema. (Fig. 15) (Fig. 16). Essa tela mostra as seguintes informações sobre os políticos: nome, CPF, sexo, data de nascimento, partido e estado, além de mostrar um botão para visualizar os dados de cada político. Por motivos de privacidade, alguns dados não são mostrados.

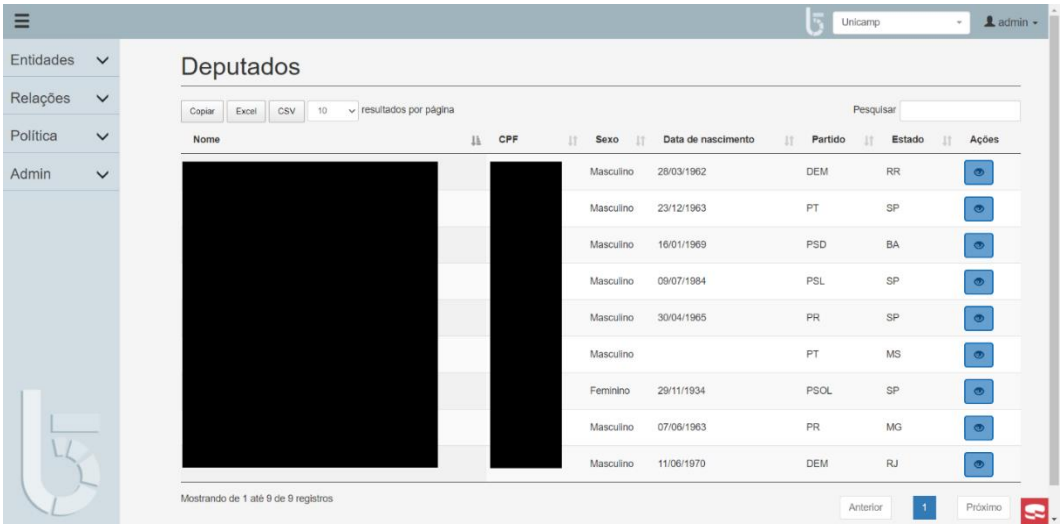

Figura 15 Relação de Deputados
Fonte: Elaborado pelo autor

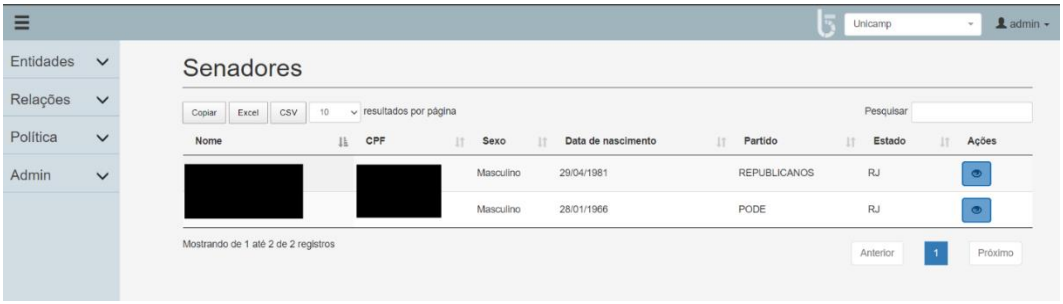

Figura 16 Relação de Senadores
Fonte: Elaborado pelo autor

Quando o usuário clica em um dos botões de visualização, a tela de detalhes do político selecionado é exibida. Esta tela mostra uma grande quantidade de informações, que são divididas em seções. Essas seções são as seguintes.

A primeira seção mostra os dados gerais do político selecionado (Fig. 17), que são divididos em dados pessoais e dados políticos. Os dados exibidos são: nome, cargo, partido, estado, alcunha, CPF, data de nascimento, formação, raça, sexo, e-mail, site, trajetória política, condição, legislatura e bancadas.

