Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Desenvolvimento, implementação e validação de um *scraper* para a captura de dados jurídicos dos tribunais do Brasil

Daniel Castro Neto Galhardo

JUIZ DE FORA
JANEIRO, 2023

Desenvolvimento, implementação e validação de um *scraper* para a captura de dados jurídicos dos tribunais do Brasil

Daniel Castro Neto Galhardo

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Priscila Vanessa Zabala Capriles Goliatt

JUIZ DE FORA

JANEIRO, 2023

Desenvolvimento, implementação e validação de um *scraper* para a captura de dados jurídicos dos tribunais do Brasil

Daniel Castro Neto Galhardo

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Priscila Vanessa Zabala Capriles Goliatt
Doutora em Modelagem Computacional

Igor de Oliveira Knop
Doutor em Modelagem Computacional

Edelberto Franco Silva
Doutor em Computação

JUIZ DE FORA

13 DE JANEIRO, 2023

Resumo

Com o advento do processo judicial eletrônico, o trabalho do advogado se tornou muito mais prático, evitando o deslocamento físico para um fórum a fim de buscar informações sobre um processo em andamento. Porém, quando é necessário lidar com uma quantidade maior de processos, a complexidade cresce com a praticidade. Dessa forma, neste trabalho foi desenvolvido uma solução que busque facilitar o acesso às informações judiciais. Para isso, foi desenvolvido um *scraper* e uma API como forma de monitorar e distribuir as informações judiciais. Após a implementação, foi criado um plano de negócio para tratar o *software* como um produto. Por fim, foi realizada uma comparação entre as atuais soluções disponíveis no mercado e o que foi desenvolvido na pesquisa para analisar a viabilidade deste possível produto.

Palavras-Chave: Advogado. Automação. *Software.Scraper. Crawler.*

Abstract

With the advent of electronic lawsuits, the lawyer's work became much more practical without the need of having to move from an office to a court just to find new information about a lawsuit. However, when a law firm needs to deal with a big quantity of lawsuits, the complexity grows with the practicality. Therefore, in this thesis the author developed an application capable of scraping three brazilian court's website to scrape it's data about the lawsuit's progress. Besides that, the author developed an API to share the data. After the development phase, the scraped data was validated by a lawyer and a business model was created so the software could be treated as a service. To finish the thesis, author made a comparison between solution available on the market and what was developed in this paper to analyze the viability of this possible product.

Keywords: Lawyer. Automation. Software. Scraper. Crawler.

Agradecimentos

Primeiramente, gostaria de agradecer aos meus pais por todo apoio dado durante toda a minha vida, em que sempre se fizeram presentes e nunca deixaram faltar nada para que eu pudesse trilhar meu caminho.

Aos meus amigos do grupo Mesa Redonda, que sempre acreditaram em mim e buscaram me ajudar, tanto profissionalmente como academicamente, sem o apoio deles essa pesquisa não teria sido a mesma. Com eles aprendi que quando há afinidade, a amizade pode ser construída com leveza, confiança e facilidade num curto período.

Também aos meus amigos da faculdade, Arthur, Caio, Camila, João, Pedro, Victor e Vinicius que me acompanharam durante todo o trajeto acadêmico, que sempre me estimularam a continuar e aprender cada vez mais, graças a eles, foi possível desenvolver esse trabalho.

Às minhas amigas, Isabella e Karina, que sempre me apoiaram e contribuíram imensamente para essa pesquisa, usando seus conhecimentos textuais para que eu pudesse melhorar minha escrita.

Aos membros da banca, por disponibilizarem seu tempo para contribuir positivamente para a apresentação desta monografia.

Por fim, gostaria de agradecer à Professora Doutora Priscila, que me orientou durante todo o desenvolvimento da pesquisa. Sua visão global de computação e suas opiniões tiveram contribuições imensuráveis ao meu trabalho.

Conteúdo

Lista de Tabelas

Lista de Figuras

Lista de Abreviações

API	*Application Programming Interface*
CNJ	Conselho Nacional de Justiça
CSV	*Comma Separated Values*
DCC	Departamento de Ciência da Computução
DDD	*Domain-Driven Design*
HTML	HyperText Markup Language
HTTP	*Hyper Text Transfer Protocol*
JDBC	*Java Database Connectivity*
JSON	*JavaScript Object Notation*
MB	*Megabytes*
PDF	*Portable Document Format*
PJE	Processo Judicial Eletrônico
RPA	*Robotic Process Automation*
STF	Supremo Tribunal Federal
TCC	Trabalho de Conclusão de Curso
TI	Tecnologia da Informação
TJCE	Tribunal de Justiça do Estado do Ceará
TJRR	Tribunal de Justiça do Estado de Roraima
TRF	Tribunal Regional Federal
TST	Tribunal Superior do Trabalho
RAM	*Random Access Memory*
UFJF	Universidade Federal de Juiz de Fora
XML	*Extensible Markup Language*
WWW	*World Wide Web*

1 Introdução

1.1 Apresentação do Tema

Com o advento da tecnologia, surgiu também a necessidade das empresas adaptarem seus processos para possibilitar a utilização de novas ferramentas que aumentem a produtividade do negócio. O principal fator individual de mudança na transformação das empresas é a tecnologia, demonstrando que não apenas é necessário adotar as novas tecnologias, mas também criar um plano de adaptação eficaz (GONCALVES, 1998).

Na área judicial, muitos fluxos dependiam da movimentação física dos advogados, como a busca de documentos relacionados a um processo (FRANCO, 2016). Porém, com o passar do tempo, os tribunais se digitalizaram e passaram a disponibilizar suas informações virtualmente. Esse tipo de mudança abriu um leque de oportunidades para o surgimento de soluções que buscassem tornar mais ágil o acesso às informações contidas nos tribunais. Assim, isso diminuiu o trabalho trivial feito pelos advogados, permitindo que eles pudessem focar na parte jurídica do trabalho, como pode ser observado na Figura 1.1.

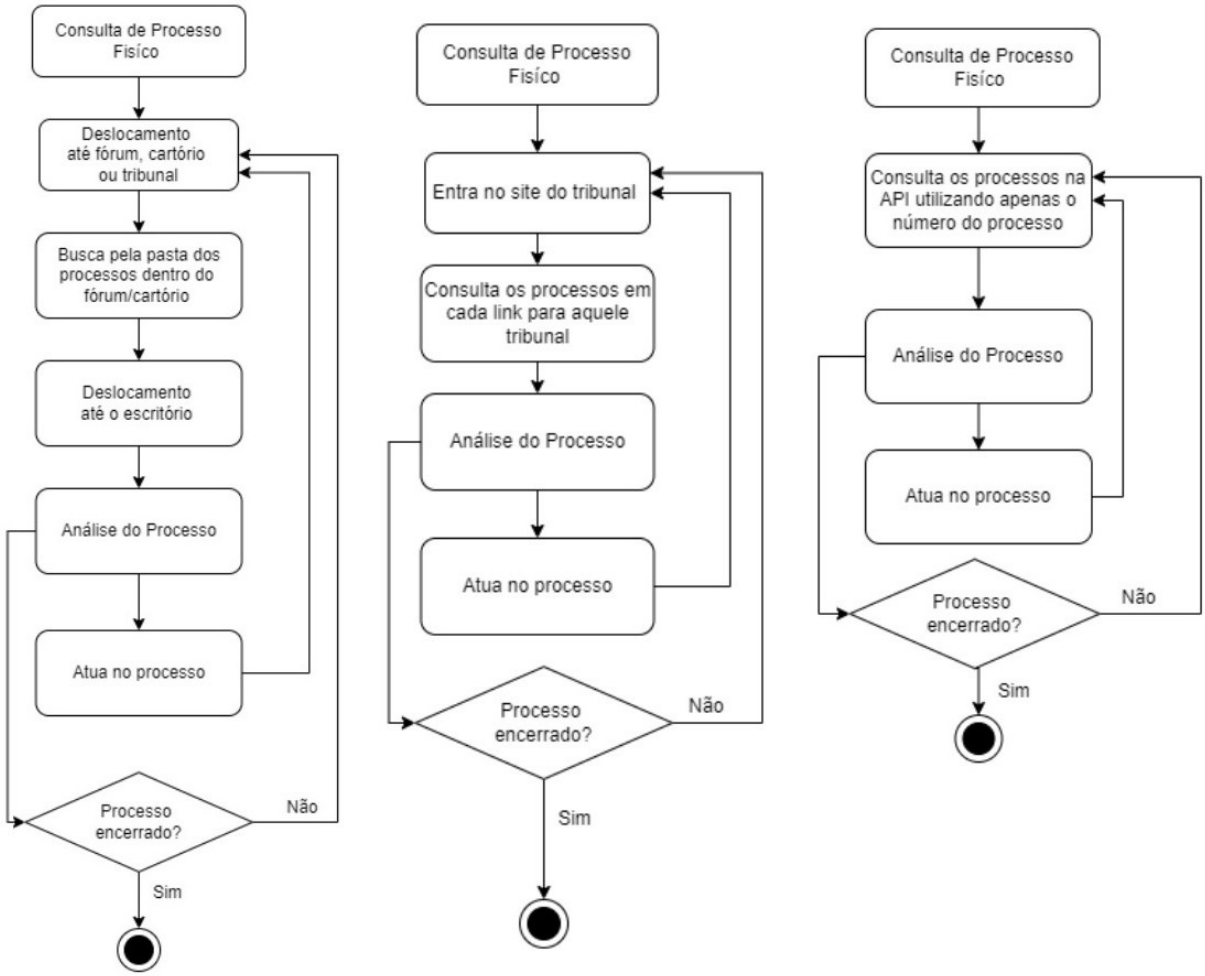

Figura 1.1: Diagrama de rotina do advogado com processos físicos, digitais e com a automação das consultas dos processos digitais. Elaborado pelo autor.

Dessa forma, nesta pesquisa, foram trabalhadas todas as etapas de criação de um *software* capaz de monitorar e coletar as informações das movimentações judiciais. Além disso, foi criado um modelo de negócio cujo objetivo é descrever a estratégia de negócios de uma *spin-off* do setor judicial. Essa empresa tem como produto principal um *software* que coletará e centralizará as informações judicais, além de realizar a distribuição dos dados através de uma *Application Programming Interface* (API).

1.2 Problema Abordado

Nesta pesquisa, foi escolhido realizar um estudo de caso utilizando uma empresa do setor judicial. A empresa em questão é chamada Ativa Gerenciamento de Recursos , que propõe soluções relacionadas aos problemas das garantias judiciais. Para isso, ela atua nas áreas

de recuperação, substituição e conciliação de depósitos judiciais. Além disso, ela também trabalha com o fornecimento de dados relacionados a esses depósitos. Para que essas atividades possam ser realizadas, é necessário que os advogados da empresa acompanhem os andamentos judiciais, algo que toma muito tempo devido à quantidade de processos monitorados pela empresa.

Atualmente, a empresa Ativa Gerenciamento de Recursos possui um sistema próprio acessado pelos advogados, no qual é possível consultar informações do processo e realizar as movimentações necessárias para seguir com a operação. Essas informações são trazidas por robôs de fornecedores. Porém, para aumentar a confiabilidade, surgiu a necessidade da criação de uma solução própria, que consiga alimentar o próprio sistema utilizado pela empresa e que tenha potencial para se tornar um novo produto.

