Instituto Federal da Bahia
Departamento Acadêmico de Computação

Programa de Pós-Graduação em Engenharia de Sistemas de Produtos

SOPHIADATA: GOVERNANÇA E PUBLICAÇÃO DE DADOS GOVERNAMENTAIS ABERTOS

Mario Jorge Pereira

DISSERTAÇÃO DE MESTRADO

Salvador
14 de fevereiro de 2025

MARIO JORGE PEREIRA

SOPHIADATA: GOVERNANÇA E PUBLICAÇÃO DE DADOS GOVERNAMENTAIS ABERTOS

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas de Produtos da Instituto Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Engenharia de Sistemas e Produtos.

Orientador: Prof. Dr. Renato Novais
Coorientador: Prof. Dr. Pablo Vieira Florentino

Salvador
14 de fevereiro de 2025

FICHA CATALOGRÁFICA ELABORADA PELO SISTEMA DE BIBLIOTECAS DO IFBA, COM OS DADOS FORNECIDOS PELO(A) AUTOR(A)

S712 Pereira, Mario Jorge

Sophiadata: Governança e Publicação de Dados Governamentais Abertos: / Mario Jorge Pereira; orientador Renato Novais; coorientador Pablo Florentino -- Salvador : IFBA, 2025.

86 p.

Dissertação (Programa de Pós-Graduação em Engenharia de Sistemas e Produtos - PPGESP) -- Instituto Federal da Bahia, 2025.

1. Governança e Publicação de Dados Governamentais Abertos. 2. Qualidade e Padronização de Metadados. 3. Conformidade com a LGPD na Publicação de Dados. I. Novais, Renato, orient. II. Florentino, Pablo, coorient. III. TÍTULO.

CDD/CDU

TERMO DE APROVAÇÃO

MARIO JORGE PEREIRA

SOPHIADATA: GOVERNANÇA E PUBLICAÇÃO DE DADOS GOVERNAMENTAIS ABERTOS

Esta Dissertação de Mestrado foi julgada adequada à obtenção do título de Mestre em Engenharia de Sistemas e Produtos e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Sistemas de Produtos da Instituto Federal da Bahia.

Salvador, 14 de fevereiro de 2025

Prof. Dr. Renato Lima Novais
Instituto Federal da Bahia

Prof. Dr. Pablo Vieira Florentino
Instituto Federal da Bahia

Prof. Dr. Francisco José da Silva Borges de Santana
Instituto Federal da Bahia

Prof. Dr. Eduardo Manuel de Freitas Jorge
Universidade Estadual da Bahia

Prof. Dr. Daniel Xavier de Sousa
Queen's University - Kingston, ON, Canada

À minha esposa, Neila, e aos meus filhos, Pedro Lucas e Maria Sofia, minha razão e inspiração.
Aos professores Renato Novais e Pablo Florentino, por sua orientação e apoio.

AGRADECIMENTOS

Agradeço, em primeiro lugar, à minha esposa, Neila Ribeiro Santos Pereira, por sua presença constante, paciência e apoio incondicional. Sem o seu amor e compreensão, essa jornada teria sido infinitamente mais difícil. Neila, você foi minha força nos momentos de dúvida e minha paz nos dias de tensão. Obrigado por ser minha parceira em todas as circunstâncias.

Aos meus filhos, Pedro Lucas e Maria Sofia, que são a razão pela qual busco sempre ser melhor e acreditar em um futuro feito de conhecimento e perseverança. Vocês foram minha inspiração diária, e este trabalho é também uma promessa de tudo o que acredito para vocês e para o nosso futuro.

Aos professores Renato Novais, meu orientador, e Pablo Florentino, meu coorientador, sou profundamente grato pela confiança e apoio. Obrigado por compartilharem seu conhecimento, por sua paciência e pela orientação generosa ao longo de todo este percurso. A experiência e o exemplo de cada um de vocês foram fundamentais para que este trabalho chegasse até aqui.

Aos meus amigos e colegas de jornada, por estarem ao meu lado, oferecendo companhia e apoio nos momentos em que as exigências pareciam maiores do que a força. Compartilhar este caminho com vocês tornou o processo mais leve e cheio de aprendizado.

Aos meus familiares, pelo incentivo e pelas palavras de apoio que me impulsionaram a seguir em frente. O carinho e a confiança de cada um de vocês foram como um porto seguro ao longo desta caminhada.

Finalmente, agradeço à Deus pela determinação e perseverança, que, mesmo nas noites mais longas, me fez acreditar que valia a pena.

As pessoas que são loucas o suficiente para achar que podem mudar o mundo são aquelas que o mudam.

—APPLE (Texto do comercial "To the crazy ones" de 1997)

RESUMO

A crescente demanda por dados abertos governamentais, impulsionada pela transparência, inovação e eficiência nos serviços públicos, tem enfrentado desafios relacionados à qualidade, governança e conformidade com as regulamentações. Este trabalho apresenta o desenvolvimento do Sophiadata, uma solução computacional *web* integrada que visa otimizar a gestão, o tratamento e a publicação de dados abertos governamentais, alinhando-se aos princípios FAIR (Findable, Accessible, Interoperable, Reusable) e à Lei Geral de Proteção de Dados (LGPD).

A ferramenta foi projetada para superar obstáculos técnicos encontrados na extração, tratamento e governança de dados provenientes de fontes heterogêneas, apoiar a conformidade regulatória e promover a transparência e a reutilização dos dados. Através de funcionalidades como a classificação de dados sensíveis e a governança de metadados, o Sophiadata facilita o processo de validação e a organização de dados em conformidade com a LGPD.

A validação da ferramenta, realizada por meio de simulações práticas, demonstrou que o Sophiadata não apenas simplifica a gestão de dados, mas também melhora a qualidade dos dados publicados em portais governamentais, tornando-os mais acessíveis e reutilizáveis.

Palavras-chave: Dados, Metadados, FAIR, Governança de Dados, LGPD.

ABSTRACT

The growing demand for open government data, driven by transparency, innovation, and efficiency in public services, has encountered challenges related to data quality, governance, and compliance with regulations. This work presents the development of Sophiadata, an integrated computational solution designed to optimize the management, processing, and publication of open government data, aligning with the FAIR principles (Findable, Accessible, Interoperable, Reusable) and the General Data Protection Law (LGPD).

The tool was designed to overcome technical challenges in the extraction, processing, and governance of data from heterogeneous sources, supporting regulatory compliance and promoting transparency and data reuse. Through functionalities such as the classification of sensitive data and metadata governance, Sophiadata facilitates the validation process and the organization of data in compliance with the LGPD.

Validation of the tool, conducted through practical simulations, showed that Sophiadata not only simplifies data management but also improves the quality of data published on government portals, making them more accessible and reusable.

Keywords: Data, Metadata, FAIR, Data Governance, LGPD

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

Capítulo 1

Neste capítulo, são apresentados o contexto e a motivação para a realização deste trabalho, bem como o problema e a justificativa do estudo. São apresentados também os objetivos, os resultados esperados, os procedimentos metodológicos e as limitações da pesquisa. Por fim, é feita uma apresentação da organização do texto.

INTRODUÇÃO

1.1 CONTEXTO E MOTIVAÇÃO

Desde o surgimento do movimento de governo aberto, que ganhou força no final dos anos 2000, diversos países adotaram práticas voltadas para a transparência, participação cidadã e colaboração. Essas políticas objetivam aumentar a responsabilização das administrações públicas, melhorar a eficiência dos serviços públicos e promover a inovação (OECD, 2014).

Com a digitalização das administrações públicas e o uso crescente de ferramentas de comunicação digital, como portais de transparência e plataformas de dados abertos, o volume de dados públicos disponíveis cresceu exponencialmente. Esses dados abrangem uma vasta gama de áreas, incluindo saúde, educação, segurança, meio ambiente e finanças públicas (EVANS; CAMPOS, 2009).

Embora o aumento do volume de dados disponíveis ofereça inúmeras oportunidades para a inovação e melhoria dos serviços públicos, ele também apresenta desafios significativos. Entre eles estão a qualidade dos dados, a proteção da privacidade dos indivíduos e a garantia de que todas as camadas da população tenham acesso igualitário a essas informações (OECD, 2014).

Segundo Pinho (2021), a disponibilização de dados abertos governamentais enfrenta uma série de barreiras que precisam ser abordadas para que a prática seja eficaz e benéfica tanto para os governos quanto para os cidadãos. As principais barreiras encontradas são:

1. Baixa qualidade dos dados publicados por órgãos governamentais;

2. Falta de recursos humanos capacitados para trabalhar com dados abertos dentro das organizações;

3. Falta de investimento dessas organizações em gestão de dados;

4. Falta de infraestrutura necessária para trabalhar com dados.

Para superar as barreiras apontadas por Pinho (2021), é fundamental adotar um processo estruturado que envolva as etapas de descoberta, extração, limpeza, pré-processamento, armazenamento e posterior disponibilização dos dados ao consumidor. Durante a fase de extração, surgem desafios significativos devido à heterogeneidade das fontes e ao formato diversificado dos dados, o que dificulta o desenvolvimento de uma solução única para sua aquisição. Além disso, a obtenção e o tratamento de dados provenientes de diferentes provedores geralmente exigem habilidades avançadas de programação, conforme destacado por Albuquerque, Alves e Maciel (2022). Essa complexidade reforça a necessidade de ferramentas e metodologias que facilitem o trabalho com dados heterogêneos.

Após a extração, para que os dados se tornem verdadeiramente úteis, não basta simplesmente coletá-los de fontes internas ou externas às instituições: é indispensável tratá-los de maneira apropriada (BHAGESHPUR, 2019). Entretanto, estabelecer um processo padronizado de tratamento de dados é um desafio complexo, já que as etapas de limpeza e preparação frequentemente precisam ser customizadas para atender aos requisitos específicos de cada caso de uso. Essa necessidade de personalização é ainda mais evidente quando se considera o impacto de dados inadequados nas análises realizadas. Dados de baixa qualidade, afetando tanto organizações empresariais quanto o governo dos Estados Unidos, conforme destacado por Ilyas e Chu (2019), podem comprometer seriamente a precisão das análises, levando a decisões erradas ou imprecisas, que resultam em prejuízos estimados em trilhões de dólares anualmente. Esse cenário ressalta a necessidade de um tratamento meticuloso, garantindo que os dados possuam qualidade e confiabilidade para suportar análises e decisões assertivas.

Além do tratamento, é importante considerar o armazenamento dos dados e entender que os dados não se esgotam ao serem processados ou entregues aos seus usuários. Eles permanecem armazenados, possibilitando seu reaproveitamento para diversas finalidades, como visualização de informações ou treinamento de modelos de aprendizado de máquina (ALBUQUERQUE; ALVES; MACIEL, 2022; SOUZA, 2021; ILYAS; CHU, 2019). Essa reutilização contínua, aliada ao crescimento exponencial dos volumes de dados, exige estratégias eficazes de gestão e armazenamento para atender às demandas crescentes e garantir que os dados tratados se tornem insumos valiosos.