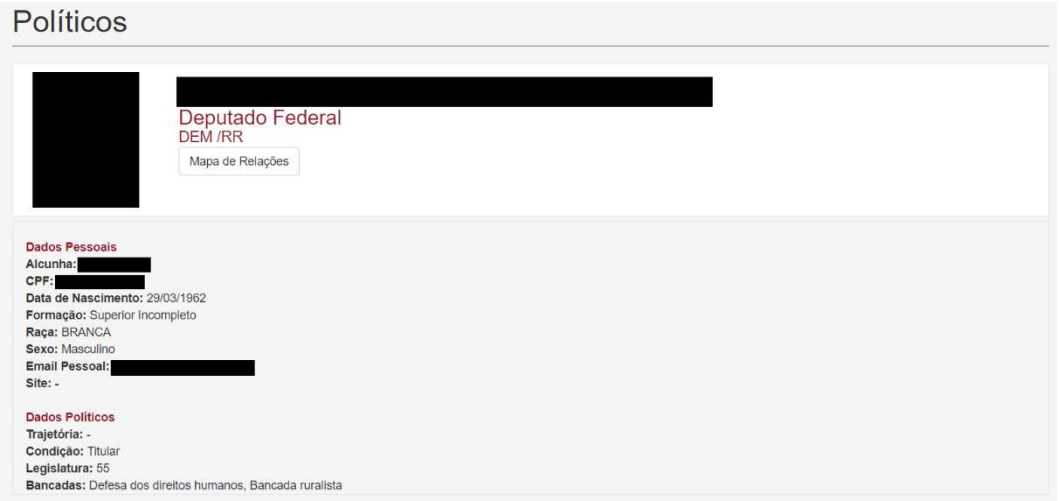

Figura 17 Dados gerais do político
Fonte: Elaborado pelo autor

Em seguida, são apresentados os dados dos discursos e declarações do político no plenário da Câmara dos Deputados ou do Senado Federal. (Fig. 18). Os dados apresentados são: o título (inclui o nome do político, a sessão e a data do discurso), o arquivo (se a opção “Salvar PDF” foi selecionada durante o processo de

coleta) e o link do discurso nos sites oficiais da Câmara dos Deputados ou do Senado Federal.

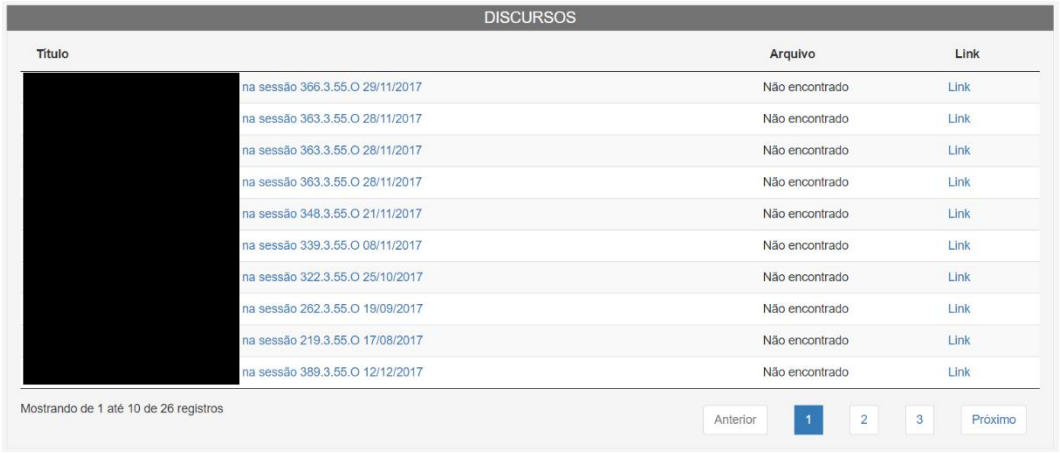

Figura 18 Dados dos discursos na Câmara ou Senado
Fonte: Elaborado pelo autor

Posteriormente, são apresentados os dados das doações recebidas nas campanhas eleitorais (Fig. 19). Os dados exibidos são: nome do doador, valor da doação e data da doação.