1.3 Revisão da Literatura

Com o intuito de enriquecer o conteúdo da pesquisa, foram buscados trabalhos que têm como objetivo a aplicação de soluções tecnológicas para a área jurídica. Foram encontrados trabalhos que abordam o desenvolvimento teórico da criação de soluções desse tipo, com o foco em entender as necessidades dos advogados, os limites éticos da automação e questões de segurança e legalidade. Também foram identificadas pesquisam que desenvolvem a parte técnica, ou seja, o desenvolvimento de sistemas e protótipos.

Nesta área de estudo, é um consenso entre diferentes trabalhos que a automação surge como uma ferramenta de grande ajuda no setor jurídico. O *software* de automação afeta juízes, promotores e advogados de maneira que não é mais possível limitar o grau de importância dessa ferramenta (RAMOS, 2020). Com a implementação da automação, os principais benefícios são a agilidade e a maior produtividade na obtenção de informações (VELOSO, 2009). Assim, é possível perceber que a automação surge como uma solução muito importante na área jurídica, em que existem muitos processos manuais que podem ser otimizados para poupar tempo para os advogados e aumentar a produtividade de seus trabalhos.

Os estudos relacionados à automatização das atividades jurídicas começaram no ambiente acadêmico e ganharam uma dimensão muito maior, culminando em uma nova re-

volução tecnológica (ALVES, 2021). Como consequência, as *startups* com enfoque jurídico que surgiram dessa revolução são chamadas de *legaltech* e *lawtech* (ALVES, 2021). Na literatura brasileira, as definições desses termos são consideradas sinônimos (CÂMARA, 2018). Na literatura estrangeira, *lawtech* é a definição de uma *startup* que implementa soluções focadas no trabalho dos advogados e *legaltech* uma *startup* que cria produtos para qualquer área relacionada ao Direito (SIMÕES, 2018). Essa definição é corroborada pelo estudo realizado por Salmerón-Manzano (2021), no qual a autora ainda define que, em uma *lawtech*, as ferramentas são desenvolvidas para substituir os advogados, enquanto numa *legaltech* os produtos e serviços criados funcionam como auxílio para o trabalho do advogado. Neste projeto de pesquisa, a *spin-off* em questão foi categorizada conforme a definição da literatura estrangeira de *legaltech*, já que possui como produto principal um sistema que será utilizado por advogados.

Alves (2021), em sua dissertação, apresentou questões relacionadas aos limites da utilização de *software* no setor jurídico. Para alcançar esse objetivo, o autor definiu as principais atividades de advogados que podem ser consideradas triviais e que exijam de seu intelecto quanto as questões judiciais. Na abordagem proposta por ele, foi traçado um panorama desde o trabalho mais trivial que o advogado faz, na parte administrativa e mecânica de coleta de informações, até as atividades jurídicas que necessitam de um humano para serem desenvolvidas.

É interessante notar que a atividade proposta pelo *software* desenvolvido nesta pesquisa está enquadrado no grupo de atividades secundárias, delegadas ou repetitivas. Para Alves (2021), a busca por informações processuais pode ser considerada uma atividade com baixa utilização do raciocínio jurídico e que possui um padrão de execução repetitivo. Dessa forma, não há nenhum tipo de prejuízo que possa ser causado por um sistema que faça a captação desse tipo de informação de maneira automatizada, como propõe o trabalho atual.

Na pesquisa de Ramos (2020), o autor procura desenvolver o protótipo de um *software* para realizar a classificação de processos. O estudo faz uma análise sobre o impacto da tecnologia no judiciário, expondo os obstáculos encontrados durante a implantação do processo judicial eletrônico no Brasil, como a dificuldade dos servidores de

lidar com esse tipo de processo. Apesar disso, o autor afirma ser imprescindível o uso da tecnologia para melhorar a eficiência do setor judiciário.

Para a elaboração do protótipo, Ramos (2020) realizou entrevistas para delimitar o escopo do sistema desenvolvido. O autor também buscou construir uma base de dados relevante. Para isso, coletou 850 petições que tramitavam no Juizado Especial Cível. Com uma base de dados concreta, o autor utilizou 30 processos judiciais para testar o *software* de classificação de processos. Vale destacar que o autor não detalha a implementação, apenas deixando explícito o uso de *machine learning*, API e de um *software* de extração de texto. Sobre a forma de realizar a classificação, o sistema recebe um arquivo de texto Portable Document Format (PDF), retira todos os caracteres especiais e converte o que sobra em letras minúsculas. Após isso, são retirados os termos considerados não relevantes. Assim, com as palavras que restaram, é realizada a classificação daquele processo conforme a base de dados montada.

Em relação aos resultados da pesquisa, Ramos conseguiu um tempo médio de 0,041 segundos gastos por página classificada, uma acurácia de 93,58% e uma precisão de 72,72%. Para realizar essas medidas, os processos foram classificados por advogados e os resultados foram comparados com os obtidos pelo sistema.

As pesquisas citadas anteriormente possuem em comum que as informações são cadastradas no sistema através da entrada do usuário, ou seja, é necessário que alguém preencha os campos de entrada para ocorrer o fluxo de informações. Ao contrário desse comportamento, na pesquisa de Guardia, Fabiano e Nascimento (2022), solicitada pelo magistrado brasileiro, foi realizada a automação de processos no Tribunal Regional Federal da 1ª Região (TRF1) para o auxílio emergencial durante a pandemia. Nesse contexto, muitas pessoas solicitaram essa assistência e tiveram seus pedidos negados, mesmo atendendo a todos os critérios sociais e financeiros. Por esse motivo, as pessoas contrataram advogados para então requererem seus direitos através de um processo judicial.

Dessa forma, para acelerar o trabalho dos advogados na consulta das movimentações do processo, foi realizada a automação das consultas utilizando a linguagem Python para o desenvolvimento do robô e SQLite para o armazenamento dos dados. A fonte das informações foi o Processo Judicial eletrônico (PJe), sistema desenvolvido pelo

Conselho Nacional de Justiça. Foram consultados inicialmente 27 mil processos, que tiveram não apenas seus autos processuais coletados como também certidões e outros documentos. Ao final do projeto, com o sucesso da implementação, haviam sido consultados mais de 38 mil ações judiciais.

Na pesquisa realizada por Calò (2014), foi desenvolvido um *web scraper* para extrair informações relacionadas a um acórdão, que pode ser definido como uma decisão do órgão do colegiado de um tribunal (CALÒ, 2014). É válido notar que, neste trabalho, o autor tinha como intuito analisar o conteúdo dessas decisões, porém, como conseguir formalmente um Comma Separated Values (CSV) para análise era muito burocrático, se tornou necessário o desenvolvimento de um *crawler*. Assim, o autor fez o uso do framework Scrappy para realizar o *scraping* das informações dos acórdãos.

Uma das dificuldades encontradas na implementação foi o fato das informações serem disponibilizadas sem um padrão, ou seja, não era possível criar uma rotina buscando através dos seletores HTML. Para contornar esse problema, Caló realizou o *download* de toda a página HTML, mantendo apenas o texto e eliminando os códigos HTML posteriormente. Uma das maiores dificuldades do autor foi lidar com o site do Supremo Tribunal Federal (STF), que não respondeu como o esperado após algumas execuções devido à falta de padronização da disponibilização das informações.

Outro projeto fortemente relacionado ao trabalho desenvolvido é o trabalho de Silva, Dias e Segundo (2021), cujo objetivo é extrair dados de patentes brasileiras para construir uma base de informações que serão futuramente analisadas localmente. É interessante observar que, na pesquisa, a motivação surgiu da necessidade de realizar análises em grandes volumes de dados mais rápida, o que não era possível quando a análise era feita sobre dados de repositórios *online*. Esse é um ponto alinhado com robô desenvolvido na pesquisa atual, em que a aplicação captura as informações sobre os autos dos processos. Outro dado apresentado pelos autores é a estimativa de tempo que um humano realizaria o trabalho. Conforme a pesquisa, seriam necessários 17.942 dias para uma pessoa que trabalha oito horas por dia. Com esse resultado, se torna explícito os benefícios da utilização da automação com o *web scraping*.

Para realizar o *web scraping* dos dados das patentes, Silva, Dias e Segundo (2021)

implementaram um algoritmo na linguagem Python para percorrer o *site* que possuía as informações. Após isso, esses dados foram utilizados para uma segunda consulta. Para tal, foi necessário desenvolver outro algoritmo para percorrer o segundo *site* de consulta. Nesse momento, houve um impedimento relacionado à necessidade de tratar os dados coletados, já que o padrão que o segundo *site* de consulta exigia era diferente. Os autores conseguiram identificar alguns padrões para poder corrigir e tratar os dados. Dessa forma, após a coleta das informações, os autores organizaram os documentos de patentes no formato *JavaScript Object Notation* (JSON). Como conclusão da pesquisa, o resultado encontrado foi de que o *web scraping*, para a construção de uma base local, tornou a análise dos dados muito mais flexível e um total de 23,7 GB de dados foram analisados, agregando um grande valor científico à pesquisa.

Com muitas similaridades com o trabalho desenvolvido, a pesquisa de Silva, Dias e Segundo (2021), assim como a de Guardia, Fabiano e Nascimento (2022) reafirma o grande impacto da automação aliada ao *webscraping*. No trabalho atual também foi utilizado o formato JSON, porém como resposta da requisição. Além disso, a estratégia de capturar dados que serão posteriormente utilizados para consultas também é uma possibilidade dentro desta pesquisa, já que é possível consultar números públicos de processos judicias eletrônicos que, em seguida, serão consultados nos sites dos tribunais.

É de comum acordo entre as pesquisas que é necessário realizar a verificação das informações de maior interesse para os advogados, como também proporcionar um modo para que o profissional possa consultar esses dados de uma forma mais intuitiva, ou seja, por um sistema que possua uma interface de fácil acesso. Ambos os pontos estão alinhados com esta pesquisa, no qual foram feitas entrevistas com advogados para entender quais são as informações relevantes para eles.

1.4 Justificativa

Após a revisão da literatura, esta pesquisa traz uma perspectiva mais atual para o desenvolvimento de *scrapers* com enfoque em dados jurídicos. Essa extração de dados proveniente dos tribunais *online* através do *web crawling* e *web scraping* pode ser bastante efetiva para escritórios de advocacia. Vale ressaltar a diferença em que, nesta pesquisa, o

software implementado será tratado como um produto.

No estudo de caso proposto, algumas áreas da empresa Ativa Gerenciamento de Recursos dependem completamente das informações relativas à movimentação dos processos, fornecidas por terceiros. Essa dependência é considerada um grande obstáculo porque certos fluxos seriam integralmente afetados caso algum dos fornecedores deixasse de prestar serviço para a empresa. Além disso, se uma nova funcionalidade é pensada, sempre é preciso verificar se os atuais fornecedores possuem a capacidade de disponibilizar os dados necessários em tempo hábil. Caso não possuam a capacidade, é preciso iniciar um processo de busca por um novo fornecedor, aumentando o grau de dependência da empresa em relação aos seus fornecedores.

Nesta pesquisa, é considerado que um produto proprietário é o ideal para que a empresa possa solucionar este problema. Outro ponto importante é a possibilidade de moldar esse novo produto a partir das necessidades já vistas, já que será utilizado como fonte de dados para seu próprio sistema e será comercializado para novos clientes. Porém, como é um produto apartado do sistema já existente, a criação de uma *spin-off* se mostra uma boa escolha para uma melhor organização da estrutura de negócio da empresa Ativa Gerenciamento de Recursos.