Outro ponto está relacionado à distribuição dos dados. Atualmente, os dados estão sujeitos a restrições legais e de acesso, como as estabelecidas pela LGPD (BRASIL, 2018), que exige um controle rigoroso sobre o propósito de uso e o consentimento para o tratamento de dados pessoais. Essas restrições tornam indispensável a implementação de mecanismos de controle individualizado que assegurem conformidade com os regulamentos legais e proteção da privacidade dos indivíduos. Paralelamente, é igualmente essencial facilitar o acesso aos dados de forma responsável, promovendo sua utilização para a descoberta de novos conhecimentos e inovação (STACH, 2023).

A relevância dos dados é amplamente reconhecida também pela comunidade acadêmica, destacando a necessidade de aprimorar os processos de gerenciamento e compartilhamento de dados provenientes de pesquisas e projetos. Esse tema ganhou maior visibilidade em

2016 com a publicação do artigo *"The FAIR Guiding Principles for scientific data management and stewardship"* no periódico *Scientific Data*. O trabalho dos autores teve como objetivo estabelecer diretrizes para garantir que os dados sejam encontráveis (*Findable*), acessíveis (*Accessible*), interoperáveis (*Interoperable*) e reutilizáveis (*Reusable*), princípios amplamente conhecidos pela sigla FAIR, que seguem sendo referência no campo da ciência de dados (GO-FAIR, 2022).

Em suma, conclui-se que lidar com os desafios relacionados aos dados requer abordagens inovadoras que integrem tratamento adequado, armazenamento eficiente e distribuição controlada, sem comprometer a acessibilidade e a segurança dos dados.

1.2 PROBLEMA E JUSTIFICATIVA

Diante do crescimento da publicação de dados abertos governamentais no Brasil, surge o desafio de garantir a conformidade com a LGPD (BRASIL, 2018) e a Lei de Acesso à Informação (LAI) (BRASIL, 2011), ao mesmo tempo em que se busca reduzir as dificuldades técnicas relacionadas à governança de metadados e ao tratamento de dados. Esse problema é ampliado pela limitação, dentro do escopo desta pesquisa, de soluções que contemplem todas as etapas do gerenciamento de dados, com foco nas fases iniciais de extração e tratamento.

Atualmente, ferramentas como o *Comprehensive Knowledge Archive Network* (CKAN) e o *DataHub* oferecem suporte robusto para catalogação e governança de dados. O CKAN facilita a publicação e organização de dados abertos, enquanto o DataHub se destaca pela descoberta e governança de dados (Open Knowledge Foundation, 2024; DataHub Project, 2024). Apesar disso, ambas as plataformas apresentam limitações importantes, como a ausência de recursos integrados para extração de dados diretamente de fontes primárias. No entanto, ambas as plataformas carecem de mecanismos integrados para a extração automatizada de dados de fontes primárias e demandam conhecimentos técnicos avançados para configuração e uso, tornando sua adoção menos acessível a usuários sem especialização na área.

A falta de ferramentas que simplifiquem a etapa de extração de dados cria uma barreira significativa. Sem processos eficientes e acessíveis, as organizações enfrentam desafios para gerenciar, atualizar, localizar e compartilhar seus dados de forma eficaz. Isso compromete a qualidade e a utilidade dos dados, limitando seu impacto em processos analíticos e na tomada de decisão.

Portanto, há uma necessidade de uma solução integrada que vá além da catalogação e governança de dados. Essa solução deve incluir mecanismos eficientes para extração e tratamento de dados, aproximando-se dos princípios FAIR. Assim é possível agregar mais valor aos ativos de dados e promover sua utilização de forma ampla e eficiente.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Desenvolver uma solução computacional materializada em software web que auxilie na gestão, no tratamento e na publicação de dados abertos governamentais.

1.3.2 Objetivos Específicos

1. Identificar as dificuldades técnicas no processo de extração e tratamento de dados em órgãos governamentais;
2. Desenvolver funcionalidades que permitam a classificação dos dados com base na LGPD;
3. Desenvolver funcionalidades que auxiliem e simplifiquem o tratamento de dados, assegurando proteção e transparência;
4. Implementar ferramentas para governança de metadados, com foco na documentação e validação;
5. Avaliar o impacto do software no processo de gestão e tratamento de dados abertos.

1.4 PROCEDIMENTOS METODOLÓGICOS

1.4.1 Cenário de Pesquisa

Nesta pesquisa, adotou-se uma abordagem bibliográfica, fundamentada em uma revisão da literatura, além de examinar publicações tradicionais como artigos, teses, livros, revistas, anais e relatórios, incluindo arquivos de órgãos públicos e bancos de dados (WAZLAWICK, 2009). Buscando uma compreensão do cenário em questão, foram empregados métodos quantitativos e qualitativos. Embora haja um caráter descritivo na apresentação do cenário, o objetivo principal é de natureza aplicada, o estudo tem uma abordagem propositiva e avaliativa, pois projeta e constrói um o software e valida empiricamente sua eficácia e eficiência. O rigor e a imparcialidade foram primordiais na análise e interpretação dos dados, conforme definido por Gil (2008).

1.4.2 Método

O desenvolvimento deste trabalho foi realizado através de um processo iterativo e incremental. Cada fase do projeto foi constantemente revisada e aprimorada, garantindo um desenvolvimento contínuo e adaptável (PRESSMAN; MAXIM, 2021). As fases incluíram:

1. **Revisão bibliográfica:** Nesta fase, foi conduzido um estudo da literatura existente nas áreas que necessitaram de maior investigação e esclarecimento;
2. **Planejamento da construção da solução:** Foram planejadas estratégias e métodos para abordar o problema identificado;
3. **Desenvolvimento de solução computacional para o problema identificado:** Uma solução tecnológica foi elaborada e constantemente aprimorada, visando atender às necessidades específicas do problema em questão;
4. **Realização de validação experimental:** Experimentos foram conduzidos para testar e validar a eficácia da solução proposta, garantindo que ela atendessem aos requisitos e expectativas;

5. **Apresentação dos resultados:** Os resultados obtidos foram analisados, compilados e apresentados, proporcionando *insights* valiosos e direcionando às futuras iterações e melhorias do projeto.

1.4.3 Validação

Para determinar o alcance dos objetivos, foram conduzidos experimentos seguidos de uma análise criteriosa dos resultados. Esta abordagem permitiu verificar efetivamente os objetivos propostos.

1.5 ORGANIZAÇÃO DO TEXTO

O trabalho está organizado da seguinte maneira: no Capítulo 2 tem-se a revisão da literatura, onde é explorado as pesquisas e desenvolvimentos anteriores relacionados à qualidade dos dados, governança e as soluções associadas. Esta revisão estabeleceu as bases da pesquisa e do desenvolvimento da solução.

Na sequencia, o Capítulo 3 apresenta o Projeto. Nesse capítulo é detalhado a arquitetura da solução, funcionalidades e a ferramenta desenvolvida para atender de forma satisfatória às necessidades levantadas na pesquisa.

Os Capítulos 4 e 5 apresentam as validações realizadas, sob a forma de experimentos. O Capítulo 4 utiliza o FAIR para avaliar o produto criado pela ferramenta, enquanto o Capítulo 5 concentra-se em analisar sua usabilidade. Esses procedimentos foram fundamentais para fornecer uma análise concreta do impacto prático do software, demonstrando sua eficácia, eficiência e benefícios.

Por fim, o Capítulo 6 apresenta as reflexões sobre o trabalho, o impacto e a relevância da ferramenta desenvolvida, considerações sobre trabalhos futuros e possíveis aprimoramentos.

Capítulo 2

Este Capítulo fornece uma visão geral da evolução do uso de dados em relação ao tempo e o volume. Apresentamos os conceitos mais importantes para elaboração do trabalho.

REVISÃO BIBLIOGRÁFICA

2.1 DADOS GOVERNAMENTAIS ABERTOS

Dados governamentais abertos podem ser definidos como informações produzidas ou mantidas pelo governo e disponibilizadas ao público de forma acessível, em formatos abertos e legíveis por máquina, para que possam ser reutilizadas e redistribuídas livremente, observadas eventuais exigências de atribuição de autoria (CRISTÓVAM; HAHN, 2020). No Brasil, a Lei de Acesso à Informação (LAI) reforça que o acesso às informações públicas deve ser garantido como direito fundamental, alinhado aos princípios de transparência e publicidade (BRASIL, 2011).

Segundo Cristóvam e Hahn (2020), os dados governamentais abertos não se limitam a simples informações disponibilizadas, mas representam um elemento estratégico para melhorar a gestão pública e fomentar o controle social, promovendo uma maior interação entre governo e sociedade.

A disponibilização de dados governamentais no Brasil ocorre majoritariamente por meio de plataformas digitais, como o Portal Brasileiro de Dados Abertos e a Infraestrutura Nacional de Dados Abertos (INDA). Essas iniciativas visam centralizar e padronizar a publicação de dados públicos, garantindo que sejam acessíveis e utilizáveis por qualquer cidadão. INDA tem como objetivo promover a integração entre os diversos órgãos públicos, facilitando a divulgação de informações em formatos reutilizáveis (CRISTÓVAM; HAHN, 2020).

Entretanto, mesmo com avanços significativos, persistem desafios técnicos e estruturais. Muitas bases de dados ainda são disponibilizadas em formatos inadequados, como PDFs, ou apresentam problemas de atualização e qualidade, dificultando sua utilização por usuários finais. Além disso, a falta de interoperabilidade entre diferentes sistemas governamentais limita o potencial de integração e análise dessas informações (FORTINI; AMARAL; CAVALCANTI, 2021).

O processo de disponibilização de dados abertos envolve uma série de demandas técnicas, legais e culturais. Entre os principais desafios estão:

1. Padronização de formatos: A adoção de padrões abertos e interoperáveis é fundamental para garantir que os dados possam ser acessados e reutilizados de forma eficiente (CRISTÓVAM; HAHN, 2020);

2. Qualidade dos dados: É necessário assegurar a consistência, completude e atualização dos dados disponibilizados. Dados incompletos ou desatualizados comprometem sua utilidade para análises e decisões baseadas em evidências (OECD, 2019);

3. Governança de dados: A falta de políticas claras para gestão e documentação de dados, incluindo metadados detalhados, limita a transparência e dificulta o controle social (FORTINI; AMARAL; CAVALCANTI, 2021);

4. Conformidade legal: A proteção de dados pessoais, conforme estabelecido pela LGPD, deve ser uma prioridade no processo de disponibilização, exigindo anonimização e categorização adequadas (BRASIL, 2018).

5. Capacitação técnica: A formação de profissionais capacitados para lidar com as diferentes etapas do ciclo de vida dos dados é essencial, especialmente em órgãos públicos de menor porte (CRISTÓVAM; HAHN, 2020);

6. Engajamento social: Para que os dados abertos atinjam seu potencial pleno, é necessário incentivar o uso por parte de cidadãos, pesquisadores e organizações, promovendo uma cultura de transparência e inovação (OECD, 2019).