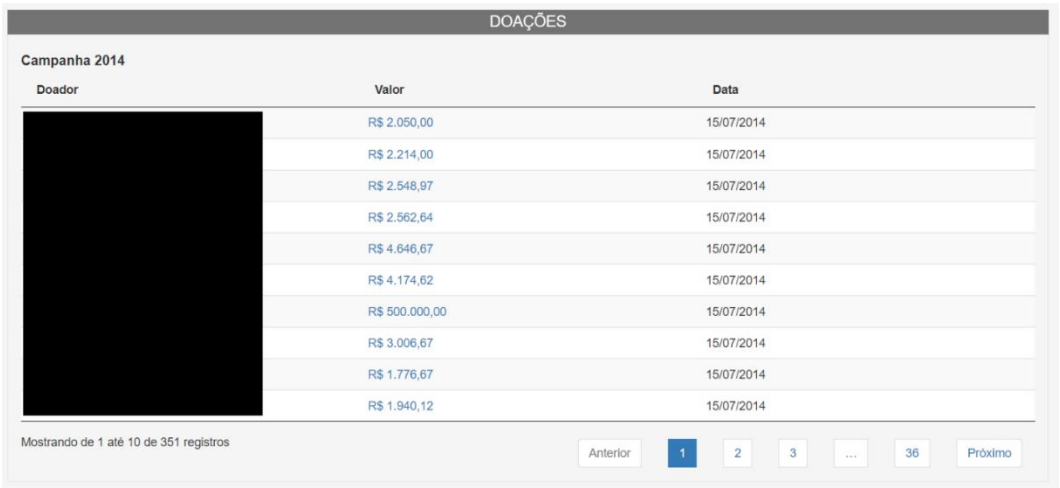

Figura 19 Dados das doações nas campanhas eleitorais
Fonte: Elaborado pelo autor

Essas doações são divididas por tipo de doador (Pessoa Física ou Pessoa Jurídica) e é calculado o percentual que cada tipo de doação representa no total de doações recebidas. Essas porcentagens são mostradas em um gráfico de pizza (Fig. 20).

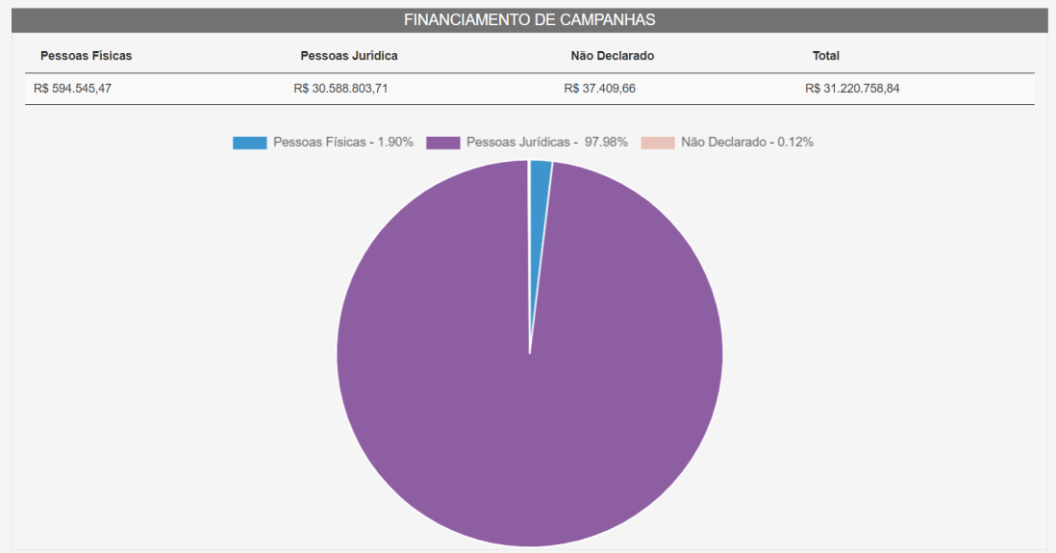

Figura 20 Gráfico do percentual de doações por tipo
Fonte: Elaborado pelo autor

Em seguida, são apresentados os dados das despesas realizadas pelo deputado ou senador durante o mandato. (Fig. 21). Os dados exibidos são: nome do fornecedor, valor da despesa, descrição da despesa e data da despesa.