1.5 Objetivos

1.5.1 Objetivo Geral

Desenvolver, implementar e validar uma solução para a automatização do monitoramento e distribuição das movimentações judiciais dos tribunais do Brasil.

1.5.2 Objetivos Específicos

- Definir os tribunais que serão utilizados como fontes de dados;
- Coletar os números de processos judiciais públicos para serem usados nas consultas;
- Desenvolver o *crawler* para automatizar a busca de informações nos tribunais;

- Desenvolver uma API para realizar a distribuição dos dados coletados para os clientes;
- Validar e testar o protótipo criado;
- Propor um modelo de negócio para a *spin-off* em questão.

2 Fundamentação Teórica

Para o entendimento desta pesquisa, essa seção busca deixar claro termos e expressões que serão utilizados com frequência. Além disso, também foram definidos os conceitos considerados de maior importância. Ao final, foram feitas considerações sobre os termos discutidos. Os conceitos utilizados nesta pesquisa são: Inovação e *Startup*; *Spin-off*, Processo Judicial, Andamento Judicial, Automação do Monitoramento de Processos Judiciais, Ferramentas de Automação, Questões Éticas da Automação e a Distribuição das Informações

2.1 Inovação e *Startup*

A presente pesquisa, conforme foi explicitado, objetiva o desenvolvimento de uma aplicação que tem como foco a inovação através da centralização de informações judiciais. Para tanto, uma das etapas a ser desenvolvida consiste em realizar a verificação da viabilidade da criação de uma *startup* para distribuir esse produto no mercado. Cabe destacar, nesse sentido, que será utilizado o conceito de *startup* proposto por Torres e Souza (2016) para definir uma empresa sujeita a diversos riscos e que tem como foco a inovação.

O conceito de inovação, no que lhe concerne, é aqui entendido como uma nova forma de resolver um problema, a qual não foi executada anteriormente. A diferença entre "invenção"e "inovação"está no fato de que a primeira se refere a concepção de algo que não existia, já a segunda pode se referir a algo que foi melhorado ou algo que teve sua forma de utilização reformulada (ALVES, 2013). Nesta pesquisa, a inovação se dá através do desenvolvimento de um *software* focado no setor jurídico, impactando diretamente o trabalho dos advogados.

2.2 *Spin-off*

Vale destacar, ainda, que o conceito de *spin-off* também se adequa à proposta aqui desenvolvida, visto que uma *spin-off* é uma empresa ligada à alta tecnologia,de porte menor, com seu surgimento a partir de uma organização já estabelecida. Todo o desenvolvimento dessa empresa é realizado fora da empresa-mãe (SEQUEIRA, 2013).

No escopo da pesquisa, considerando o estudo de caso da empresa Ativa Gerenciamento de Recursos, a ideia da criação de uma *spin-off* é adequada. Isso se dá pelo fato de que é comum que as *spin-offs* apareçam em um momento em que a empresa-mãe possui um novo produto. Nessa situação, como alternativa, pode ser necessário desvincular o produto desenvolvido de sua marca principal. Com isso, um público diferente é atingido simultaneamente e experiências distintas são proporcionadas aos clientes.

Assim, também é importante considerar que *spin-offs* são consideradas *startups* nos dez primeiros anos (SEQUEIRA, 2013). Cabe a esta pesquisa realizar a criação do modelo de gestão da *spin-off*, considerando as diferentes formas de organização entre uma *startup* e uma empresa convencional.

2.3 Processo Judicial

Com os conceitos empresariais bem definidos, é de suma importância entender os termos relacionados ao setor jurídico, que foi o setor de atuação da *spin-off* desenvolvida. Para isso, a definição de um processo judicial é a de um instrumento que o Estado utiliza para, no exercício da função jurisdicional, resolver os conflitos apresentados pelas partes (PINHO, 2019). É no processo que se encontra as informações relacionadas às decisões tomadas pelo juiz e as movimentações dos advogados.

No contexto da pesquisa atual, o processo judicial utilizado será o processo eletrônico público, que surgiu devido à necessidade da digitalização do judiciário brasileiro. Essa escolha foi feita devido às principais características desse tipo de processo, sendo elas a sua disponibilização *online* de forma rápida, a facilidade de acesso às informações e a possibilidade da automação das rotinas. O fato do processo também ser público visa evitar qualquer impasse relacionado a proteção de dados, permitindo que os

dados possam ser utilizados livremente.

Portanto, como visto por Montenegro (2022), com a aprovação da Resolução CNJ n. 420/2021, a restrição a processos físicos irá acelerar a extinção dos processos em papel, reduzindo os custos e aumentando o acesso à prestação jurisdicional. Além disso, a virtualização de processos é uma realidade inexorável, de maneira que os processos físicos irão eventualmente se extinguir (OLIVEIRA, 2017). Esse é um fato importante que garante a longevidade do *software* que foi desenvolvido no trabalho, visto que o processo judicial eletrônico é a matéria-prima para as consultas realizadas por esse sistema.

2.4 Andamento Judicial

Um andamento ou movimentação judicial, pode ser definido como a "pegada do caminhar"processual, deixada por atos praticados dentro dos autos, sejam eles conclusão ao juiz, certificação de prazo, despachos, entre outros. Essas informações são de suma importância para que os advogados possam acompanhar e monitorar a movimentação do processo em que estão atuando.

No presente trabalho, o andamento é a informação principal a ser coletada, além das informações relativas às partes. O conteúdo de um andamento consiste na data de ocorrência e de um ato processual disponibilizado pelo judiciário.

2.5 Automação do Monitoramento de Processos Judiciais

Com o surgimento do processo judicial eletrônico, uma das principais vantagens se tornou a automação de rotinas (OLIVEIRA, 2017). A automatização é um conjunto de ferramentas que simula o uso de um sistema computacional por um ser humano (AALST; BICHLER; HEINZL, 2018). Esse método opera a partir do mapeamento de ações em cima de um software, repetindo essas ações conforme o programado. Sua principal função é reduzir o trabalho repetitivo, aumentando a produtividade daqueles que dependem dos resultados oriundos dessa repetição.

É necessário que esse tipo de solução seja implementada em três fases, sendo elas: prova de conceito, o piloto e os testes de caso de uso, já que uma solução mal implementada além de ser complexa, pode custar caro e ser lenta (ALBERTH; MATTERN, 2017). Com o desenvolvimento da aplicação que percorrerá os tribunais, esse conceito será utilizado para definir o nível de qualidade do *software* desenvolvido.

Atualmente é possível acessar as informações processuais *online*. Porém, como existem diversos tribunais e um tribunal pode possuir mais de uma fonte de informação, nessa pesquisa o *Robotic Process Automation* é utilizado como uma ferramenta para a criação de uma solução cuja principal vantagem é a não limitação de tempo de trabalho para os robôs, são escaláveis e aumentam a produtividade, tudo isso sendo considerada uma solução barata.

2.6 Ferramentas de Automação

Com a definição da utilização do *Robotic Process Automation* (RPA), é necessário desenvolver uma ferramenta capaz de percorrer os tribunais. Para isso, nesta pesquisa será implementado um *webcrawler*. Um *web crawler* é um *software* ou *script* programado que navega pela internet de maneira sistemática e automatizada, com o intuito de reconhecer informações (KAUSAR; DHAKA; SINGH, 2013).

Existem diversas técnicas de *crawling*. Nesta pesquisa, a técnica utilizada será a definida por Kausar, Dhaka e Singh (2013) como *focused crawling*, em que serão buscadas apenas as informações de um tópico específico que, nesse caso, são as informações sobre os processos jurídicos nos sites dos tribunais.

Um ponto importante é que os *crawlers* podem acessar informações de maneira muito rápida, gerando um enorme impacto no desempenho do sistema em que ele está atuando (DHENAKARAN; SAMBANTHAN, 2011). Portanto, é necessário que, durante o desenvolvimento do *crawler*, sejam realizados testes para entender os limites do sistema e a quantidade de dados que podem ser obtidos, de modo a assimilar os limites de eficiência do *software* que será desenvolvido na pesquisa.

Enquanto o *web crawling* é uma forma de percorrer as páginas da *web*, o *web scraping* é considerado uma técnica para a mineração de dados, ou seja, é realizado o

download dessas informações para uso posterior. Esse conceito é definido como uma técnica de extração de dados do *World Wide Web* (WWW) que serão salvos no formato de um arquivo em uma base de dados para futura análise (ZHAO, 2017).

Um *web scraper* pode ser utilizado para diversas finalidades, a comparação e monitoramento de dados e a detecção de mudanças em *websites* são alguns exemplos. Essa prática pode ser realizada manualmente, com o usuário realizando os cliques e todo o processo de armazenamento daquela informação.

2.7 Questões Éticas da Automação

Nos dias de hoje, com o *Robotic Process Automation*, o *web scraper* se associou ao *web crawler*, de forma que a automatização do reconhecimento de informações se alia a automatização do *download* de dados e das rotinas de armazenamento, tornando muito mais prática a obtenção e o armazenamento dessas informações em uma base de dados. Porém, a utilização dessa ferramenta de maneira automatizada esbarra em algumas questões legais.

Um *web scraper* consegue copiar livremente fragmentos de informações provenientes de uma página sem infringir nenhum direito legal, já que apenas uma parte específica daquelas informações estaria protegida legalmente (ZHAO, 2017). Com isso, uma ferramenta de *web scraping* ética manteria uma quantidade razoável de requisições.

No escopo da pesquisa, é extremamente necessário que seja mantida uma quantidade razoável de requisições, já que os *sites* dos tribunais são acessados diariamente por milhares de pessoas que precisam consultar processos, não só advogados. Assim, realizar muitas requisições poderia não apenas causar uma lentidão no sistema em geral como também acarretar o bloqueio acesso da máquina em que essa ferramenta se encontra.

2.8 Distribuição das Informações

Com as técnicas de automatização bem estabelecidas, foi definida a utilização de uma API para a disponibilização das informações coletadas. Uma API pode ser conceituada como um conjunto de definições e protocolos para construir e integrar uma aplicação. As

Web APIs utilizam do protocolo HTTP (*Hyper Text Transfer Protocol*) para realizar as requisições. Como resposta da requisição, são geralmente utilizados os formatos *JavaScript Object Notation* (JSON) e *Extensible Markup Language* (XML) (FIELDING, 2000), por possuírem como característica uma maior facilidade de manipulação e integração dos dados por outros sistemas.

Nesta pesquisa, foi utilizado o conceito de API Remota, ou seja, uma API criada para interagir através da comunicação pela rede, em que uma requisição vem de um computador diferente daquele em que a API se encontra. Essa API faz a distribuição das informações, através das requisições que possuem as respostas no formato JSON.

3 Material e Métodos

Para o desenvolvimento do presente trabalho, foi necessário criar um ambiente compatível com todas as tecnologias fundamentais para o desenvolvimento das aplicações que constituem o sistema responsável pela captura e entrega de movimentações judiciais. Para isso, o computador utilizado possuía as seguintes configurações:

- Processador AMD Ryzen 3 3300X.
- Memória RAM de 16 GB DDR4 2333Mhz.
- Placa de Vídeo NVIDIA GTX 1050 TI 4 GB.
- Sistema Operacional Windows 11 64 Bits.