Em síntese, os dados governamentais abertos representam uma oportunidade significativa para aprimorar a gestão pública e fortalecer a cidadania. No entanto, sua efetiva implementação requer esforços coordenados para superar barreiras técnicas, culturais e legais, promovendo um ecossistema de dados mais acessível, transparente e confiável.

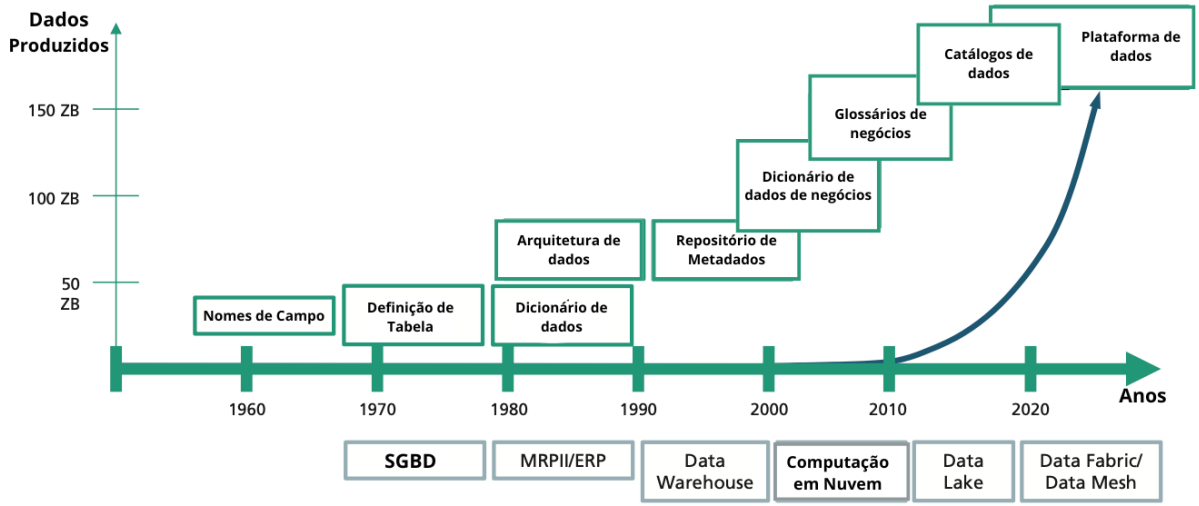
2.2 EVOLUÇÃO DO USO DE DADOS

A utilização de dados na pesquisa, na indústria, nos governos e nas empresas tem se tornado fundamental para auxiliar o desenho de politicas públicas, criação de tecnologias sociais, para negócios e para o desenvolvimento da sociedade. Entretanto, é preciso garantir que esses dados possam ser descobertos, extraídos e organizados. Para que possam ser utilizados e aproveitados, é necessário que os envolvidos consigam acessar e compreender os dados (JAHNKE; SPIEKERMANN; RAMOUZEH, 2022).

A Figura 2.1 apresenta de forma visual um resumo da evolução das estratégias e soluções em relação a produção de dados ao longo do tempo. Os primeiros conceitos de *Field names* e a *Table definition* são a base para o armazenamento de dados e o desenvolvimento dos Sistemas de Gerenciamento de Bancos de Dados (SGBD). Por sua vez, foi necessária a criação de dicionários de dados para documentar informações técnicas de tabelas em aplicações de banco de dados(JAHNKE; SPIEKERMANN; RAMOUZEH, 2022).

O surgimento de *Enterprise Resource Planning* (ERP) e *Material Requirements Planning* (MRP e MRPII), integrando os dados aos processos de negócios, resultou em uma

Figura 2.1 Histórico e Evolução do uso dos dados Jahnke, Spiekermann e Ramouzeh (2022) apud Korte et al. (2019)



Fonte: Adaptado de ISST DataCatalogs Report (JAHNKE; SPIEKERMANN; RAMOUZEH, 2022).

maior complexidade dos sistemas. Foi necessário então planejar e integrar sistemas e dados de acordo com as necessidades empresariais através da *Data Architecture* (JAHNKE; SPIEKERMANN; RAMOUZEH, 2022).

Os *Data Warehouses* surgem como uma solução para a necessidade de integração dos dados produzidos pelos diversos sistemas e para facilitar a tomada de decisões baseada em dados (JAHNKE; SPIEKERMANN; RAMOUZEH, 2022). Como os *Data Warehouses* integram dados de diferentes formatos, estruturas e semânticas, e fornecem os dados consolidados para usuários com expectativas diferentes, os dados precisavam ser documentados em Repositórios de Metadados(JAHNKE; SPIEKERMANN; RAMOUZEH, 2022).

A tecnologia de *Cloud Compuiting* propicia que os dados e informações se tornem mais dispersos, criando a necessidade de dicionários de dados institucionais para definir metadados técnicos para toda instituição. Em um segundo momento, foram desenvolvidos glossários empresariais para organizar e compreender os dados de uma forma mais semântica, com definições claras dentro das empresas(JAHNKE; SPIEKERMANN; RAMOUZEH, 2022).

Com o reconhecimento do valor dos dados e o surgimento das tecnologias de *Big Data*, foram implementadas arquiteturas de *Data Lake* para tratar as grandes quantidades e variedades de dados (JAHNKE; SPIEKERMANN; RAMOUZEH, 2022). Para atender a demanda de extração, surgiram os Catálogos de Dados, com o objetivo de que os dados sejam encontrados, compreendidos e confiáveis, fornecendo curadoria para os metadados.

Novas abordagens como *Data Fabric* e o *Data Mesh* estão sendo apresentadas para atender as arquiteturas de dados descentralizadas e federadas que estão ganhando espaço, principalmente por conta das desvantagens do uso de *Data Lake* monolíticas (JAHNKE; SPIEKERMANN; RAMOUZEH, 2022).

2.3 METADADOS

A definição de metadados segundo a ISO/IEC (2023) é que metadados são "dados que definem e descrevem outros dados". Em uma definição na perspectiva da Ciência da Informação, o conceito de metadados pode ser definido como "uma informação estruturada para as ações de identificação, descoberta, seleção, uso, acesso e gerenciamento" (ARAKAKI; ARAKAKI, 2020). Na literatura, os metadados podem ter diversas interpretações e a depender da interpretação e do contexto em que estão sendo tratados, os próprios metadados podem ser vistos como dados (ARAKAKI; ARAKAKI, 2020).

Arakaki e Arakaki (2020) usam um ambiente de biblioteca para explicar o metadado. Nesse contexto, por exemplo, o bibliotecário define os metadados de um livro. Um usuário da biblioteca faz a busca por um livro a partir de metadados. No entanto, quando o bibliotecário usa os metadados para relatórios ou mesmo pesquisa referentes ao acervo, essas informações estão sendo tratadas como dados. No contexto atual de *BigData*, os metadados são vistos por Arakaki e Arakaki (2020) como fundamentais para possibilitar o gerenciamento de grandes volumes de dados.

2.4 DATA CATALOG

O *Data Catalog*, ou catálogo de dados, é defino por Consortium et al. (2014) como "uma coleção selecionada de metadados sobre conjuntos de dados". Segundo Jahnke, Spiekermann e Ramouzeh (2022), *Data Catalog* é uma plataforma para curadoria de dados que provê uma ligação entre as fontes de dados e os usuários que precisam dos dados. De uma forma geral a plataforma deve oferecer funcionalidades para registrar, recuperar e utilizar os dados. Além da possibilidade de avaliar e analisar dados.

Em outra definição de catálogo, temos que é "... basicamente um banco de dados com metadados que foram inseridos ou extraídos de fontes de dados que fazem parte de uma instituição" (OLESEN-BAGNEUX, 2023).

Um *Data Catalog* deve prover um inventário de dados e recursos para descoberta de dados como componentes-chave. Outros recursos adicionais devem apoiar a governança de dados, avaliação de dados e análise de dados, juntamente com recursos adequados para administração de catálogos e colaboração de dados (JAHNKE; SPIEKERMANN; RAMOUZEH, 2022).

Segundo Olesen-Bagneux (2023), um catálogo de dados é essencialmente um repositório centralizado que permite às organizações armazenarem, gerenciarem e encontrarem seus dados. De uma forma simplificada, podemos enxergar como um sistema de uma biblioteca, onde temos informações sobre os livros que nos permitem encontrar de acordo com o assunto e saber sua localização.

O catálogo de dados armazena informações sobre os dados, como localização da fonte de dados, proprietário dos dados, qualidade dos dados e termos de negócios associados. Podemos dizer então que esse tipo de solução organiza a documentação sobre os dados.

A medida que inserimos ou extraímos os metadados das fontes de dados e os organizamos no catálogo de dados, eles passam a ser vistos como mais um ativo para organização. Um ativo pode ser definido como "uma entidade de dados que existe em sua organização"

(OLESEN-BAGNEUX, 2023), por exemplo, um arquivo, pasta ou tabela armazenado em uma fonte de dados, sistema de arquivos ou banco de dados.

Um conjunto de ativos que logicamente estão interligados pode-se dizer que pertencem ao mesmo domínio, mesmo que tenha origem em uma ou mais fontes de dados (OLESEN-BAGNEUX, 2023). Por exemplo, um domínio com dados sobre educação pode ter tanto fontes de dados ligadas ao ambiente virtual de aprendizagem quanto a fontes de dados ligadas ao sistema financeiro.

Para organizar os ativos é necessário adicionar informações aos metadados que facilitem a descoberta desses dados. Olesen-Bagneux (2023) definem que os ativos podem ter metadados derivados da fonte de dados, metadados adicionados no catálogo de dados ou ambos.

2.4.1 Metadados derivados da fonte de dados

Estes tipos de metadados podem ser vistos como a descrição mínima de seus ativos e são sempre derivados da fonte de dados (OLESEN-BAGNEUX, 2023). Eles são muito importantes, mas podem não contextualizar o suficiente para fazer seu catálogo de dados, pois em geral não fornecem informações ligadas aos domínios.

Os metadados técnicos informam qual a fonte de dados o ativo está armazenado, quem criou, quando foi criado, o formato do arquivo do ativo ou tipo de fonte, etc. Os metadados de negócios também são metadados que descrevem o ativo, por exemplo, nomes de tabelas e colunas, descrições e definições de tipos de dados, etc. (JAHNKE; SPIEKERMANN; RAMOUZEH, 2022).

2.4.2 Metadados adicionados no catálogo de dados

Este tipo de metadado é adicionado de forma independente e não é derivado da fonte de dados. São utilizados para expressar o contexto da organização. Podem ser apenas descrições e também podem estar associados a pessoas e termos de glossário.

As descrições são os metadados de uso primário. São informações de alto nível sobre os ativos, como uma breve explicação sobre seu objetivo ou contexto de onde foi extraído ou como foi inserido (OLESEN-BAGNEUX, 2023).