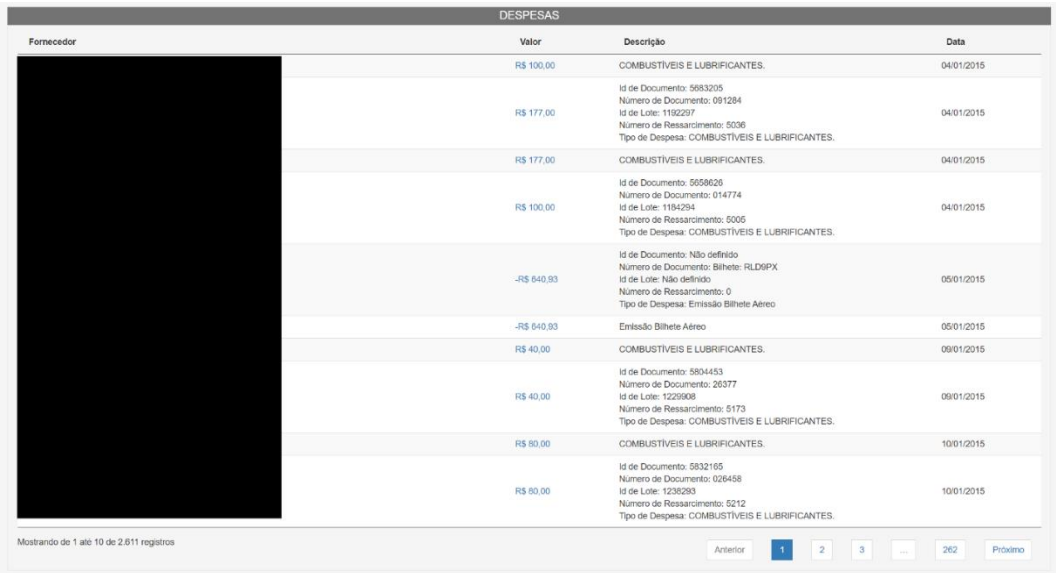

Figura 21 Dados de despesas do político
Fonte: Elaborado pelo autor

Essas despesas são divididas por tipo de despesa (Fig. 22).

Figura 22 Despesas divididas por tipo
Fonte: Elaborado pelo autor

Em seguida, são apresentados os dados dos bens declarados pelo deputado ou senador durante as campanhas eleitorais (Fig. 23). Os dados exibidos são: nome do bem, tipo de bem e valor do bem.

Figura 23 Dados dos bens declarados nas campanhas eleitorais
Fonte: Elaborado pelo autor

Por fim, são apresentados os dados dos processos judiciais, caso existam, com os quais o deputado ou senador possa estar relacionado (Fig. 24). Os dados exibidos são: nome do processo, o arquivo (se a opção “Salvar PDF” foi selecionada durante o processo de coleta) e o link do processo nos sites oficiais correspondentes. Neste caso não há dados sobre processos judiciais.

Figura 24 Dados dos processos judiciais
Fonte: Elaborado pelo autor

Na primeira seção mencionada acima, existe um botão para exibir um Mapa de Relações entre o deputado ou senador e seus principais doadores (Fig. 25).