Além disso, como *crawling* e o *scraping* são métodos que necessitam da conexão com a internet, é importante detalhar qual foi a configuração de rede utilizada, sendo ela

- Conexão do Protocolo Ethernet.
- Velocidade de Conexão de 650 Mbps, fibra ótica.
- Driver de Conexão Realtek PCIe GbE Family Controller.

Para o desenvolvimento do *crawler* foi escolhida a linguagem C#[1] utilizando o *framework* .NET 6. Em relação ao banco de dados, foi criado um servidor local utilizando o MySQL Workbench[2].

3.1 Obtenção dos Dados

Para a obtenção dos dados, foi cedido o acesso a uma base de processos públicos disponibilizados pela empresa Ativa Gerenciamento de Recursos. Dessa forma, foi possível ter acesso à relação de processos, com os tribunais que pertenciam, além das fontes em que

[1] https://learn.microsoft.com/en-us/dotnet/csharp/
[2] https://www.mysql.com/products/workbench/

eles se encontravam. Foram definidos três tribunais e, no total, 1130 processos para terem suas movimentações buscadas.

É valido reafirmar o caráter público dos processos utilizados na pesquisa de modo a evitar conflitos relacionados a consulta de dados terceiros. Como todas essas informações podem ser consultadas livremente, foi possível o uso dessas informações no atual trabalho.

3.2 Crawler

3.2.1 Framework

Para o desenvolvimento do *crawler* foi utilizado o Selenium WebDriver[3]. Esse *web framework* permite automatizar rotinas de acesso à página *web*, facilitando principalmente a criação de testes automatizados. Apesar disso, por familiaridade do autor com a ferramenta, ela foi utilizada para executar o *web scraping* das informações-alvo da pesquisa.

3.2.2 Tribunais

Durante a implementação, foram escolhidos três diferentes tribunais, sendo eles:

- Tribunal de Justiça de Roraima [4]
- Tribunal de Justiça do Ceará [5]
- Tribunal Superior do Trabalho [6]

Essas escolhas ocorreram considerando testes iniciais para validar a possibilidade de utilizá-los na pesquisa de modo satisfatório, isto é, *sites* que não possuam o recurso de segurança captcha e informações dispostas sem padrão, como visto na revisão da literatura.

Com os tribunais definidos, o primeiro passo foi verificar como as informações são dispostas em cada tribunal para buscar formas de tornar certas partes do código mais genéricas, buscando atender todos os tribunais escolhidos.

[3]https://www.selenium.dev/documentation/webdriver/
[4]https://projudi.tjrr.jus.br/projudi/processo/consultaPublica.do?actionType=iniciar
[5]https://esaj.tjce.jus.br/cpopg/open.do
[6]https://consultaprocessual.tst.jus.br/consultaProcessual/

Neste caso, foi verificado que informações equivalentes foram dispostas em campos equivalentes, mas com nomes diferentes. Com isso, foi necessário desenvolver uma rotina para cada tribunal. Para isso, foram implementadas funções respectivas a cada *site*, realizando a rotina conforme a identificação do processo.

3.2.3 Identificadores Únicos

Para a identificação dos campos de interesse, foi necessário utilizar um identificador único para que as informações fossem coletadas de maneira mais ágil. Dessa forma, foi utilizado o XPATH (*XML Path Language*) como ferramenta de identificação.

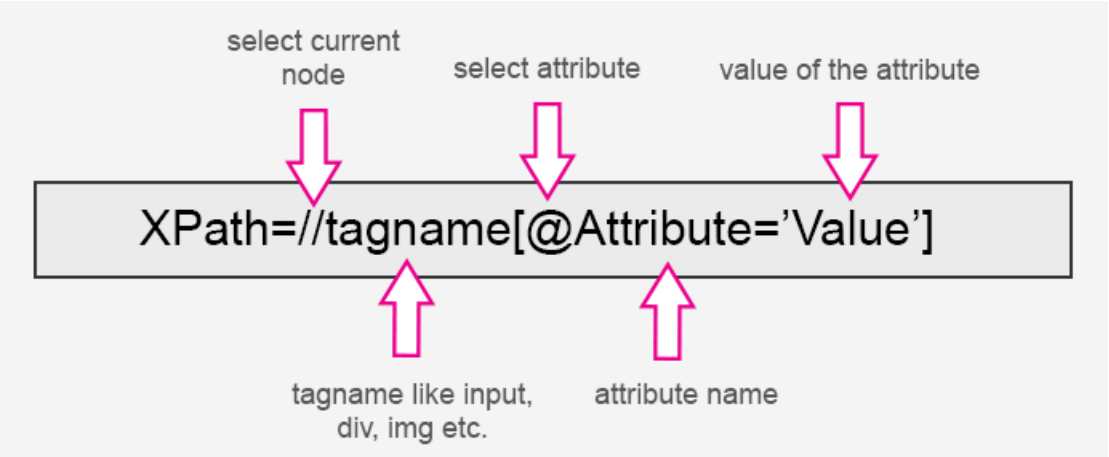

Figura 3.1: Descrição da estrutura de um XPATH. Fonte: (PROVAR, 2022)

Utilizando o Selenium, é necessário realizar a inspeção do código HTML e encontrar o elemento necessário e, a partir dele, criar a lógica para percorrer todas as informações. É possível ter acesso a diversas propriedades utilizando o Selenium WebDriver, dentre elas a que acessa as propriedades textuais do elemento. É através da manipulação dos XPATHs que se torna viável realizar o *scraping* das informações da tela.

Apesar de Gunawan et al. (2019/03) ter chegado a conclusão de que o XPATH é um método mais lento se comparado ao uso de expressões regulares e do HTML DOM para o *scraping* de informações, ainda assim foi utilizada a identificação através do XPATH, pois a maneira como os *websites* se comportavam acabou tornando a manipulação dos dados na tela mais difícil. Em relação às expressões regulares, por falta de familiaridade com o uso delas para *scraping*, essa alternativa foi descartada.

3.3 API

Para este trabalho, um dos objetivos gerais era criar uma API com a intenção de distribuir as informações através do formato JSON.

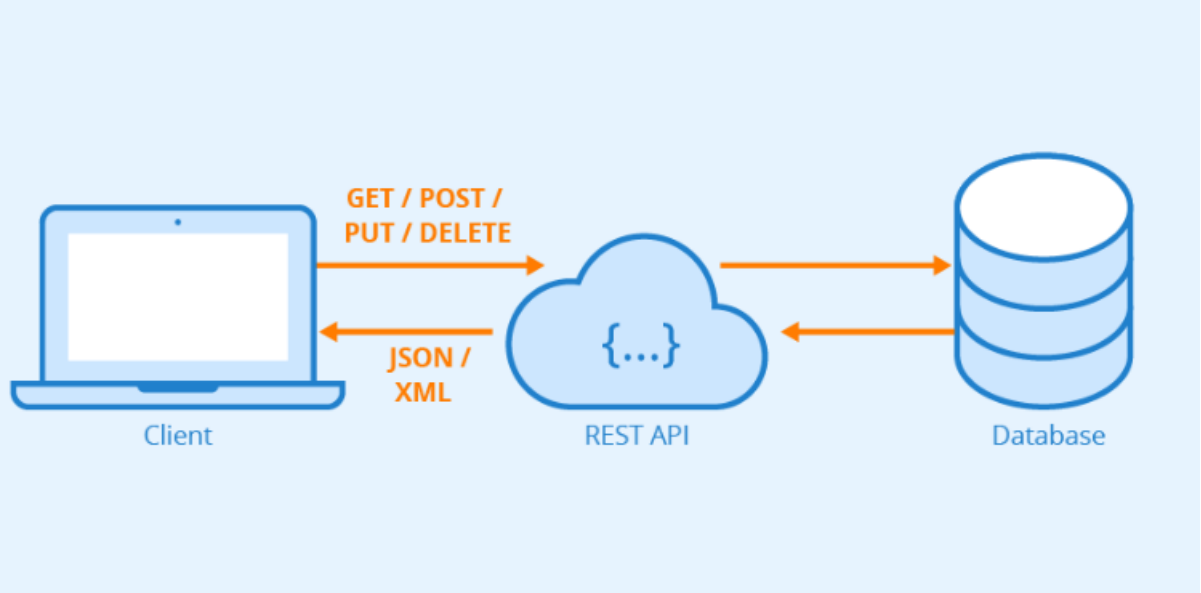

Figura 3.2: Comunicação entre API e Cliente. Fonte: (NAEEM, 2020)

Dessa forma, foi desenvolvida uma API REST utilizando ASP.NET[7] com o conjunto de ferramentas Swagger[8].

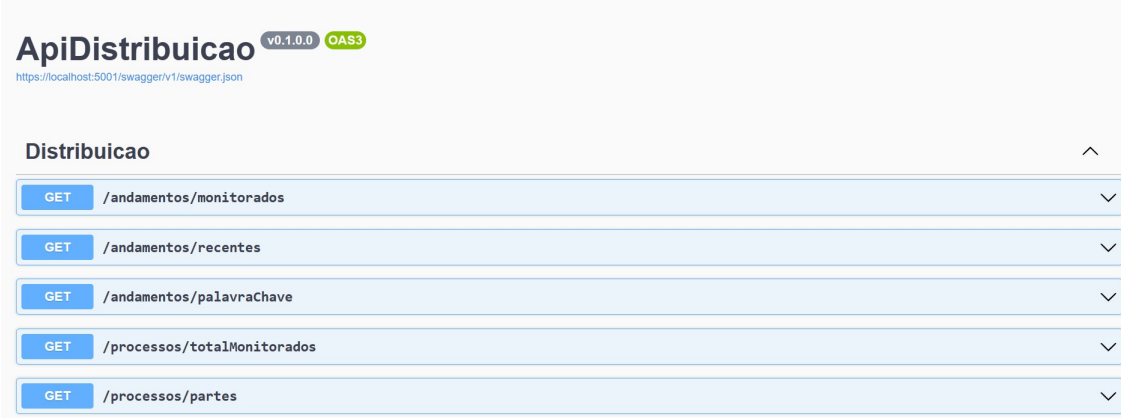

Figura 3.3: Utilização do Swagger. Elaborado pelo autor.

Também foi utilizada a abordagem *Domain-Driven Design* (DDD), aplicando seus principais conceitos, como:

[7]https://dotnet.microsoft.com/en-us/apps/aspnet
[8]https://swagger.io/tools/swagger-ui/

- Apresentação: Como a aplicação irá interagir com o usuário. Neste trabalho, através da API.
- Domínio: Implementação dos serviços.
- Infraestrutura: Implementação dos repositórios, no qual ocorre a comunicação com o banco de dados da aplicação, além de realizar o mapeamento de dados, como a leitura de informações do appsettings.json.
- Aplicação: Comunicação com o domínio.