Os metadados considerados de uso secundário são sugestões do provedor de dados para potenciais consumidores sobre o que o ativo pode ser usado. Alguns possíveis exemplos de metadados secundários são: pessoas ligadas aos dados, como o proprietário do ativo, o administrador do ativo, o responsável por alimentar a fonte de dados (OLESEN-BAGNEUX, 2023).

Ainda como metadados secundários, tem-se os termos de glossário, que são palavras dos glossários pertencentes ao catálogo de dados. Os glossários são listas de palavras que descrevem sua organização ou domínio ao qual o metadado pertence. Os catálogos de dados permitem associar seu ativo aos termos de um ou mais glossários.

Olesen-Bagneux (2023) destaca tipos de glossários segundo o nível de controle para serem usados em um *Data Catalog*:

- Glossário livre: Glossário livre é uma *folksonomia*. *Folksonomias* são glossários,

gerados pelo usuário, que organizam ativos pelo uso de *tags*. Em redes sociais é comum postagens relacionadas a um certo tópico marcado com uma *hashtag*. Não existem regras formais em uma *folksonomia*; todos podem criar *tags* e usá-las como quiserem. Em um catálogo de dados, pode-se usar uma *folksonomia* para descrever ativos de maneiras completamente únicas;

- Glossário de domínio O Glossário de domínio é uma taxonomia. Taxonomias têm uma hierarquia e são criadas por um pequeno grupo de pessoas dentro de um domínio, criando assim uma linguagem mais formal. As taxonomias são específicas do domínio e representam uma descrição formal do domínio em um glossário;

- Glossário global Glossário global é uma ontologia. Diferente da taxonomia, a ontologia se afasta do pensamento hierárquico e se volta para o pensamento em *clusters*. A ontologia é altamente controlada e tem muito potencial para melhorar a pesquisa.

2.5 SISTEMA DE 5 ESTRELAS PARA DADOS ABERTOS

O sistema de 5 estrelas para Dados Abertos proposto por Berners-Lee (2009) é uma escala que classifica os dados com base em sua abertura e interoperabilidade. As estrelas são atribuídas com base na facilidade com que os dados podem ser acessados, interpretados e reutilizados. As estrelas são categorizadas da seguinte forma:

- ($\star$) Dados sob licença aberta, formato qualquer;
- ($\star\star$) Dados estruturados em formato legível por máquina;
- ($\star\star\star$) Formato não proprietário;
- ($\star\star\star\star$) URIs utilizadas para denotar coisas;
- ($\star\star\star\star\star$) Dados interligados a outros conjuntos de dados.

Para Batista (2021), o sistema de 5 estrelas representa uma abordagem valiosa para orientar as instituições que publicam dados, incentivando-as a disponibilizar informações que ofereçam maiores vantagens para os usuários desses dados. Embora isso possa resultar em um esforço adicional para os publicadores, o modelo se destaca como uma estratégia eficaz para promover aprimoramento da qualidade e maximizar dos benefícios dos dados disponibilizados.

2.6 FAIR

Os princípios FAIR – acrônimo para "*Findable*", "*Accessible*", "*Interoperable*"e "*Reusable*-- representam um conjunto de diretrizes reconhecidas internacionalmente (GO-FAIR, 2022). Esses princípios surgiram em janeiro de 2014 na conferência internacional *Jointly designing the data FAIRPORT*, com o objetivo de debater sobre a gestão, utilização e reutilização de dados de pesquisa dentro do contexto da *e-Science*. O intuito principal da conferência foi o de estabelecer as bases para uma infraestrutura global de dados,

alinhada com os avanços tecnológicos e as necessidades científicas contemporâneas (BONETTI, 2023).

Esses princípios ganharam um maior reconhecimento após sua publicação no periódico *Scientific Data* da *Nature* em 2016. A consolidação dos princípios FAIR se consolida no ano seguinte com a Comissão Europeia passando a exigir a adoção de um plano de gestão de dados com base no FAIR para financiar projetos (BONETTI, 2023).

Outro ponto importante dos princípios FAIR é legibilidade dos dados por máquinas, visando facilitar o processamento automático da vasta quantidade de dados disponíveis. A iniciativa *Metadata for Machines* (M4M) é lançada em 2018 por membros da GO FAIR e da *Research Data Alliance* (RDA), com o objetivo de estimular a criação e reutilização de componentes de metadados FAIR e modelos de metadados prontos para máquinas. Esses modelos buscam definir um conjunto de metadados minimamente viáveis para leitura por máquinas, que sejam modulares e possam ser expandidos conforme necessário (GO-FAIR, 2022).

Vale ressaltar que a maturidade da implementação dos princípios FAIR pode variar, o que permite que pesquisadores e instituições adaptem sua implementação conforme suas necessidades e contextos específicos. Isso significa que os dados podem estar em diferentes níveis de *FAIRness* (WILKINSON et al., 2016). Para verificação da aderência dos dados aos princípios foram propostas várias ferramentas. Neste trabalho nos utilizamos da *SATIFYD*, disponível em https://fairaware.dans.knaw.nl/, que a partir de respostas em um *checklist*, consegue verificar o nível de *FAIRness* dos dados.

2.7 LEGISLAÇÃO E CONFORMIDADE

A convergência entre a LAI e a LGPD apresenta desafios e oportunidades para a administração pública contemporânea. De um lado, a LAI estabelece a transparência como preceito geral, determinando que informações de interesse público sejam amplamente acessíveis. Por outro lado, a LGPD assegura a proteção de dados pessoais, impondo limites e responsabilidades no tratamento de informações sensíveis. Essa interação exige que gestores públicos conciliem esses princípios, promovendo tanto a transparência quanto a privacidade (FORTINI; AMARAL; CAVALCANTI, 2021).

No entanto, Fortini, Amaral e Cavalcanti (2021) argumentam que a LGPD não se coloca como antagonista à LAI, mas sim como uma legislação complementar, com o objetivo de proteger os direitos fundamentais de liberdade e privacidade. A legislação estabelece que, ao tratar dados pessoais, a administração pública deve garantir a transparência e prestar contas sobre como essas informações são utilizadas. Essa exigência reforça a necessidade de soluções que atendam as políticas públicas, proteção de dados e transparência ativa (FORTINI; AMARAL; CAVALCANTI, 2021).

A gestão de dados na administração pública envolve diferentes categorias de informações, como dados anonimizados, dados sensíveis e dados pessoais. Cada uma dessas categorias requer tratamentos específicos, conforme descrito na LGPD, que inclui procedimentos de anonimização e pseudoanimização para minimizar riscos de violação da privacidade (FORTINI; AMARAL; CAVALCANTI, 2021).

O tratamento de dados abertos exige que os gestores sejam capazes de identificar quais

informações podem ser disponibilizadas publicamente sem comprometer a privacidade ou a segurança, o que demanda ferramentas para avaliação e categorização de dados (OECD, 2019).

A tecnologia desempenha um papel essencial na superação desses desafios. Soluções como as propostas neste trabalho, que incluem funcionalidades de governança de metadados e tratamento de dados, indispensáveis para garantir a conformidade com as legislações vigentes. Tais ferramentas não apenas promovem a eficiência administrativa, mas também reforçam a confiança dos cidadãos na administração pública, ao garantir que suas informações sejam tratadas de forma ética e segura (FORTINI; AMARAL; CAVALCANTI, 2021).

Em síntese, a relação entre a LAI e a LGPD é complexa, mas também oferece oportunidades para a construção de uma administração pública mais transparente e responsável. A implementação de soluções tecnológicas que conciliem essas legislações é um passo fundamental para que os princípios de transparência e privacidade sejam plenamente atendidos, promovendo um Estado mais eficiente e democrático (CRISTÓVAM; HAHN, 2020).

2.8 TRABALHOS RELACIONADOS

A patente "US 20240028606 A1" denominada "*Data Catalog and Retrieval Systema*" apresenta sistemas e métodos para catalogação e recuperação de dados. Assim como nosso trabalho, foca em aprimorar a gestão de dados, identificar tipos de dados, gerando configurações para o ingresso de dados e validação em ambiente de teste, transforma o formato dos dados, cataloga, extrai partes dos dados, e gera metadados associados. No entanto, o propósito dos metadados são para melhorar a eficiência na resposta a consultas aos dados (CHAPLIN et al., 2024). Por outro lado, a nossa solução prioriza o metadado para o melhor entendimento das bases valorizando os dados como ativos da instituição.

O trabalho de Czajkowski, Kesselman e Schuler (2017) discute a importância de criar e manter descrições de ativos de dados, para tornar os dados localizáveis, acessíveis, interoperáveis e reutilizáveis (FAIR) e apresenta o ERMrest como um serviço web de dados relacionais que permite a modelagem e manipulação de dados via *API RESTful*, servindo como um catálogo dinâmico para a descrição e vinculação de ativos científicos. Como o nosso trabalho, busca possibilitar que usuários não especialistas gerenciem os dados e os metadados.

A Tabela 2.1 apresenta um resumo comparativo entre as soluções com base nos trabalhos e ferramentas correlatas abordadas: A Tabela 2.1 resume, de maneira comparativa, as principais características das soluções discutidas neste trabalho. Nela, são destacadas as funcionalidades relacionadas ao registro de dados, gerenciamento de metadados, disponibilidade de glossário e dicionário de dados, fornecimento de amostras de dados e a classificação conforme a LGPD. Essa comparação permite visualizar, de forma concisa, as capacidades de cada ferramenta abordada na revisão da literatura.

Embora as soluções possuam funcionalidades relevantes, o Sophiadata busca integra as capacidades essenciais para o gerenciamento dos dados, desde o registro inicial até a classificação conforme a LGPD. Essa integração é fundamental para que as instituições

Tabela 2.1 Comparativo das Ferramentas e Trabalhos Correlatos

Ferramenta	**Registro de Dados**	**Gerenciamento de Metadados**	**Glossário e Dicionário de Dados**	**Data Samples**	**Classificação LGPD nativa**
CKAN	SIM*	SIM	NÃO	NÃO	NÃO
DATAHUB	SIM*	SIM	SIM	SIM	NÃO
Patente	SIM*	SIM	SIM	NÃO	NÃO
ERMrest	NÃO	SIM	SIM*	NÃO	NÃO
Sophiadata	SIM	SIM	SIM	SIM	SIM

possam não só gerenciar seus dados de forma eficiente, mas também garantir a conformidade com as exigências legais e promover a transparência na divulgação dos dados governamentais.

Em síntese, o referencial teórico apresentado destaca a evolução das ferramentas de catalogação e gestão de dados, evidenciando tanto os avanços tecnológicos quanto as limitações das abordagens existentes. A integração de múltiplas funcionalidades pode representar um avanço significativo para a administração pública, contribuindo para a melhoria da qualidade, acessibilidade e segurança dos dados. Esses estudos buscam fundamentar a importância do desenvolvimento de soluções para enfrentar os desafios contemporâneos na gestão de dados governamentais.