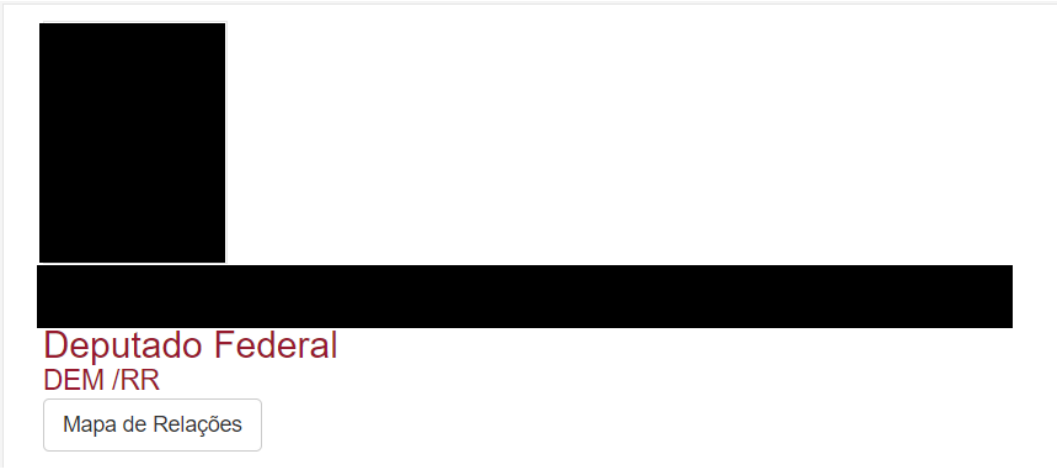

Figura 25 Botão do Mapa de Relações
Fonte: Elaborado pelo autor

Este mapa de relacionamento mostra os seguintes dados (Fig. 26): nome do deputado ou senador, os 10 principais doadores em campanhas eleitorais, o nome dos doadores e o valor acumulado de todas as doações. O político e seus doadores são exibidos usando os ícones selecionados durante o processo de coleta.

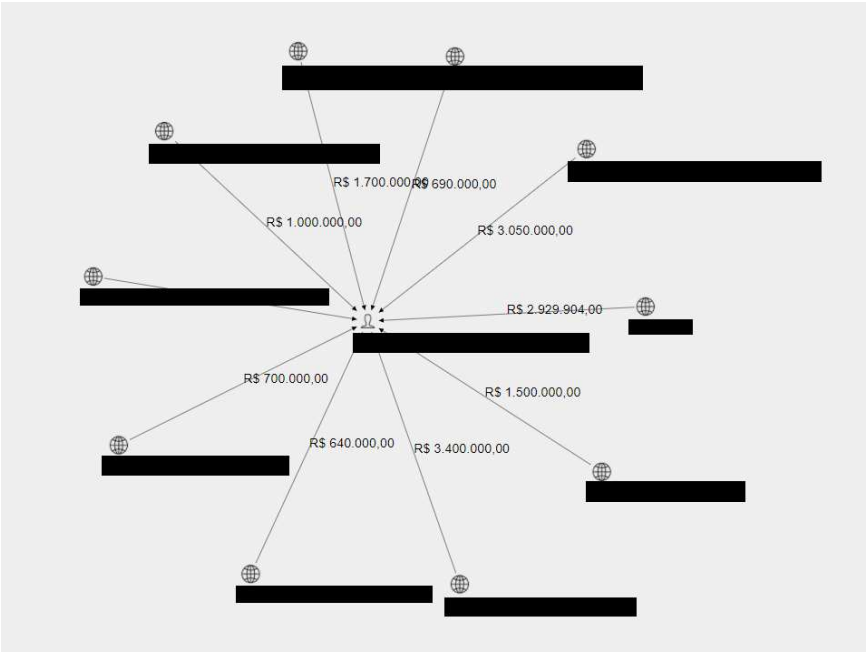

Figura 26 Mapa de relacionamento entre o político e os 10 maiores doadores
Fonte: Elaborado pelo autor

6 CONCLUSÕES

Um dos principais objetivos deste trabalho de mestrado consistiu no desenvolvimento de um sistema de recuperação de informações públicas sobre os políticos brasileiros, especificamente os membros da Câmara dos Deputados e do Senado. O outro objetivo fundamental consistiu num estudo da Lei Geral da Proteção de Dados Pessoais (LGPD) para identificar o impacto desta lei no desenvolvimento do referido sistema, bem como possíveis soluções para o cumprimento da lei.