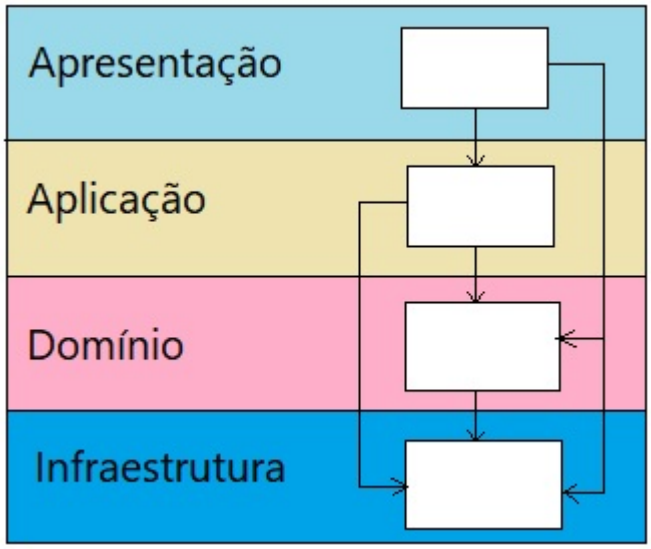

Figura 3.4: Comunicação entre as camadas utilizando o *design* DDD. Fonte: (MACORATTI, 2020)

3.4 Banco de Dados

Para armazenar as informações coletadas, foi implementado um servidor local de banco de dados relacional MySQL[9]. Foi utilizado o DBeaver[10] como *client* para as operações relacionadas ao banco de dados. Esse *software* utiliza a API para interagir com o banco de dados através do driver *Java Database Connectivity* (JDBC).

Na Figura 3.5, é possível observar o diagrama do modelo Entidade Relacionamento, visando tornar mais claro os relacionamentos entre cada tabela desenvolvida para o funcionamento do *crawler*.

[9]https://www.mysql.com/
[10]https://dbeaver.io/

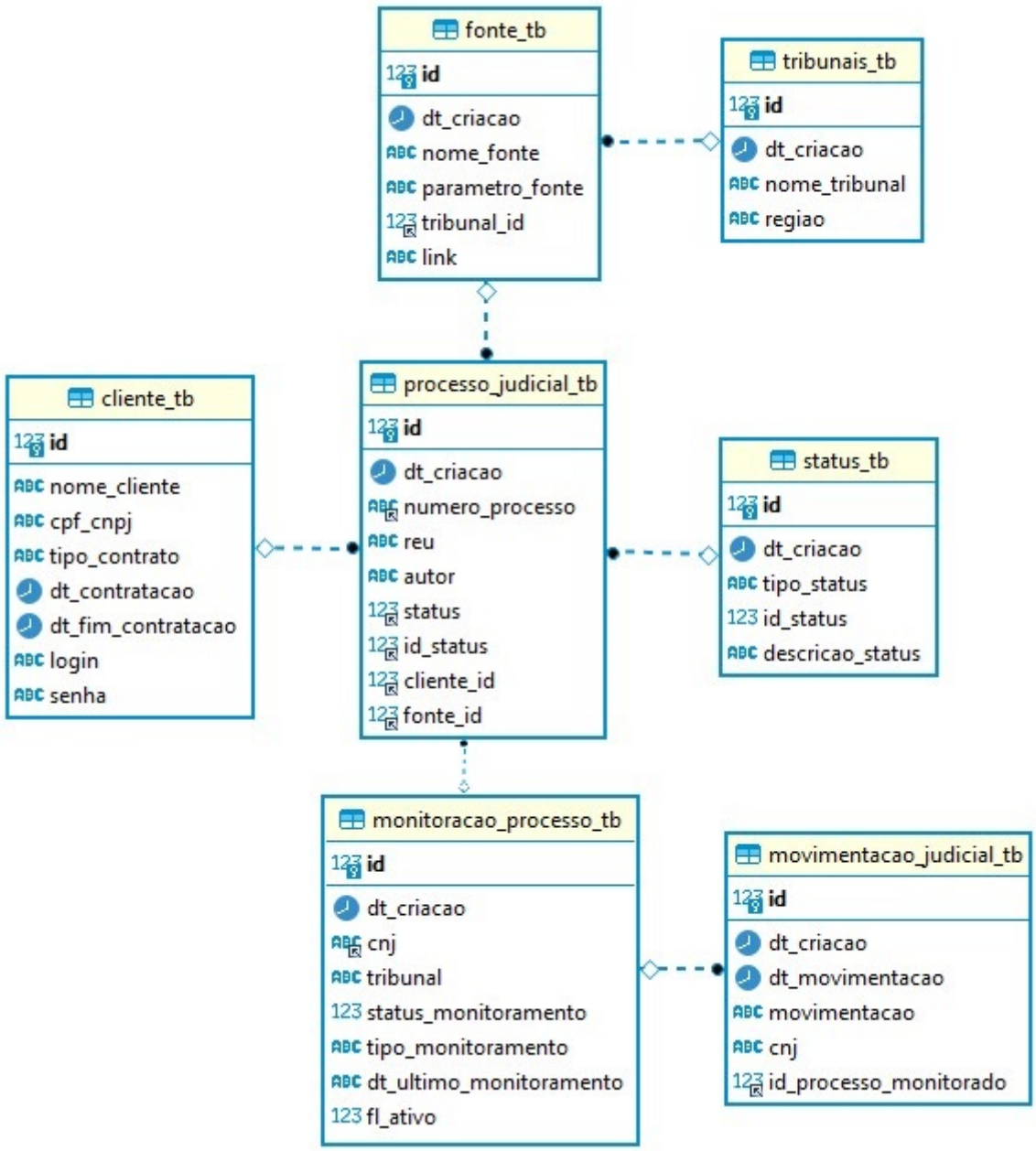

Figura 3.5: Diagrama Entidade Relacionamento do Banco de Dados utilizado no trabalho. Elaborado pelo autor.

Com a ideia de seguir uma padronização, alguns protocolos foram criados:

- Tabelas sempre serão nomeadas com o sufixo -_tb (*table)*;
- *Procedures* que possuem como objetivo atualizar dados sempre serão nomeadas com o sufixo -_spu (*storage procedure update*);
- *Procedures* que possuem como objetivo selecionar dados sempre serão nomeadas com o sufixo -_sps (*storage procedure select*);

- *Procedures* que possuem como objetivo deletar dados sempre serão nomeadas com o sufixo -_spd (*storage procedure delete*);

A tabela processo_judicial_tb é a tabela utilizada para manter as informações principais do processo, ou seja, informações relacionadas às partes e ao processo em si, como o seu número único. A partir dela, as relações são estabelecidas. A tabela fonte_tb é responsável por armazenar as informações relacionadas ao *link* em qual o processo será buscado, neste caso, o *site* que o *scraper* irá buscar as informações. Dentro das fontes, existem os tribunais para a identificação das informações referentes ao processo. A tabela status_tb foi utilizada para padronizar os *status* que podem aparecer dentro das demais tabelas e de tabelas implementadas no futuro.

Em relação às movimentações judiciais, foi criada a tabela monitoracao_processo_tb. Essa tabela é responsável por guardar as informações relativas aos dados necessários para o monitoramento do processo. Com a informação da fonte, é possível buscar diretamente da tabela de fontes o link para que o *scraper* possa buscar as movimentações no site do tribunal. Quando as movimentações são buscadas, elas são inseridas na tabela movimentacao_judicial_tb, contendo seus principais campos, que são a data de ocorrência e a movimentação em si, dados de suma importância para análise do andamento do processo na justiça.

Em relação às informações do cliente, foi criada a tabela cliente_tb. Essa tabela guarda informações de login e senha para um posterior acesso a uma tela em que será possível o acesso das informações por uma interface, além da resposta da requisição.

3.5 Monitoração

O principal objetivo do *crawler* a longo prazo é monitorar os processos e sempre manter atualizada a base de processos com as novas movimentações conforme os tribunais disponibilizam. Dessa forma, para simular o uso do *crawler*, foi programado através do gerenciador de tarefas para que o robô executasse três vezes por semana. Por preocupação com a detecção do uso de um *crawler* pelo sistema de proteção, que utiliza o recurso de *captcha*, quando a tarefa é disparada, é sorteado um entre os três tribunais para que os

andamentos sejam buscados, tornando assim o processo de consulta aleatorizado.

Para garantir que movimentações antigas não sejam inseridas novamente, foi implementado um algoritmo que realiza a comparação entre a data da movimentação e o conteúdo da movimentação mais recente para o processo já existente na base com os dados disponibilizados no site do tribunal. Caso os dados sejam divergentes, as movimentações são coletadas, se não, é considerado que não há novas movimentações para aquele processo.

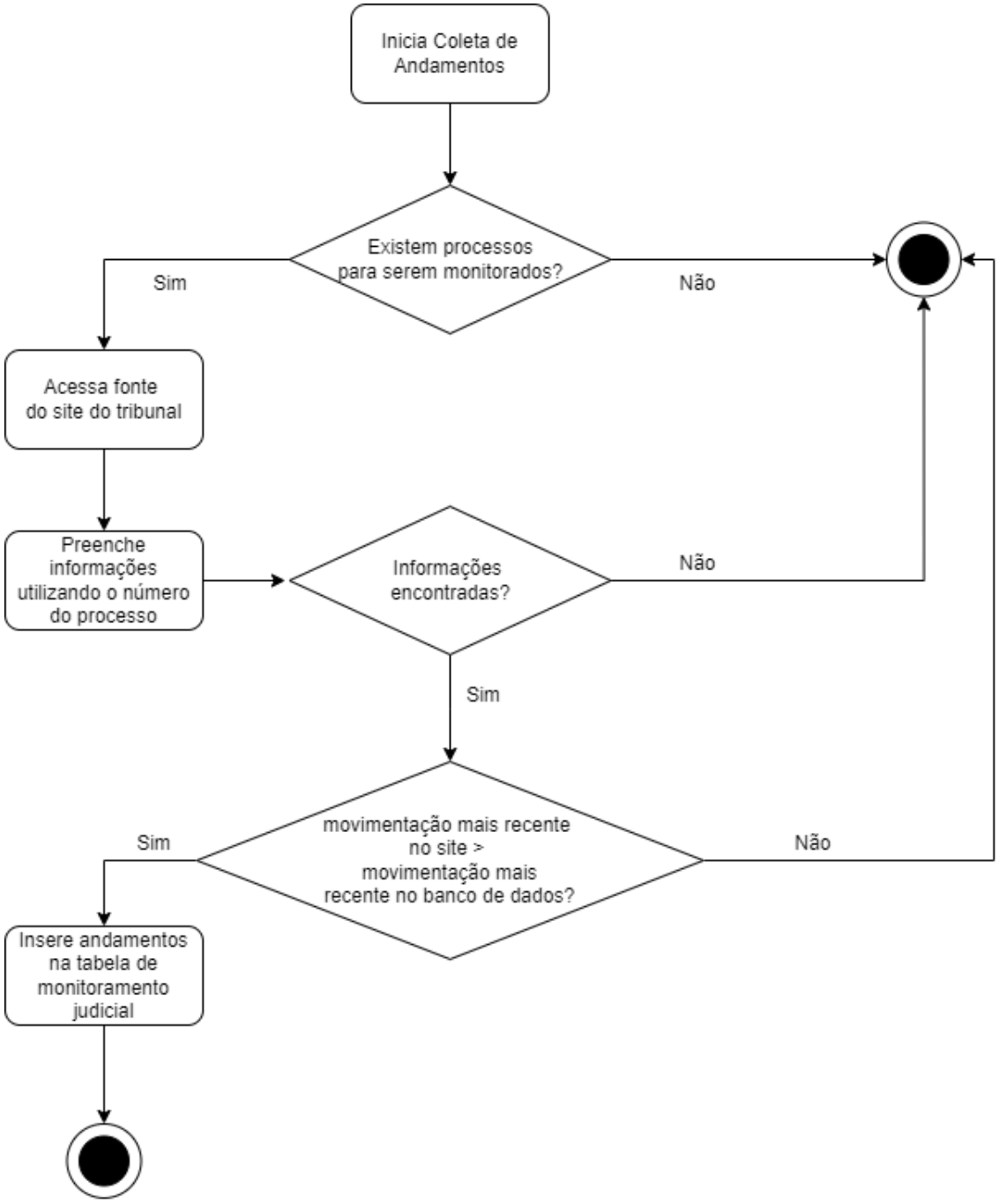

Figura 3.6: Fluxo da monitoramento de processos. Elaborado pelo autor.