Capítulo 3

SOPHIADATA

Este capítulo apresenta a solução computacional materializada em software web que auxilia na gestão, no tratamento e na publicação de dados abertos governamentais, denominada Sophiadata. O Sophiadata representa uma solução no campo de gerenciamento de dados, desenvolvida para estar em conformidade com os princípios FAIR (*Findable, Accessible, Interoperable, and Reusable*), oferecendo um conjunto de técnicas para a carga, extração, limpeza, enriquecimento e administração de dados e metadados.

3.1 VISÃO GERAL

A solução é estruturada em três módulos principais:

Módulo de Ingestão de Dados e Metadados: Esta primeira fase do processo envolve a coleta de dados brutos e metadados de diversas fontes. O módulo é projetado para ser flexível e eficiente, capaz de lidar com uma ampla gama de formatos de dados e estruturas de metadados, garantindo que as informações sejam extraídas de maneira íntegra e confiável.

Módulo de Tratamento e Enriquecimento de Metadados: Após a coleta, os dados passam por um processo de refinamento. Este módulo se concentra na limpeza e na organização dos dados, removendo inconsistências e enriquecendo os metadados com informações adicionais para aumentar sua utilidade e precisão. Este passo é fundamental para garantir que os dados sejam não apenas utilizáveis, mas também valiosos para análises futuras.

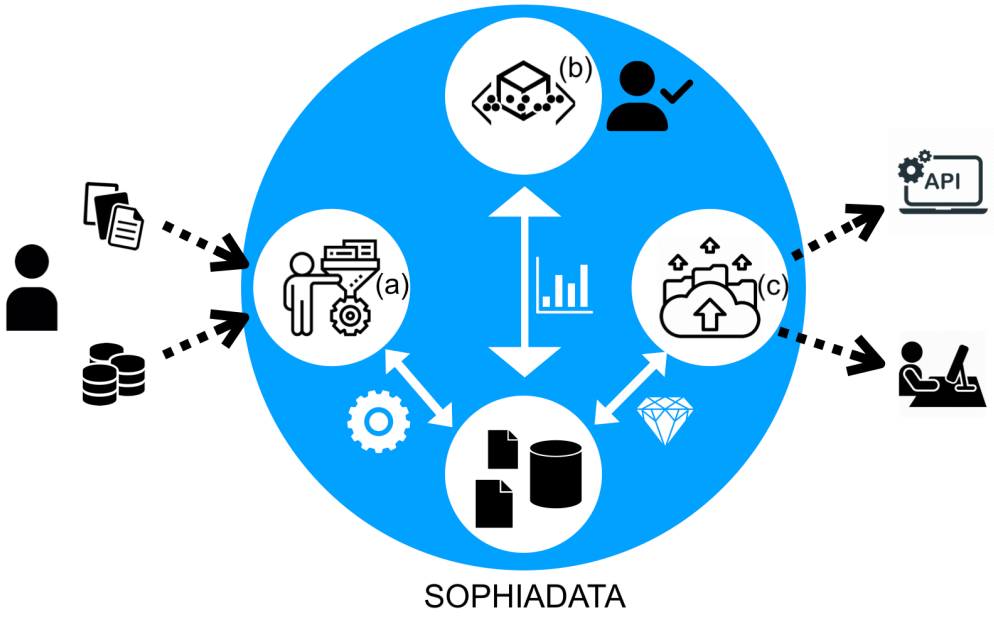
Módulo de Extração e Compartilhamento de Metadados e Dados M2M: O último módulo é dedicado à distribuição e compartilhamento de dados. Este módulo disponibiliza informações estruturadas sobre os metadados em diferentes formatos, atendendo a demanda de pessoas e máquinas. Isso facilita a interoperabilidade entre diferentes sistemas e plataformas com o objetivo de cumprir os princípios FAIR.

Os três módulos trabalham em conjunto para garantir que o Sophiadata não seja apenas uma ferramenta de armazenamento de dados, mas uma solução integrada para a gestão de dados. Solução esta que respeita os princípios de ser localizável, acessível,

interoperável e reutilizável de dados, importantes no contexto atual da ciência de dados e da gestão da informação.

A Figura 3.1 ilustra, de maneira objetiva, uma visão geral do sistema Sophiadata, destacando a disposição e interconexão dos seus módulos, bem como o fluxo principal de comunicação entre eles.

Figura 3.1 Visão Geral do Sophiadata: (a) Módulo de Ingestão de Dados e Metadados (b) Módulo de Tratamento e Enriquecimento de Metadados e (c) Módulo de Extração e Compartilhamento de Metadados e Dados M2M.



Fonte: O Autor, 2024.

A Figura 3.1 representa um processo que se inicia com o usuário importando conjuntos de dados que podem estar em diferentes formatos (como por exemplo, no formato CSV ou de um banco de dados). Estes conjuntos de dados são submetidos a um processo de validação, limpeza e normalização, assegurando a qualidade e a consistência dos dados. Após essa etapa, procede-se à extração dos metadados e estatísticas iniciais, os quais são posteriormente submetidos a uma avaliação e enriquecimento detalhado. Finalmente, o módulo de Extração e Compartilhamento de Metadados e Dados M2M permite a divulgação dos dados e metadados enriquecidos, tanto através de uma *Application Programming Interface* (API) – respeitando o paradigma M2M – quanto por meio de *download* de arquivos, facilitando a integração com outras ferramentas e sistemas. Este processo reflete a abordagem sistemática e integrada do Sophiadata no gerenciamento eficiente de dados e metadados.

3.1.1 Transformação

Um ponto importante da ferramenta é o processamento que a fonte de dados sofre ao se tornar um *dataset*. Detalhamos as transformações necessárias para extração dos metadados e normalização dos dados. No caso de um arquivo CSV, o sistema identifica a primeira linha do arquivo e a utiliza para nomear as colunas do conjunto de dados. Essas ações garantem a identificação única das colunas na estrutura, prevenindo erros na importação.

Para as colunas, as seguintes transformações são aplicadas:

1. Colocar todos os caracteres em maiúsculo;
2. Remover todos os acentos;
3. Remover espaços iniciais e finais;
4. Remover espaços duplos entre as palavas;
5. Substituir os espaços por *underline*;
6. Identificar colunas unicamente.

Para os dados em si, as seguintes transformações são aplicadas:

1. Colocar todos os caracteres em maiúsculo;
2. Remover todos os acentos;
3. Remover espaços iniciais e finais;
4. Remover espaços duplos entre as palavras;
5. Identificar o tipo de dado;
6. Calcular estatísticas descritivas para cada coluna.

Esses procedimentos aprimoram a qualidade dos dados, aplicando um conjunto inicial de regras de transformação.

Com o intuito de exemplificar e detalhar como funciona a transformação inicial do ponto de vista dos dados, apresentamos um conjunto de dados de um arquivo no formato CSV, ilustrado parcialmente na Figura 3.2. Podemos reconhecer alguns dos problemas mais comuns em arquivos de dados como espaço antes, depois e entre os textos, diferença na grafia dos textos e colunas diferentes com o mesmo nome.

Na Figura 3.3, temos o resultado da aplicação das regras para preparação de dados. É possível observar as mudanças na linha de cabeçalho, onde todas as letras ficaram em maiúsculas. Caso existam acentos, eles são removidos, espaços iniciais e finais são removidos, espaços duplos entre as palavas são substituídos por espaços simples, depois os espaços simples são substituídos por *underline*; e por último, adição de um número

Figura 3.2 Exemplo de conteúdo de arquivo CSV.

```
Nome,Idade,Curso  Universidade, Cidade Estado ,data,data
João   Silva  ,25,Engenharia   Civil,  São Paulo,SP,01/05/2010,01/05/2014
 Maria   Oliveira  ,30,  Administração,   São Paulo,SP,01/04/2011,01/04/2015
Carlos Pereira,28,  Medicina, SÃO PAULO  ,sp,01/07/2015,01/07/2019
Ana Souza  ,22,  Psicologia  , Salvador  ,BA,01/04/2015,01/04/2020
Pedro Santos,27,   Design   Gráfico,   salvador   ,Ba,01/02/2016,01/02/2020
```

Fonte: O Autor, 2023.

Figura 3.3 Exemplo de conteúdo de arquivo CSV processado.

```
NOME,IDADE,CURSO_UNIVERSIDADE,CIDADE,ESTADO,DATA,DATA1
JOAO SILVA,25,ENGENHARIA CIVIL,SAO PAULO,SP,01/05/2010,01/05/2014
MARIA OLIVEIRA,30,ADMINISTRACAO,SAO PAULO,SP,01/04/2011,01/04/2015
CARLOS PEREIRA,28,MEDICINA,SAO PAULO,SP,01/10/2011,01/10/2015
ANA SOUZA,22,PSICOLOGIA,SALVADOR,BA,01/04/2015,01/04/2020
PEDRO SANTOS,27,DESIGN GRAFICO,SALVADOR,BA,01/02/2016,01/02/2020
```

Fonte: O Autor, 2023.

às colunas com o mesmo identificador para fazer com que as colunas sejam identificadas unicamente. A estatística descritiva para cada coluna foi calculada e armazenada na ferramenta.

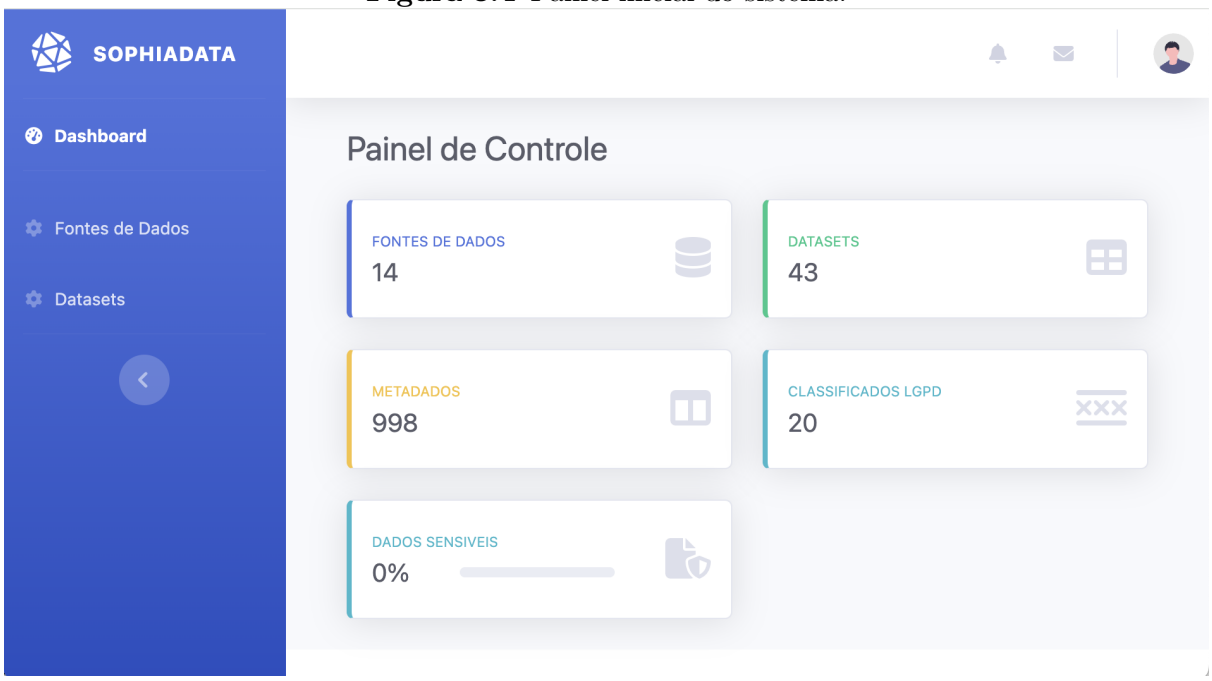
Ainda na Figura 3.3, é possível observar que os demais dados também sofreram mudanças em relação aos originais. Eles foram alterados para maiúscula. Os acentos foram removidos, assim como os espaços iniciais e finais, os espaços duplos entre as palavas são substituídos por espaços simples. Por exemplo, o nome "Maria Oliveira "apresenta espaços antes do texto, depois do texto e espaço duplo entre as palavras. Caso outra linha apresente o "Maria Oliveira" sem os espaços, a solução de análise de dados não identificaria os valores como iguais.