No Capítulo 2 foi realizada uma análise da Lei Geral de Proteção de Dados Pessoais (LGPD) e o impacto desta lei no desenvolvimento do projeto. Primeiramente, foi feita uma descrição dos principais objetivos, conceitos e regulamentações presentes na lei. Foram definidas as atividades e dados aos quais se aplica a lei, os direitos dos titulares dos dados, quais são as bases jurídicas sob as quais são permitidas as atividades reguladas pela lei e as penalidades que são impostas por violações da lei. Posteriormente, foram analisados os artigos que afetam diretamente o desenvolvimento do sistema, foram analisados os diferentes cenários aos quais a lei se aplica e as possíveis exceções que podem ser aplicadas ao projeto. Em seguida, foram definidos os requisitos que o sistema deve considerar para cumprir a lei, analisando os artigos relevantes. Foram também mencionadas algumas soluções que podem ser implementadas para este fim. Finalmente, foi apresentado um exemplo das possíveis consequências do não cumprimento da lei, tendo como referência a lei europeia de proteção de dados, a General Data Protection Regulation (GDPR), que tem muitas semelhanças com a LGPD.

No Capítulo 3 deste trabalho, foi abordada a fundamentação teórica sobre a qual se baseia o desenvolvimento do sistema de recuperação de informação L5P. Primeiramente, o processo de Mineração de Dados foi analisado, este é um processo cíclico, dinâmico e iterativo composto por 6 etapas principais. O principal objetivo do processo consiste na geração de um modelo de Mineração de Dados, que pode ser preditivo ou descritivo, de forma a identificar padrões, tendências e relações existentes nos dados e que não podem ser detectados através de técnicas tradicionais de exploração.

Neste mesmo capítulo, foi analisado o processo de Mineração de Dados na Web, que pode ser definido como uma subárea da Mineração de Dados tradicional, que visa extrair informações úteis e relevantes, bem como gerar novos conhecimentos sobre os dados capturados de fontes de informação localizadas na web. A Mineração de Dados na Web é composta por 4 etapas fundamentais, que têm relação direta com as etapas presentes na Mineração de Dados tradicional, podendo ser divididas em 3 áreas de interesse, dependendo das fontes de informação utilizadas. Dentre essas áreas de interesse, a área mais relevante para o desenvolvimento do sistema de recuperação de informação L5P é a Mineração de Conteúdo, a qual foi analisada em maior detalhe neste capítulo. A Mineração de Conteúdo da Web consiste na exploração, transformação e extração de informações e novos conhecimentos presentes no conteúdo das páginas da web. Para a realização desse processo, são utilizadas técnicas tradicionais de Mineração de Dados, e processos e técnicas desenvolvidas especificamente para a extração de informações presentes na estrutura de páginas da web, como *crawlers* e *wrappers*.

O Capítulo 4 descreve o processo de implementação do sistema de recuperação de informações L5P. Primeiramente, foi apresentada a Metodologia de desenvolvimento utilizada. O projeto foi desenvolvido usando uma metodologia ágil de desenvolvimento de software conhecida como *Scrum*. É uma das metodologias ágil mais usadas, e presenta várias vantagens em comparação com as metodologias tradicionais de desenvolvimento. Depois foram definidas as tecnologias utilizadas para o desenvolvimento do sistema. Quando se decidiu realizar a implementação do L5P como um sistema web, foram escolhidas tecnologias de desenvolvimento web de código aberto e multiplataforma, para poder aproveitar o grande número de soluções disponíveis e as grandes comunidades que estas tecnologias reúnem. Posteriormente, foram definidas as fontes de informação a serem utilizadas. Foram encontradas fontes de informação pública, incluindo fontes oficiais do governo, bem como outras fontes de informação complementares. Essas fontes de informação estão disponíveis em diversos formatos (serviços web, planilhas, informações presentes nas páginas, etc.), o que requer diferentes abordagens específicas para cada formato para poder realizar a extração e captura dos dados. Em seguida, foi descrito o processo de limpeza dos dados e a estrutura utilizada para armazenar eles. Foi utilizada uma estrutura com entidades genéricas que permite, através das

relações entre elas, representar qualquer entidade conhecida. Por fim, foram apresentadas diferentes medidas de segurança implementadas para a proteção do acesso aos dados.