3.6 Validação

Para assegurar a confiabilidade dos dados coletados, foi gerada uma planilha a partir da base coletada e enviada para uma advogada. Nessa planilha constavam as informações relacionadas ao processo, como o número do processo, a data da movimentação e o conteúdo das movimentações.

Como forma de validar as informações, foram separados 10% do total de processos de cada tribunal, com o *link* de cada tribunal para a consulta das movimentações. Assim, a advogada conseguiu cruzar as informações coletadas e as informações disponibilizadas no tribunal. Essa porcentagem foi definida para encaixar dentro da disponibilidade da advogada que realizou a validação, já que, para validar todos os processos, seria necessária uma quantidade maior de tempo.

4 Resultados e Discussão

4.1 Resultados da Implementação

Nesta pesquisa, o *crawler* precisava de duas informações fundamentais: o número do processo e o *link* para a busca desses processos. Com essas informações, através dos identificadores únicos, foi possível coletar as movimentações judiciais e as informações relacionadas às partes.

Os dados da coleta podem ser verificados na Tabela 4.1, em que o tempo de execução foi definido em minutos para todo o processo de *scraping* das movimentações judiciais, ou seja, desde a busca dos processos no banco de dados, seguido da coleta das movimentações até a finalização com a inserção no banco de dados. O pacote utilizado para a medição do tempo foi o Serilog Timings[11].

Tabela 4.1: Tabela comparativa entre os dados coletados por tribunal.

Tribunal	Quantidade de Processos	Quantidade de Movimentações	Tempo de Execução (m)
TJCE	360	1491	60
TJRR	200	57826	120
TST	570	7682	150

É necessário pontuar que os tempos de execução maiores não são proporcionais a uma quantidade maior de movimentações, já que pode haver muito mais movimentações em um processo do que em outro. Além disso, as particularidades de cada tribunal e como o *crawler* lida com elas também influencia no tempo de execução.

Outra questão relacionada ao tempo foi a necessidade da utilização de tempos de espera entre as requisições nos *sites* dos tribunais. Esse método é utilizado para suspender a execução do *thread* por um período específico de tempo. Essa escolha foi feita com dois propósitos: não sobrecarregar os *sites* dos tribunais e não ativar o *captcha*.

Neste caso, foram encontrados mais movimentações por processo no tribunal de Roraima. Apesar de possuir a maior quantidade de processos disponíveis para consulta,

[11] https://github.com/nblumhardt/serilog-timings

o Tribunal Superior do Trabalho não encontrou uma quantidade tão expressiva de movimentações se comparado com o tribunal de Roraima, como é possível perceber na Tabela 4.2.

Tabela 4.2: Tabela da quantidade de movimentações encontradas por processo.

Tribunal	Quantidade de processos encontrados	Quantidade de processos não encontrados	Média de movimentações por processo
TJCE	235	125	6
TJRR	130	70	445
TST	569	1	15

Com a validação, foi constatado que os tribunais TST e TJRR tiveram 100% das informações coletadas corretamente, com as datas de ocorrência e o contéudo das movimentações sendo exatamente os disponíveis nos *sites* dos tribunais.

No caso do TJCE, ocorreu, em alguns casos, do *scraper* coletar as informações relacionadas as petições disponíveis no site ao invés das movimentações judiciais. Esse comportamento se deve ao fato do *site* do tribunal se comportar de uma maneira em que, dependendo das informações disponíveis para a exibição, os elementos se organizam de maneira diferente, sendo necessário modificar a lógica de busca das movimentações.

A advogada notificou que, apesar de nesses casos as movimentações judiciais não terem sido coletadas, todas as informações coletadas pertenciam ao processo que estava sendo consultado, ou seja, não houve casos de informações de um processo pertencendo a outro. Como é possível observar na Tabela 4.3, segue a relação da validação feita pela advogada, com a assertividade sendo processos que tiveram suas movimentações judiciais coletadas.

Tabela 4.3: Tabela com o resultado da validação do conteúdo das movimentações judiciais.

Tribunal	Quantidade de Dados Validados	Taxa de Assertividade (%)	Movimentações corretas
TJCE	36	67,81	24
TJRR	20	100	20
TST	57	100	57

Com essa implementação, foi possível desenvolver uma API cujo objetivo é servir como uma plataforma de acesso às informações. Dentro da API, foram criadas as rotas

para que os usuários tenham acesso à diferentes informações relacionadas aos processos monitorados e as informações coletadas. Cada rota tem uma função, sendo elas:

- Busca por movimentações de um processo. Parâmetro da requisição: número único do processo. Resposta da requisição: número do processo e lista de andamentos, contendo a data de ocorrência da movimentação e o conteúdo da movimentação.


```
{
    "NumeroProcesso": "00620062320088060001",
    "ListaMovimentacoes": [
        {
            "DataCriacao": "2022-11-23T23:18:07",
            "DataMovimentacao": "2022-11-22T00:00:00",
            "Conteudo": " Certidão emitida"
        },
        {
            "DataCriacao": "2022-11-23T23:18:07",
            "DataMovimentacao": "2022-11-22T00:00:00",
            "Conteudo": " Concluso para Despacho"
        },
        {
            "DataCriacao": "2022-11-23T23:18:07",
            "DataMovimentacao": "2022-11-22T00:00:00",
            "Conteudo": " Conversão para Processo Digital"
        },
        {
            "DataCriacao": "2022-11-23T23:18:07",
            "DataMovimentacao": "2019-07-17T00:00:00",
            "Conteudo": " Arquivado Definitivamente"
        },
        {
            "DataCriacao": "2022-11-23T23:18:07",
            "DataMovimentacao": "2019-07-17T00:00:00",
            "Conteudo": " Expedição de Certidão de Arquivamento"
        }
    ]
}
```


Figura 4.1: Rota de consulta de andamentos. Elaborado pelo autor.

- Busca por movimentações recentes, ou seja, movimentações que ocorreram ou foram coletados na última semana. Parâmetro da requisição: número único do processo. Resposta da requisição: número do processo e lista de andamentos, contendo a data de ocorrência da movimentação e o conteúdo da movimentação.

```
{
    "NumeroProcesso": "01525370920188060001",
    "ListaMovimentacoes": [
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2018-08-22T00:00:00",
            "Conteudo": " Pedido de Juntada de Guia de Recolhimento"
        },
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2018-10-22T00:00:00",
            "Conteudo": " Pedido de Juntada de Guia de Recolhimento"
        },
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2019-02-28T00:00:00",
            "Conteudo": " Pedido de Suspensão"
        },
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2020-01-24T00:00:00",
            "Conteudo": " Pedido de Desarquivamento"
        },
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2020-08-25T00:00:00",
            "Conteudo": " Petições Intermediárias Diversas"
        },
```


Figura 4.2: Rota de andamentos recentes. Elaborado pelo autor.

- Consulta pela quantidade de processos monitorados para aquele cliente. Parâmetro da requisição: número do contrato. Resposta da requisição: quantidade de processos monitorados e lista com os processos.

```
{
    "NumeroProcesso": "01525370920188060001",
    "ListaMovimentacoes": [
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2018-08-22T00:00:00",
            "Conteudo": " Pedido de Juntada de Guia de Recolhimento"
        },
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2018-10-22T00:00:00",
            "Conteudo": " Pedido de Juntada de Guia de Recolhimento"
        },
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2019-02-28T00:00:00",
            "Conteudo": " Pedido de Suspensão"
        },
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2020-01-24T00:00:00",
            "Conteudo": " Pedido de Desarquivamento"
        },
        {
            "DataCriacao": "2023-01-02T17:57:48",
            "DataMovimentacao": "2020-08-25T00:00:00",
            "Conteudo": " Petições Intermediárias Diversas"
        },
```


Figura 4.3: Rota de consulta de processos monitorados. Elaborado pelo autor.

- Busca por andamentos com uma determinada palavra-chave. Parâmetro da requisição: número único do processo e palavra-chave. Resposta da requisição: número do processo e lista de andamentos contendo a data de ocorrência da movimentação, o conteúdo da movimentação com o filtro da palavra-chave.

```
{
    "NumeroProcesso": "08029106720138230010",
    "ListaMovimentacoes": [
        {
            "DataCriacao": "2022-11-07T13:35:27",
            "DataMovimentacao": "2015-10-30T00:00:00",
            "Conteudo": "ARQUIVADO DEFINITIVAMENTE"
        },
        {
            "DataCriacao": "2022-11-07T13:35:27",
            "DataMovimentacao": "2015-09-02T00:00:00",
            "Conteudo": "PROCESSO DESARQUIVADO"
        },
        {
            "DataCriacao": "2022-11-07T13:35:27",
            "DataMovimentacao": "2015-03-09T00:00:00",
            "Conteudo": "ARQUIVADO DEFINITIVAMENTE"
        },
        {
            "DataCriacao": "2022-11-07T14:34:22",
            "DataMovimentacao": "2015-10-30T00:00:00",
            "Conteudo": "ARQUIVADO DEFINITIVAMENTE"
        },
        {
            "DataCriacao": "2022-11-07T14:34:22",
            "DataMovimentacao": "2015-09-02T00:00:00",
            "Conteudo": "PROCESSO DESARQUIVADO"
        },
        {
            "DataCriacao": "2022-11-07T14:34:22",
            "DataMovimentacao": "2015-03-09T00:00:00",
            "Conteudo": "ARQUIVADO DEFINITIVAMENTE"
        }
    ]
}
```


Figura 4.4: Rota de busca por movimentações com palavra-chave. Palavra utilizada: Arquivado. Elaborado pelo autor.

- Consulta das informações relativas ao processo, como as informações de réu e autor, além do número do processo. Parâmetro da requisição: número do contrato. Resposta da Requisição: número do processo, informação sobre as partes (autor e réu), vara e comarca.

```
{
    "NumeroProcesso": "08029106720138230010",
    "Comarca": "BOA VISTA",
    "Vara": "NÃO DISPONÍVEL",
    "Reu": [
        "ANTONIO ZITO DE BARROS",
        "FRANCISCA RODRIGUES NETA"
    ],
    "Autor": [
        "TAM Linhas Aéreas S.A."
    ]
}
```


Figura 4.5: Rota de busca por informações do processo. Elaborado pelo autor.

4.2 Modelo de Negócios

Como o intuito da pesquisa é realizar um estudo de caso de uma empresa, trazendo um novo produto, se tornou necessária a ideia da criação de uma *spin-off* para que esse produto fizesse sentido dentro do ambiente atual da empresa.

4.2.1 *Business Model Canvas*

Assim, para conseguir entender a lógica para a criação e entrega de valor por essa nova empresa, foi utilizado o *Business Model Canvas*, buscando tornar mais visível o modelo de negócio da *spin-off*.