3.2 FUNCIONALIDADES

Ao acessar o sistema Sophiadata, o usuário é imediatamente apresentado a um painel (Figura 3.4) consolidado que resume as principais métricas referentes ao gerenciamento de dados. Este painel central exibe informações como o total de fontes de dados integradas, a quantidade de *datasets* gerados a partir dessas fontes, além da contagem de metadados distintas abrangendo todos os *datasets*. Detalha também a quantidade de registros classificados segundo a LGPD e o percentual de metadados classificados como sensíveis em relação ao total de dados classificados.

Um menu lateral permite o acesso direto e organizado às seções de Fontes de Dados e *Datasets*. É importante destacar que se trata de um sistema web que foi desenvolvido buscando uma interface simplificada e responsiva para ter seu acesso possível em diferentes dispositivos.

Figura 3.4 Painel inicial do sistema.



Fonte: O Autor, 2024.

3.2.1 Ingestão de dados

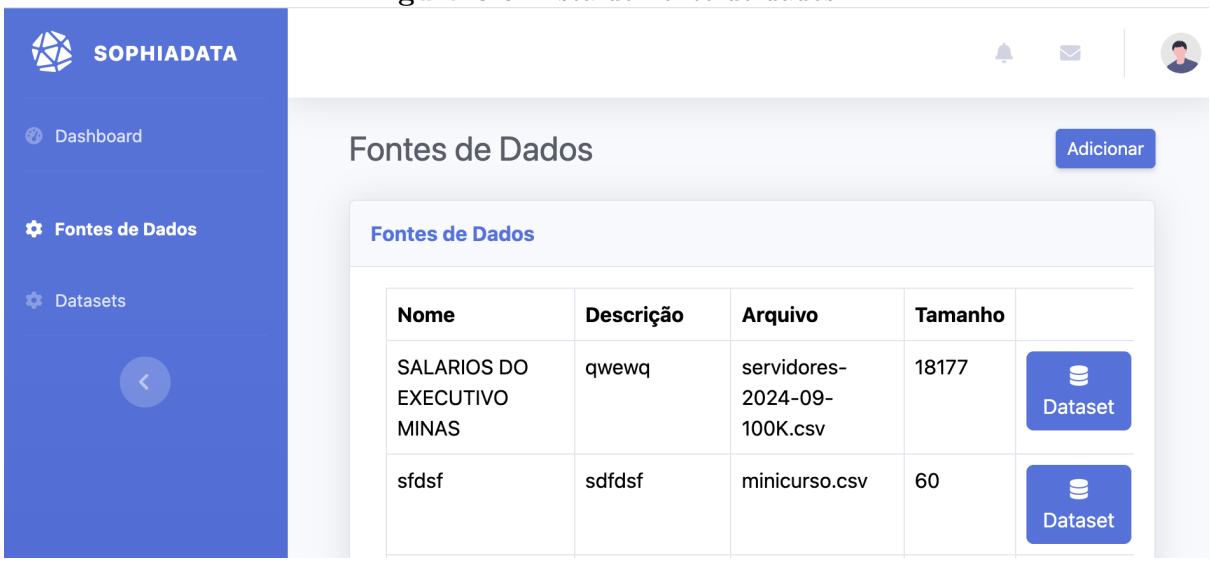
A solução permite o gerenciamento de fontes de dados. Neste local, serão exibidas as fontes de dados já cadastradas (Figura 3.5). A inclusão de novas fontes de dados requer o preenchimento de um formulário (Figura 3.6) com dados descritivos do novo *dataset*, inclusive um atributo de "Visibilidade" da fonte de dados: privada, acesso interno, pública (quando está disponível de forma pública), ou compartilhada (caso tenha sido disponibilizada mediante alguma forma de convênio). Neste formulário ainda é necessário identificar o "Tipo" de fonte: Arquivo do tipo CSV ou uma conexão com um Banco de dados.

No caso de arquivos CSV (Figura 3.7) como tipo de fonte de dados, além da escolha do arquivo correspondente, devem ser informados com ele três campos adicionais ligados ao formato do arquivo CSV: o Delimitador, que define como é a separação entre os campos no arquivo, o Conjunto de Caracteres, que define a codificação utilizada para o arquivo e o campo Cabeçalho que define número de linhas utilizadas para o cabeçalho do arquivo. Ao selecionar o arquivo, esses campos são identificados e preenchidos automaticamente de acordo com o arquivo.

Após a submissão do arquivo, o sistema realiza uma checagem automática, analisando o tipo do arquivo e conferindo a validade dos dados quanto à consistência de linhas e colunas. Em caso de detecção de qualquer erro de leitura, o sistema indica os problemas, para que os mesmos possam ser corrigidos.

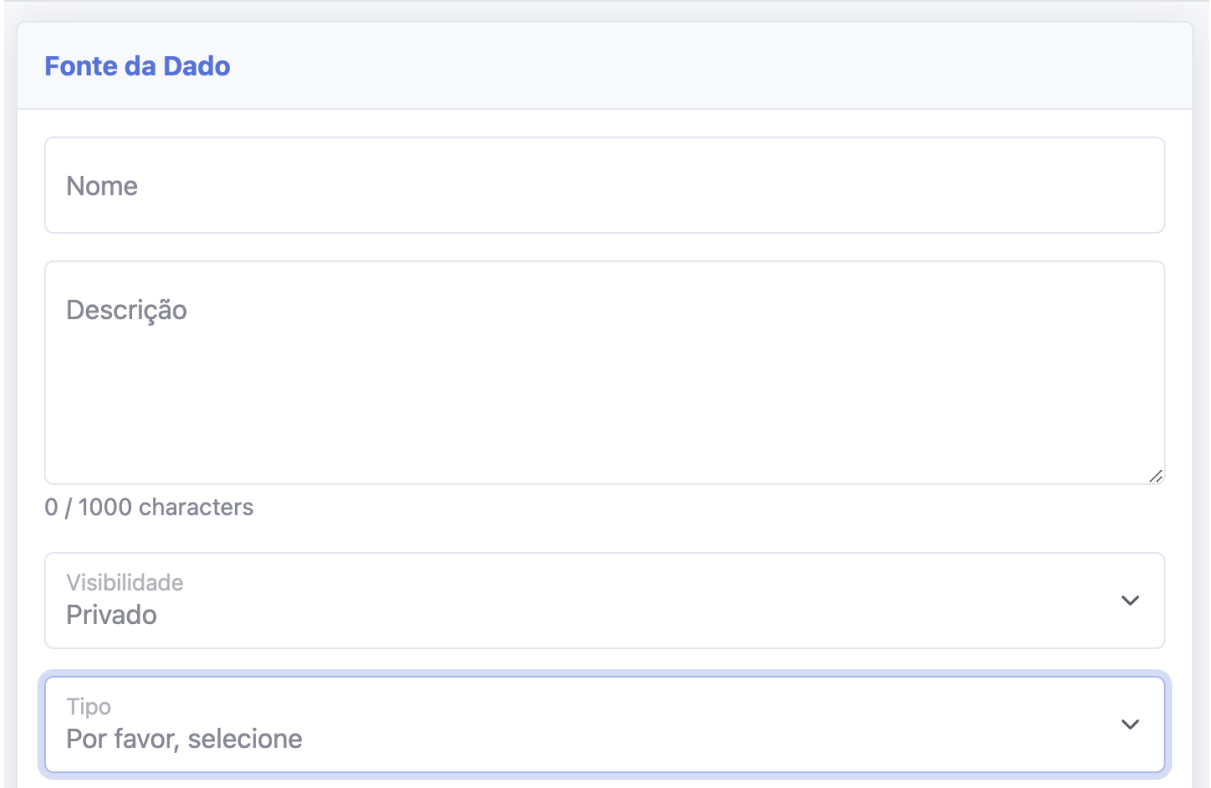
Uma vez realizado o *upload*, o arquivo é incorporado ao sistema como uma fonte de dados, sendo listado com o nome e a descrição fornecidos, os quais podem ser modificados a qualquer momento. O sistema atribui um código *hash* ao arquivo, que serve como

Figura 3.5 Lista de Fonte de dados.



Fonte: O Autor, 2024.

Figura 3.6 Fonte de dados Formulário Inicial.



Fonte: O Autor, 2024.

mecanismo de validação da origem dos dados e facilita o rastreamento do trajeto dos dados desde sua inserção no sistema. Na listagem, são exibidos o nome atribuído a fonte de dados, a descrição correspondente, o nome original do arquivo e o tamanho em *bytes*. A partir disso, o sistema permite a criação do *dataset* para ajustes e compartilhamento.

Figura 3.7 Fonte de dados Formulário CSV.

Fonte: O Autor, 2024.