No Capítulo 5, foi realizada uma apresentação do sistema L5P. Foram mostradas várias telas do sistema e foi detalhado seu funcionamento.

6.1 Trabalhos futuros

Após a realização deste trabalho podem ser feitas várias recomendações para trabalhos futuros

- Recomenda-se continuar o desenvolvimento do L5P, implementando técnicas de Mineração de Dados que permitam extrair padrões, tendências, relacionamentos e novos conhecimentos dos dados coletados.
- Recomenda-se implementar técnicas de Mineração de Textos que permitam analisar os discursos e declarações públicas de políticos para obter novos conhecimentos.
- Recomenda-se a implementação de técnicas de Análise de Sentimentos para a análise de declarações e publicações em redes sociais e poder comparar essas declarações com as ações do político.
- Recomenda-se criar vários métodos de visualizações (gráficos, mapas de relações, etc.) que capturem a atenção dos usuários e que facilitem a compreensão dos dados.
- Recomenda-se implementar técnicas de Mineração de Uso da Web para analisar a interação de usuários com L5P e personalizar a ferramenta com suas necessidades.
- Recomenda-se implementar um sistema automatizado de aviso aos titulares dos dados capturados pelo L5P
- Recomenda-se implementar um sistema de anonimização dos dados para obter conformidade com a LGPD

- Recomenda-se implementar um sistema de atenção automática às petições de confirmação de tratamento e acesso simplificado dos dados dos titulares.

7 REFERÊNCIAS

AGGARWAL, C. C. ***Data mining: the textbook***. Springer, 2015

BIN, W. e ZHIJING, L. **"Web mining research."** *Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003*. IEEE, 2003. DOI: https://doi.org/10.1109/ICCIMA.2003.1238105

Bloomberg. **No One Has Ever Made a Corruption Machine Like This One**. 2017. Disponível em: https://www.bloomberg.com/news/features/2017-06-08/no-one-has-ever-made-a-corruption-machine-like-this-one. Acesso em: 21 sep. 2020.

BRAMER, M. ***Principles of data mining***. Vol. 180. London: Springer, 2007

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. **Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet)**. Diário Oficial da União: seção 1, Brasília, DF, ano 155, n. 157, p. 59-64, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 21 sep. 2020.

CAKEPHP. **CakePHP at a Glance**. Disponível em: https://book.cakephp.org/4/en/intro.html. 2020. Acesso em: 31 dec. 2020

CNN – Cable News Network (Internet Archive). ***What's REALLY behind the Brazilian riots?*** 2013. Disponível em: https://web.archive.org/web/20170723002811/http://ireport.cnn.com/docs/DOC-988431. Acesso em: 21 sep. 2020.

DW - Deutsche Welle. **Odebrecht bribed across Latin America**. 2016. Disponível em: https://www.dw.com/en/odebrecht-bribed-across-latin-america/a-36887600. Acesso em: 21 sep. 2020.

ETZIONI, O. **The World-Wide Web: quagmire or gold mine?** Commun. ACM 39, 11, 65–68. nov. 1996. DOI: https://doi.org/10.1145/240455.240473

FAYYAD, U.; PIATETSKY-SHAPIRO, G. e SMYTH, P. **"From data mining to knowledge discovery in databases."** *AI magazine* 17.3, 37-37. 1996. DOI: https://doi.org/10.1609/aimag.v17i3.1230

GDPR.EU. **What is the LGPD? Brazil's version of the GDPR**. Disponível em: https://gdpr.eu/gdpr-vs-lgpd/?cn-reloaded=1. 2020. Acesso em: 21 sep. 2020.

GERRISH LEGAL. **Processing Publicly Available Data: What are Your Obligations?** Disponível em: https://www.gerrishlegal.com/legal-blog/2019/7/2/processing-publicly-available-data-what-are-your-obligations. 2019. Acesso em: 21 sep. 2020

GITHUB. **The State of the Octoverse**. Disponível em: https://octoverse.github.com/. 2020. Acesso em: 21 sep. 2020.