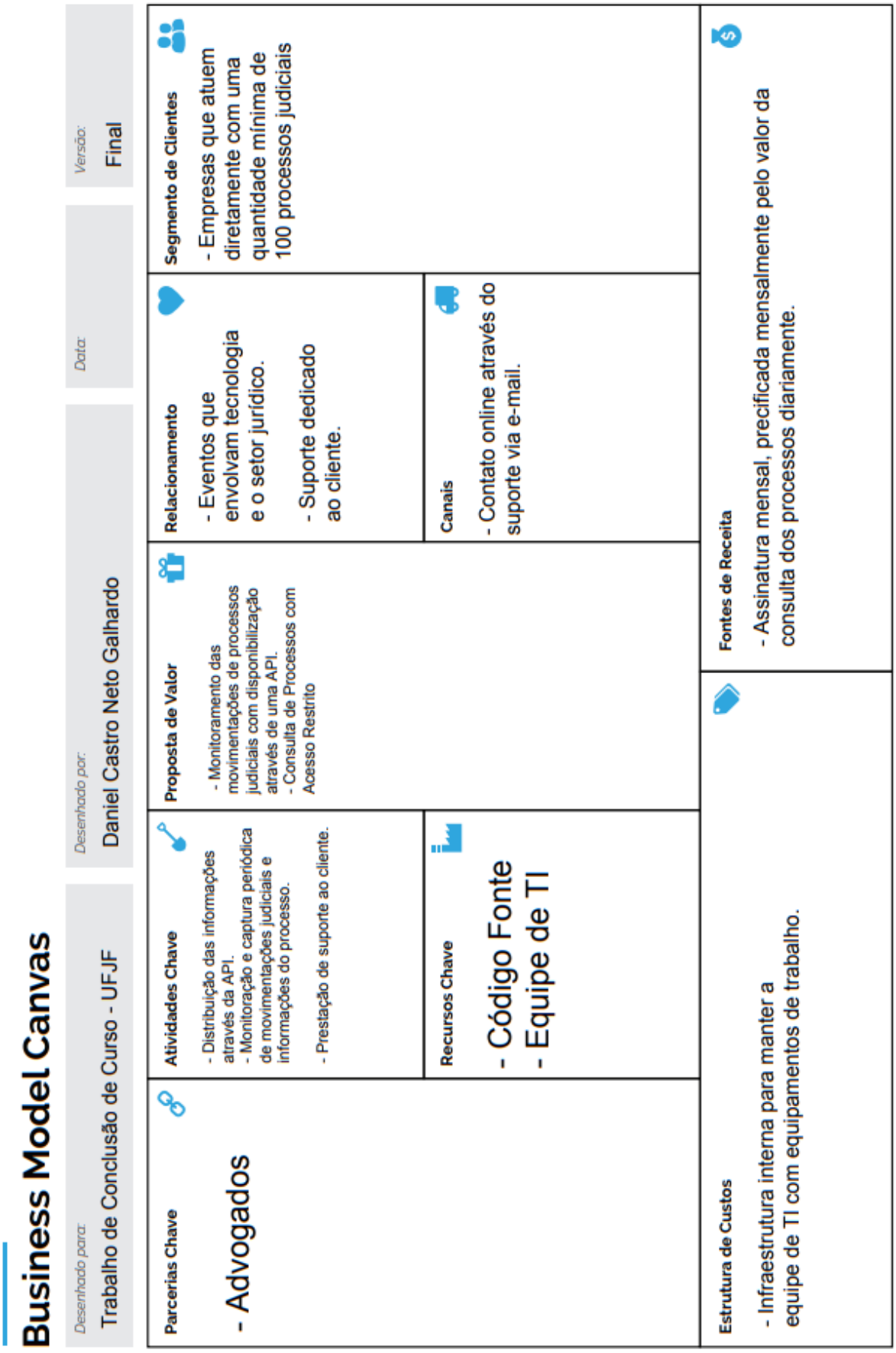

Figura 4.6: *Business Model Canvas*. Elaborado pelo autor.

Conforme o modelo de negócio traçado, o fluxo de receita giraria em torno da contratação de uma assinatura mensal do serviço de monitoração dos processos. Essa maneira se mostra mais interessante, já que permite precificar o valor considerando duas variáveis: a periodicidade da monitoração e a quantidade de processos. É importante considerar que tanto o monitoramento mensal quanto o diário possuem quantidades minímas específicas para serem contratados.

Como atividades-chave, foi proposto todo o serviço de monitoração de processos, desde a verificação nos tribunais, até a coleta e a distribuição das informações. Para isso, foram estabelecidos como recursos-chave o código-fonte e a equipe de Tecnologia de Informação (TI), sendo a proposta de valor a disponibilização das rotas de fácil integração para empresas que possuam um sistema proprietário e precisam das informações judiciais atualizadas.

4.2.2 Concorrência

Com a automação de processos se tornando algo cada vez mais presente dentro de diversas áreas da sociedade, é possível ver os benefícios que essa prática traz para o setor jurídico, para o qual há um grande volume de dados disponível para consulta. Atualmente, é possível encontrar no mercado soluções que realizam a captura de dados jurídicos, tal qual o que foi proposto neste trabalho.

Dessa forma, para efeito de comparação e verificação da viabilidade do desenvolvimento proposto nesta pesquisa, foram levantados orçamentos com três empresas do ramo, Advise, Advbox[12] e Digesto. Essas empresas cobram mensalmente por consulta diária de processo. Na Tabela 4.4, é possível visualizar as principais características consideradas para a comparação.

Tabela 4.4: Tabela comparativa do orçamento dos provedores do serviço de monitoramento processual.

Empresa	Valor Cobrado (R$)	Acesso Restrito	Mínimo de Processos
Advise	0,30	Não	Não Informado
Advbox	0,35	Não informado	100
Digesto	0,50	Não Informado	Não Informado

[12]https://blog.advbox.com.br/robos-andamentos/

Como é possível verificar na tabela, o acesso restrito não está presente ou não é informado pelas empresas que realizam esse tipo de serviço. Já nesta pesquisa, como visto no modelo do plano de negócios, há atuação de um advogado com a equipe de tecnologia justamente para viabilizar esse diferencial, já que é possível obter informações judiciais que só são disponibilizadas com este tipo de acesso.

Em relação à precificação, é necessário considerar o escopo trabalhado dentro da pesquisa, no qual a mineração de dados judiciais foi realizada em apenas três tribunais. Assim, para possuir um preço competitivo, foi considerada a cobrança por volume de processo em vez do valor por requisição. Dessa forma, as seguintes variáveis foram consideradas:

- Gasto com plano de internet;
- Gasto com AWS Lambda;
- Volume de processos a serem monitorados.

Para calcular o valor do AWS Lambda para o *scraper*, foram consideradas:

- 50 execuções por dia;
- Tempo de execução: 180 minutos;
- Alocação de memória de 256 MB.

Foram consideradas 50 execuções por dia, já que haveria repetição da execução para busca de movimentações mais atualizadas. Também foi utilizada a alocação de 256 MB considerando as operações com banco de dados e a utilização do navegador pelo Selenium para acessar os sites dos tribunais.

Já em relação à API, foram consideradas:

- 1440 requisições por dia (1 requisição por minuto);
- Tempo de execução: 30 de segundos;
- Alocação de memória de 128 MB.

Em relação à API, apesar de haver mais execuções por dia, o tempo de execução é menor, já que a única operação realizada é a de busca. Essa operação tende a ser mais rápida, mas, mesmo assim, foi considerado um tempo médio de 30 segundos para a resposta da requisição, já que a quantidade de dados na tabela pode acabar aumentando o tempo da busca.

Com esses valores, usando o calculador de custo do AWS Lambda [13], os custos seriam:

- *Scraper*: R$356,88;
- API: R$14,33;
- Plano de Internet de 650 MB: R$120,00.

Dessa forma, o valor total do custo para a execução da proposta implementada nesta pesquisa seria de R$491,21. Para conseguir cobrir os custos de execução, considerando o mínimo de 100 processos monitorados diariamente nos dias úteis, o valor da consulta diária seria de R$0,25 centavos.

É importante pontuar que, para precificar o *software* de forma mais assertiva, seria necessário considerar muitas variáveis que afetariam diretamente esse cálculo, já que precisariam ser definidos todos os custos relacionados ao desenvolvimento de um *software* como produto, o que está além do escopo da pesquisa. Além disso, foi calculado apenas o valor mínimo para a execução da rotina de monitoramento, já que o valor para lucro não faz parte do escopo desta pesquisa.

4.3 Dificuldades

Um dos principais empecilhos enfrentados durante a implementação do *crawler* foi a falta de padronização entre os tribunais utilizados. Todos eles possuíam as mesmas informações relacionadas às partes do processo, o processo e as movimentações judiciais, porém a disposição das informações era totalmente diferente.

[13] https://dashbird.io/lambda-cost-calculator/

A dificuldade de tornar o código mais versátil expôs que, para abranger uma gama ampla de tribunais, é necessária uma implementação bastante robusta, pois é preciso lidar com todas as exceções que podem ser geradas pelo *scraping* das informações em tela. Um caso pôde ser percebido no Tribunal Superior do Trabalho, em que, quando certam informações não eram preenchidas, o campo era ocultado da página. Nesse caso, a identificação através do XPATH não funcionava ou capturava outra informação da página. Com isso, foi necessário adaptar como eram estruturados os *XPATHs* para a identificação das movimentações ocorresse corretamente.

Outro empecilho encontrado durante a implementação foi a utilização de *captcha* pelos tribunais como recurso de proteção. Primeiramente foram realizados testes com o PJe, porém, o sistema de proteção utilizado por eles identificava com facilidade a utilização de um *crawler* e logo pedia a validação através do *recaptcha*. O Tribunal de Justiça de Roraima também possui um sistema, mas para contornar essa situação foi criada uma função capaz de simular a digitação humana, digitando com tempo entre cada carácter inserido.

Além disso, durante a realização de testes usando o Tribunal de Justiça do Rio de Janeiro, foi verificado que a disponibilização era feita de maneira direta, em um HTML praticamente puro, sem a utilização de divisões que facilitam a identificação das informações. Para contornar esse problema seria necessário detectar palavras que apontassem uma movimentação e identificar o fim das informações relacionadas a movimentação. Por questões relacionadas a otimização de tempo, esse tribunal foi descartado.

Por fim, a última dificuldade relacionada à implementação dos *crawlers* foi referente ao Tribunal de Justiça do Distrito Federal. Como a abordagem utilizada nesse trabalho para *scraping* foi através dos identificadores únicos, era necessário que os *XPATHs* fossem visíveis para a captura das informações. Nesse tribunal, porém, quando o Selenium utilizava o XPATH para identificar uma informação disponibilizada na página, ele era disponibilizado com toda a sintaxe do XPATH substituida pela palavra *hidden*, impossibilitando assim o *scraping* das informações.

5 Conclusões

Este trabalho buscou desenvolver um *software* capaz de monitorar e capturar as movimentações judiciais, realizando um estudo de caso para verificar a viabilidade do produto, as fases de implementação, testes e criação de um modelo de negócio para entender a melhor forma de capturar valor com o *software* desenvolvido.

Primeiramente, foram definidos os três tribunais que seriam buscados, sendo eles o TJCE, TJRR e o TST. A partir dessa escolha, foram escolhidos os números de processos públicos cedidos pela empresa Ativa Gerenciamento de Recursos. A partir daí, foram consultados os *links* para os *sites* para a busca das movimentações judiciais.

Durante o desenvolvimento do *scraper* foram encontrados diversas adversidades, sendo a principal delas o fato dos tribunais não possuírem muitos elementos em comum. Para resolver essa questão, foi necessário manipular os identificadores únicos de maneira a atender o comportamento do site para a coleta das movimentações corretamente.