Caso a opção seja por adicionar uma conexão direta a um banco de dados (Figura 3.8), um conjunto de campos necessários para estabelecer uma conexão são necessários:

1. "URL JDBC", recebe o *Uniform Resource Locator* (URL) do *Java Database Connectivity* (JDBC), que especifica os o caminho de localização para acesso a um banco de dados, seguindo o formato:

   ```
   jdbc:<protocolo>:<subprotocolo>://<host>:<porta>/<banco>?<parametros>
   ```

2. Usuário e a respectiva senha para acesso ao banco de dados.

3. O último campo trata da consulta que deve ser executada em formato *Structured Query Language* (SQL).

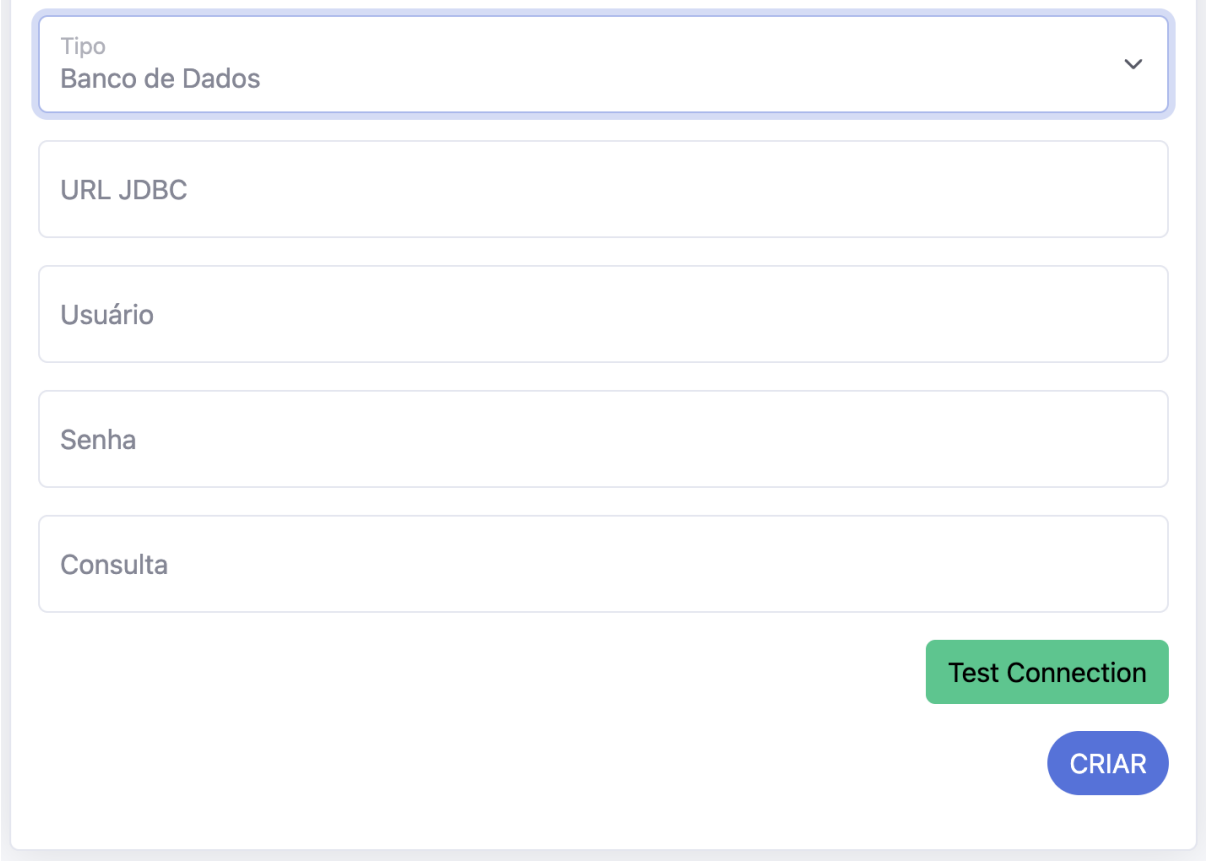
A partir desses dados, a interface permite a execução de testes para verificar a conectividade com a base de dados e se a consulta pode ser executada.

Assim como no caso de Arquivos CSV, ao submeter os dados de conexão, o sistema adiciona as informações da conexão à lista de fonte de dados (Figura 3.5), com o nome e a descrição fornecidos, os quais podem ser modificados a qualquer momento.

3.2.2 Gestão dos Metadados

Para criar um *dataset*, o sistema faz o processamento de uma fonte de dados cadastrada. O Sophiadata inicialmente obtém a fonte de dados com os metadados e dados correspondentes atualizando o *dashboard* e apresentando em uma lista de *datasets* ao usuário

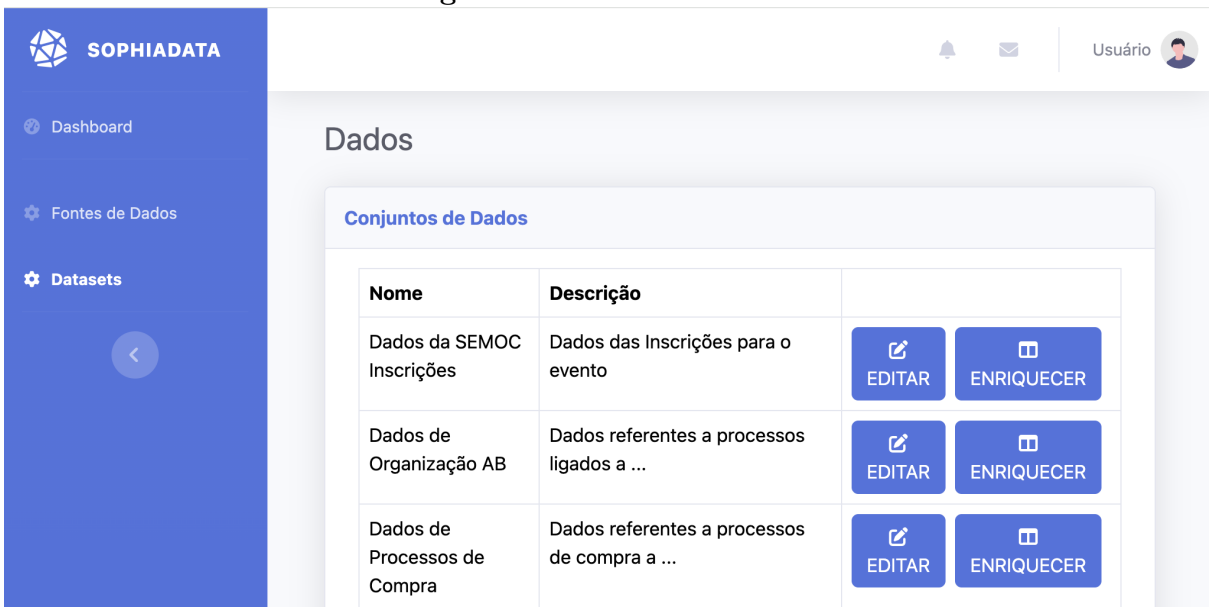
Figura 3.8 Fonte de dados Formulário Banco de Dados.



Fonte: O Autor, 2024.

(Figura 3.9). Essa lista permite modificações no nome e na descrição dos *datasets* para melhor compreensão do seu conteúdo e propósito. Cada *dataset* está associado a um conjunto de metadados.

Figura 3.9 Lista de Datasets.



Fonte: O Autor, 2024.

O sistema permite a criação de múltiplos *datasets* a partir de uma única fonte de dados, para que os termos usados possam variar conforme o contexto em que esse conjunto de dados será utilizado. Essa funcionalidade tem o objetivo de assegurar a flexibilidade no compartilhamento desses dados, possibilitando que diferentes aspectos ou dimensões dos dados sejam explorados e gerenciados de maneira independente.

No sistema, ao acessar a opção de enriquecer (Figura: 3.9) o *dataset*, os usuários tem a possibilidade de visualizar (Figura: 3.10) e editar (Figura: 3.11) cada campo individualmente na aba *Schema*.

Para cada campo à uma coluna equivalente no banco de dados é possível observar (Figura 3.10) "nome" que pode ser alterado e definido pelo usuário, seu identificador no banco de dados e o tipo do dado. Na coluna "descrição", temos a descrição extraída e na coluna *tags*, a possibilidade do uso dessa técnica para rotular o dado. Ao clicar no campo é possível fazer a edição dos seguintes itens (Figura: 3.11): rótulo, descrição, tipo de dado, classificação LGPD e *tags*, para melhorar ainda mais a compreensão do dado.

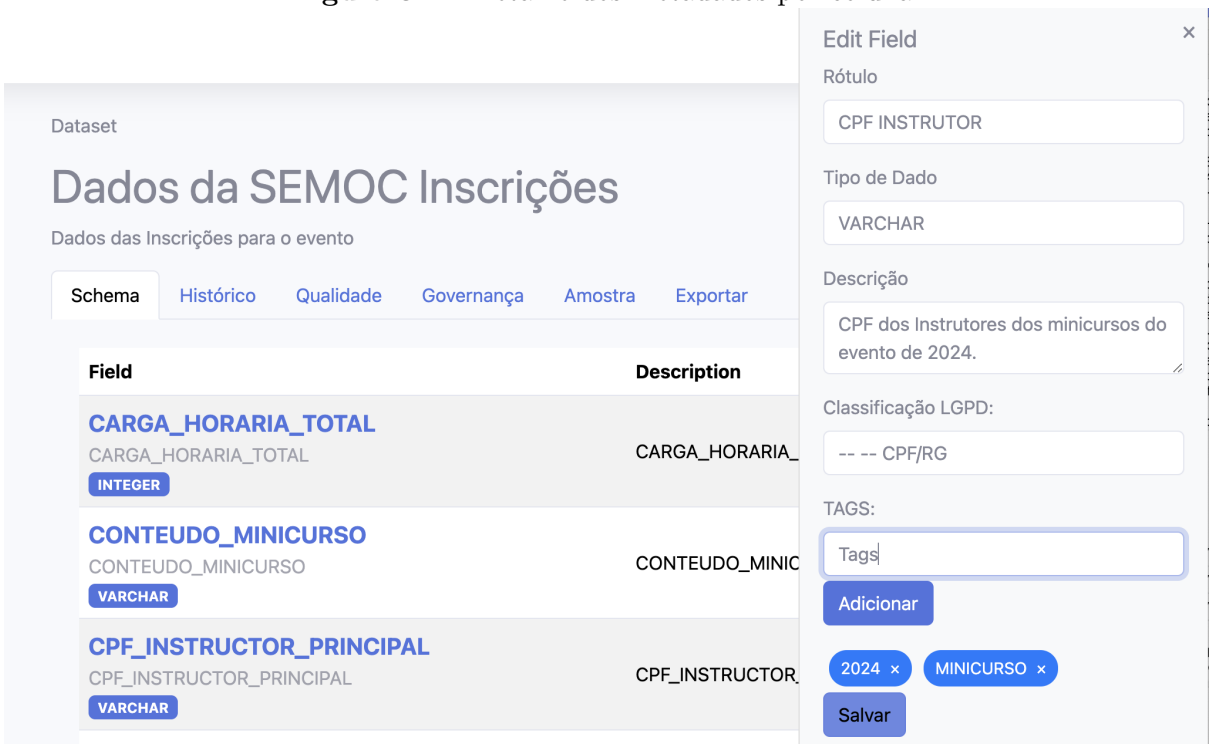
Na Figura 3.12, apresentamos o resultado em que o metadado que inicialmente foi extraído apenas como CPF_INSTRUCTOR_PRINCIPAL foi editado para CPF INSTRUTOR. Este dado foi classificado segundo a LGPD como Dado Pessoal, que identifica a pessoa. A descrição foi alterada para melhor entendimento do dado e foram adicionadas duas *tags* para indicar que tem relação com "2024" e com "MINICURSO".