GOOGLE. **How Google Anonymizes Data**. Disponível em: https://policies.google.com/technologies/anonymization?hl=en-US. 2020. Acesso em: 21 sep. 2020

HAND, D. J.; MANNILA, H. e SMYTH, P. ***Principles of data mining (adaptive computation and machine learning)***. MIT Press, 2001. DOI: https://doi.org/10.2165/00002018-200730070-00010

HAWKINS, D. M. ***Identification of outliers***. Vol. 11. London: Chapman and Hall, 1980.

IMPERVA. **Anonymization**. Disponível em: https://www.imperva.com/learn/data-security/anonymization/. 2020. Acesso em: 21 sep. 2020.

JACKSON, J. **"Data mining; a conceptual overview."** *Communications of the Association for Information Systems* 8.1, 19. 2002. DOI: https://doi.org/10.17705/1CAIS.00819

KOSALA, R.; e BLOCKEEL, H. **"Web mining research: A survey."** *ACM Sigkdd Explorations Newsletter* 2.1, 1-15. 2000. DOI: https://doi.org/10.1145/360402.360406

MARISCAL, G.; MARBAN, O., e FERNANDEZ, C. **"A survey of data mining and knowledge discovery process models and methodologies."** *The Knowledge*

Engineering Review 25.2, 137. 2010. DOI: https://doi.org/10.1017/S0269888910000032

MENDOZA, M. Minería de datos en la Web. CACHEDA, F.; FERNANDEZ, J. e HUETE, J. **Recuperación de información: Un enfoque práctico y multidisciplinar**, capítulo 19, **Minería de datos en la web**, Editorial Ra-Ma. 2011.

MICHENER, G. e PEREIRA, C. **A Great Leap Forward for Democracy and the Rule of Law? Brazil's Mensalão Trial**. Journal of Latin American Studies, 48(3), 477-507. 2016. DOI: https://doi.org/10.1017/S0022216X16000377

MOZILLA MSDN. **Types of attacks**. Disponível em: https://developer.mozilla.org/en-US/docs/Web/Security/Types_of_attacks#Cross-site_scripting_XSS. 2020. Acesso em: 21 sep. 2020

NETCRAFT. **August 2020 Web Server Survey**. Disponível em: https://news.netcraft.com/archives/2020/08/26/august-2020-web-server-survey.html. 2020. Acesso em: 21 sep. 2020.

POSTGRESQL. **About**. 2020. Disponível em: https://www.postgresql.org/about/. Acesso em: 21 sep. 2020.

SHARMA, K.; SHRIVASTAVA, G. e KUMAR, V. **"Web mining: Today and tomorrow."** *2011 3rd International Conference on Electronics Computer Technology*. Vol. 1. IEEE. 2011. DOI: https://doi.org/10.1109/ICECTECH.2011.5941631

SINGH, B. e SING, H. **Web Data Mining research: A survey**. 2010 IEEE International Conference on Computational Intelligence and Computing Research, 1-10. 2010. DOI: https://doi.org/10.1109/ICCIC.2010.5705856

SUTHERLAND, J. e SCHWABER, K. **"The Scrum Papers."** *Nuts, Bolts and Origins of an Agile Process* (2007).

TECHNOLOGY LAW DISPATCH. **Processing publicly available personal data without telling data subjects? The Polish data protection authority has (bad) news for you…** Disponível em: https://www.technologylawdispatch.com/2019/04/privacy-data-protection/processing-

publically-available-personal-data-without-telling-data-subjects-the-polish-data-protection-authority-has-bad-news-for-you/. 2019. Acesso em: 21 sep. 2020

TRANSPARENCY INTERNATIONAL. **Corruption Perceptions Index 2019**. 2019. Disponível em: https://www.transparency.org/en/cpi/2019. Acesso em: 21 sep. 2020.

W3TECHSa. **World Wide Web Technology Surveys**. Disponível em: https://w3techs.com/technologies/overview/programming_language. 2020. Acesso em: 21 sep. 2020.

W3TECHSb. **World Wide Web Technology Surveys**. Disponível em: https://w3techs.com/technologies/overview/javascript_library. 2020. Acesso em: 31 dec. 2020