Com a identificação e coleta dos dados funcionando corretamente, estes foram armazenados no serviço desenvolvido para uso posterior. Assim, foi criado o banco de dados estruturado para manter os relacionamentos entre as tabelas o mais simples e objetivo possível. Com isso, foi possível executar o processo de monitoração e coleta dos dados judiciais, armazenando as informações no banco de dados.

Com os dados coletados, foi criada uma API utilizando o design DDD. Essa API possuía rotas para a distribuição das informações. Assim, foi finalizada a etapa de implementação do trabalho.

Referente ao *software* como serviço, foi estabelecido uma quantidade mínima de processos para que fosse encontrado um valor mínimo que cobrisse os gastos da execução. Para que uma análise mais detalhada do *software* como um produto fosse feita, seria necessário o desenvolvimento da ferramenta mais completa, abrangendo todos os tribunais e suas respectivas fontes, já que o tempo em que está pesquisa foi desenvolvida não contemplou toda a implementação. Por fim, foi criado um plano de negócio para buscar a melhor maneira de capturar e entregar valor para a empresa Ativa Gerenciamento

de Recursos, através da criação de uma *spin-off* no caso da utilização deste *software* como serviço. Além disso, foi feita uma comparação entre os orçamentos de diferentes prestadoras do serviço proposto neste trabalho para verificar o custo mínimo para que o produto se tornasse viável.

Em suma, foi visto que é possível implementar uma solução para a monitoração de andamentos judiciais, coletando uma quantidade expressiva de informações num período muito mais ágil do que uma pessoa faria. Com essa implementação, a empresa resolveria o problema referente a dependência de fornecedores, já que ela mesma conseguiria trazer essas informações.

6 Perspectivas Futuras e Desafios

Para trabalhos posteriores, existem várias oportunidades que podem ser exploradas dentro do cenário do *scraping* de informações jurídicas. No contexto deste trabalho, o foco foi a monitoração de movimentações judiciais oriundas de processos públicos. Para esse caso, existem algumas sugestões de possíveis melhorias e implementações que possam ser feitas de modo a contribuir com o que foi proposto nesta pesquisa.

Primeiramente, uma parte que pode ser explorada é a de identificação dos tribunais a partir da numeração única do processo. É possível criar maneiras de identificar a qual tribunal o processo pertence realizando a quebra do número. Com isso, não seria necessário a indicação de qual tribunal o processo pertence, possibilitando vincular o processo a um tribunal e posteriormente encontrando em qual *link* do tribunal as informações daquele processo podem ser encontradas.

Com a implementação acima, outro desenvolvimento seria o de identificar a fonte correta do processo. Hoje em dia não é possível saber qual o *link* correto para acessar as informações dos processos, dessa forma, é necessário por tentativa e erro buscar em quais *sites* dos tribunais esses processos se encontram.

Outra oportunidade é a da criação de uma interface em que o usuário possa interagir, ao invés apenas da distribuição em formato JSON pela API. Com isso, se tornaria muito mais intuitivo e muito mais atrativo para o usuário, um sistema em que ele possa visualizar de forma mais fácil as informações relacionadas aos processos monitorados.

Por fim, outra sugestão seria utilizar do processamento de linguagem natural para realizar uma análise do texto. Como grandes volumes de dados em formato de texto são capturados, seria interessante a utilização desse tipo de processamento para identificar e interpretar o contexto de certos dados de movimentação, conseguindo identificar, por exemplo, uma situação em que um processo está prestes a ser arquivado, algo que pode ser de bastante interesse para um advogado.

Bibliografia

AALST, W. M. van der; BICHLER, M.; HEINZL, A. Robotic process automation. In: . Germany: Springer Nature, 2018. v. 60, p. 269–272.

ALBERTH, M.; MATTERN, M. Understanding robotic process automation (rpa). In: . Antwerp, Belgium: Capco Institute, 2017. v. 46, p. 54–61.

ALVES, F. K. T. Advocacia 4.0: o uso de softwares que produzam conteúdo jurídico nos escritórios de advocacia. In: . Brasília, Brazil: Instituto Brasiliense de Direito Público, 2021. v. 1.

ALVES, F. S. Um estudo das startups no brasil. In: . Bahia, Brazil: Universidade Federal da Bahia, 2013.

CALÒ, A. Extração e análise de informações jurídicas públicas. *Trabalho de Conclusão de Curso-Bacharelado em Ciência da Computação no Instituto de Matemática e Estatística da Universidade de São Paulo*, 2014.

DHENAKARAN, S.; SAMBANTHAN, K. T. Abstract web crawler - an overview. In: . California, United States: Computer Science & Electronics Journals, 2011. v. 2.

FIELDING, R. T. Architectural styles and the design of network-based software architectures. In: . California, United States of America: University of California, 2000.

FRANCO, I. S. A. A influência da tecnologia na busca pela celeridade e efetividade processual, à luz da lei n.11.419-06. In: . Goiânia, Brazil: Faculdade Alfredo Nasser, 2016. v. 3.

GONCALVES, J. E. L. A necessidade de reinventar as empresas. In: . São Paulo, Brazil: Escola de Administração de Empresas de São Paulo da Fundação Getúlio Vargas, 1998. v. 38.

GUARDIA, G.; FABIANO, I. A.; NASCIMENTO, M. B. do. Iarpje: Automação de processos no trf1 para o auxílio emergencial. *Concilium*, v. 22, n. 1, p. 96–109, 2022.

GUNAWAN, R.; RAHMATULLOH, A.; DARMAWAN, I.; FIRDAUS, F. Comparison of web scraping techniques : Regular expression, html dom and xpath. In: *Proceedings of the 2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018)*. Atlantis Press, 2019/03. p. 283–287. ISBN 978-94-6252-689-1. ISSN 2589-4943. Disponível em: ⟨https://doi.org/10.2991/icoiese-18.2019.50⟩.

KAUSAR, M. A.; DHAKA, V. S.; SINGH, S. K. Article: Web crawler: A review. In: . [S.l.]: IJCA Journal, 2013. v. 63, p. 31–36.

MACORATTI, J. C. *ASP .NET Core - API com VS Code e conceitos do DDD*. 2020. Disponível em: ⟨https://www.macoratti.net/20/07/aspnc_ucddd1.htm⟩.

MONTENEGRO, M. C. Exigência do uso de processo eletrônico deve acelerar extinção dos processos em papel. In: . Brazil: Escola Superior da Magistratura do Amazonas Noticias, 2022.

NAEEM, T. *Compreendendo os fundamentos das APIs REST*. 2020. Disponível em: ⟨https://www.astera.com/pt/tipo/blog/definicao-de-api⟩.

OLIVEIRA, E. S. de. Breves considerações sobre o processo judicial eletrônico. In: . Brazil: Tribunal de Justiça do Estado do Amazonas, 2017.

PINHO, H. D. B. de. Manual de direito processual civil contemporâneo. In: . Brasília, Brazil: Saraiva Educação, 2019. v. 1.

PROVAR. *Compreendendo os fundamentos das APIs REST*. 2022. Disponível em: ⟨https://www.provartesting.com/documentation/page-objects/introduction-to-xpaths/⟩.

RAMOS, J. D. A. Protótipo de um software para a classificação de processos, conforme as tabelas processuais unificadas do conselho nacional de justiça. In: . Palmas, Tocatins: UNIVERSIDADE FEDERAL DO TOCANTINS, 2020. v. 1.

SALMERÓN-MANZANO, E. Legaltech and lawtech: Global perspectives, challenges, and opportunities. *Laws*, MDPI, v. 10, n. 2, p. 24, 2021.

SEQUEIRA, A. L. R. O. Spin-off em pequenas e médias empresas : estudo de caso. In: . Coimbra, Portugal: Univeristy of Coimbra, 2013.

SILVA, R.; DIAS, T.; SEGUNDO, W. C. Uma estratégia para a identificação e extração de dados de patentes brasileiras. In: *Anais do XVIII Congresso Latino-Americano de Software Livre e Tecnologias Abertas*. Porto Alegre, RS, Brasil: SBC, 2021. p. 31–36. ISSN 0000-0000. Disponível em: ⟨https://sol.sbc.org.br/index.php/latinoware/article/view/19902⟩.

VELOSO, P. R. N. Desenvolvimento e aplicação de sistemas de conhecimento no setor jurídico. In: . Florianópolis, SC , Brazil: Centro Tecnológico da Universidade Federal de Santa Catarina., 2009. v. 1.

ZHAO, B. Web scraping. In: . Corvallis, OR,United States of America: Springer Internacional, 2017. p. 1–3.

A Apêndice - Termo de Concessão de Uso de Imagem

Termo Cessão de Uso de Imagem

Eu, Guilherme Ribeiro Martins, na qualidade de Presidente da empresa Ativa Gerenciamento de Recursos (“Empresa”), autorizo Daniel Castro Neto Galhardo (“Pesquisador”) da Universidade Federal de Juiz de Fora, a citar, no âmbito do projeto de pesquisa intitulado Desenvolvimento, implementação e validação de um scraper para a captura de dados jurídicos dos tribunais do Brasil, o nome da Empresa.

Juiz de Fora, 12 de janeiro de 2023

ATIVA GERENCIAMENTO DE RECURSOS LTDA:03871618000115

Assinado de forma digital por ATIVA GERENCIAMENTO DE RECURSOS LTDA:03871618000115
Dados: 2023.01.12 15:44:32 -03'00'

Guilherme Ribeiro Martins

gov.br
Documento assinado digitalmente
DANIEL CASTRO NETO GALHARDO
Data: 16/01/2023 15:09:08-0300
Verifique em https://verificador.iti.br

Daniel Castro Neto Galhardo

B Apêndice - Termo de Concessão de Dados

Termo de Compromisso de Cessão de Dados Para Pesquisa

Eu, Guilherme Ribeiro Martins, na qualidade de Presidente da empresa Ativa Gerenciamento de Recursos ("Empresa"), autorizo Daniel Castro Neto Galhardo ("Pesquisador") da Universidade Federal de Juiz de Fora, a utilizar, no âmbito do projeto de pesquisa intitulado Desenvolvimento, implementação e validação de um scraper para a captura de dados jurídicos dos tribunais do Brasil, numerações de processos no padrão CNJ, escolhidos de forma aleatória do Banco de Dados da Empresa.

O Pesquisador se compromete em manter a confidencialidade dos dados coletados, bem como com a privacidade de seus conteúdos.

O Pesquisador declara entender que é de sua a responsabilidade de cuidar da integridade das informações e de garantir a confidencialidade dos dados.

Por fim, o Pesquisador compromete-se com a guarda, cuidado e utilização das informações apenas para cumprimento dos objetivos previstos na pesquisa acima referida, não os expondo no trabalho final.

Juiz de Fora, 12 de janeiro de 2023

ATIVA GERENCIAMENTO DE RECURSOS LTDA:03871618000115

Assinado de forma digital por ATIVA GERENCIAMENTO DE RECURSOS LTDA:03871618000115
Dados: 2023.01.12 16:20:51 -03'00'

Guilherme Ribeiro Martins

Documento assinado digitalmente
gov.br DANIEL CASTRO NETO GALHARDO
Data: 16/01/2023 15:09:08-0300
Verifique em https://verificador.iti.br

Daniel Castro Neto Galhardo