Na aba Histórico (Figura 3.13), é possível visualizar todas as transformações aplicadas ao conjunto de metadados. Essas informações são organizadas cronologicamente,

Figura 3.10 Detalhe dos metadados por coluna.

Fonte: O Autor, 2024.

Figura 3.11 Detalhe dos metadados por coluna.



Fonte: O Autor, 2024.

Figura 3.12 Metadado editado e enriquecido.

Fonte: O Autor, 2024.

detalhando a data, a hora e o tipo de transformação realizada. O objetivo é facilitar o entendimento geral do conteúdo dos dados e garantir a rastreabilidade e o histórico (*lineage*) do *dataset*. Por exemplo, na Figura 3.13, observa-se:

- Em 09/10/2024 às 20:09, foi realizada a ação CREATE_DATASET, indicando que o conjunto de dados foi criado;
- Em 13/12/2024 às 08:36, ocorreu uma transformação personalizada (CUSTOM_TRANSFORM), na qual o metadado foi alterado, especificando as mudanças realizadas no campo CPF_INSTRUCTOR_PRINCIPAL, que recebeu o rótulo "CPF INSTRUTOR".

A funcionalidade de Histórico dos dados é essencial para garantir a rastreabilidade das alterações feitas no conjunto de dados, permitindo um histórico completo de transformações. Assim, os usuários conseguem acompanhar as mudanças realizadas, compreender o impacto das alterações nos metadados e validar as modificações para assegurar a consistência e a governança dos dados.

Figura 3.13 Histórico de transformações.

Fonte: O Autor, 2024.

Na aba Qualidade (Figura 3.14), são apresentadas estatísticas descritivas que permitem avaliar a integridade e a consistência do conjunto de dados. Essas estatísticas incluem a distribuição dos valores de cada campo, a identificação de dados válidos e a detecção de possíveis inconsistências ou dados ausentes. Esse tipo de análise oferece uma visão geral da qualidade dos dados. Ela permite que inconsistências sejam detectadas de forma antecipada, assegurando a confiabilidade do conjunto de dados. Por exemplo, na Figura 3.14, o campo CARGA_HORARIA_TOTAL exibe uma distribuição em que 95% dos valores corresponde a 4 horas, enquanto 5% corresponde 8 horas. Nesse conjunto de dados existem 19 registros, nenhum ausente ou inconsistente. Além da exibir

de forma textual também é representado visualmente por meio de uma barra horizontal, destacando a proporção de valores válidos e indicando que o campo está completo.

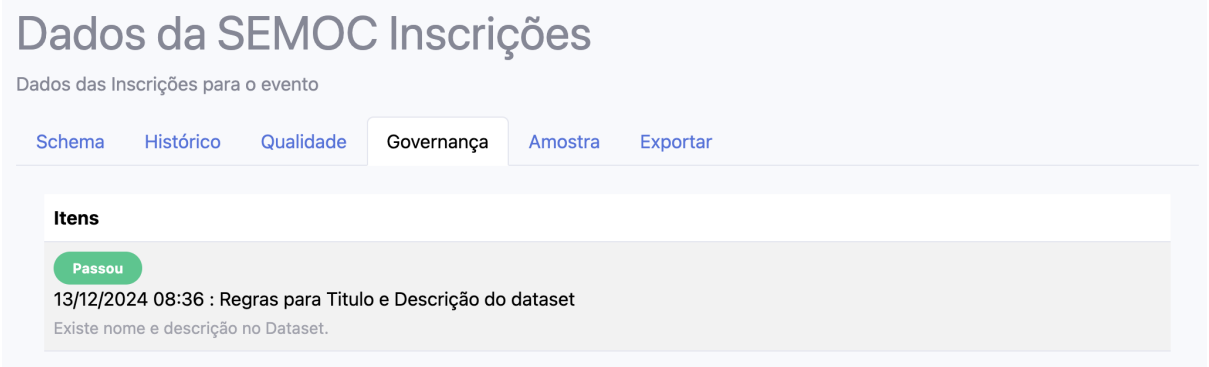
Figura 3.14 Aba Qualidade.

Fonte: O Autor, 2024.

A funcionalidade de Governança (Figura 3.15) apresenta as verificações realizadas para garantir que o conjunto de dados atende às regras estabelecidas. Como exemplo, destaca-se a aplicação da regra que verifica se as informações de título e descrição do *dataset* estão presentes e em conformidade com os critérios previamente definidos, assegurando a consistência e a padronização dos dados.

Devido ao uso de padrões de projeto bem definidos no sistema, adicionar novas regras de governança é um processo simples para desenvolvedores com conhecimento na linguagem Java. Essa modularidade e facilidade de extensão tornam o sistema altamente adaptável às necessidades específicas de governança, permitindo a implementação de validações adicionais com pouco esforço.

Figura 3.15 Aba Governança.

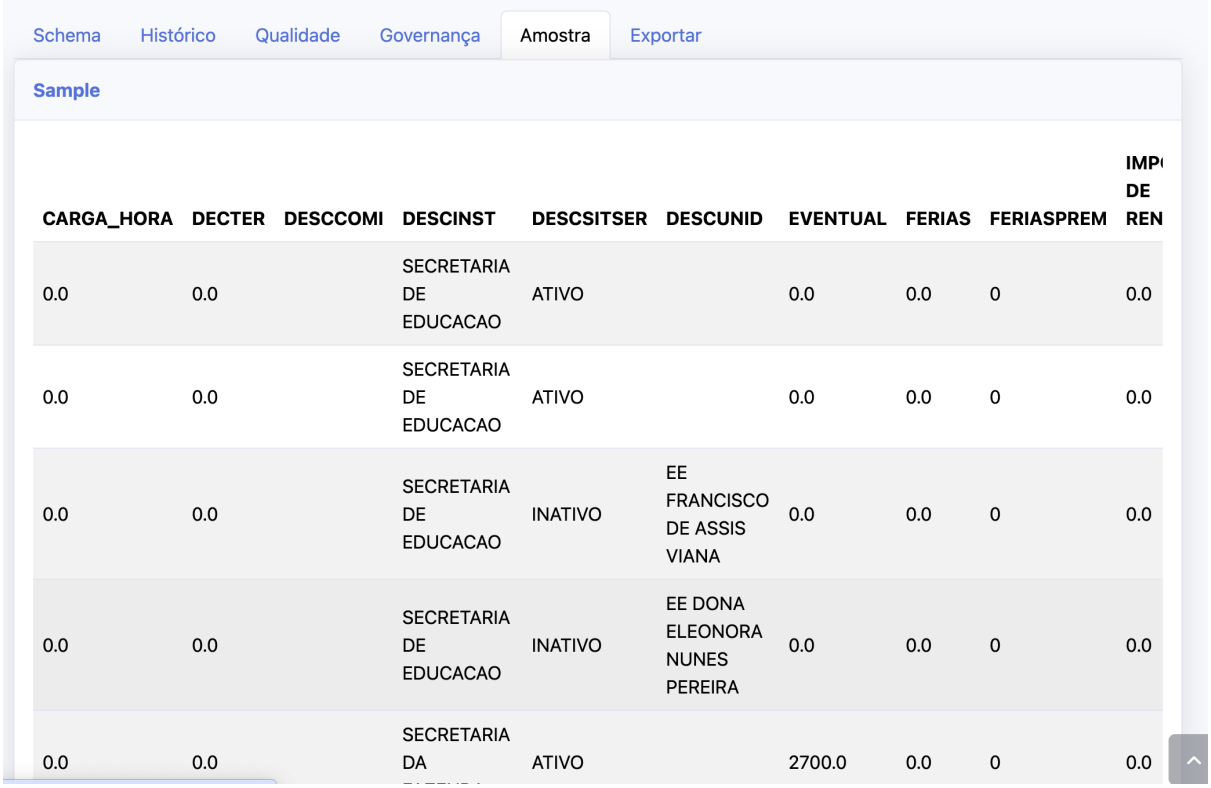


Fonte: O Autor, 2024.

Na aba Amostra (Figura 3.16), é exibido um pequeno conjunto de dados que permite ao usuário visualizar de forma prática o conteúdo do *dataset*. Essa funcionalidade tem

como objetivo fornecer uma visão inicial dos registros armazenados, facilitando a compreensão e a inspeção rápida das informações antes de realizar análises mais detalhadas ou operações sobre o conjunto completo de dados.

Figura 3.16 Aba Exportar.



Fonte: O Autor, 2024.

A funcionalidade de Exportação (Figura 3.17) apresenta uma interface que disponibiliza diversas opções para exportar os metadados e os dados associados em formatos amplamente utilizados, como CSV, PDF, JSON e *Excel Open XML Spreadsheet* (XLSX). Essas opções permitem aos usuários transferir as informações com facilidade para outros sistemas ou ferramentas, promovendo a portabilidade e integração em diferentes contextos de uso.

A possibilidade de escolha entre os formatos facilita a adaptação às necessidades específicas de cada usuário ou projeto. Por exemplo, o formato PDF pode ser utilizado para documentação ou relatórios oficiais, enquanto os formatos CSV e XLSX são ideais para análises em planilhas ou ferramentas de visualização de dados. O formato JSON, por sua vez, é amplamente utilizado em integrações e troca de dados entre sistemas, especialmente em aplicações web. Essa flexibilidade torna a funcionalidade de exportação essencial para garantir a usabilidade e acessibilidade das informações em múltiplos cenários.

Figura 3.17 Aba Exportar.

Fonte: O Autor, 2024.

3.2.3 Exportação e compartilhamento

Dentro da opção de *datasets* do menu, o usuário tem a possibilidade de acessar os dados completos ou uma amostra dos dados. Neste ponto, há a opção de exportar tanto os dados quanto os metadados no formato CSV (Figura 3.18), facilitando a manipulação e análise externa das informações.

Figura 3.18 Exemplo Metadados da Coluna no formato CSV.

```
NOME,DESCRICAO,TIPO
CIDADE,CIDADE,VARCHAR
CURSO,CURSO,VARCHAR
DATA,DATA,VARCHAR
DATA_1,DATA_1,VARCHAR
ESTADO,ESTADO,VARCHAR
IDADE,IDADE,INTEGER
NOME,NOME,VARCHAR
```

Fonte: O Autor, 2023.

O sistema oferece mecanismos para obter os metadados, disponibilizando-os para *download* ou via API em formato JSON (Figura 3.19), assim como o acesso direto aos dados. Essa funcionalidade permite a integração e o acesso M4M, assegurando que os dados possam ser consultados e utilizados de maneira distribuída, ampliando as capacidades de interconexão e colaboração entre sistemas distintos.

Uma versão do arquivo completo de metadados em um formato PDF também está disponível no sistema (Figura 3.20).

Figura 3.19 Trecho do conteúdo de arquivo metadados no formato JSON.

Fonte: O Autor, 2023.

Figura 3.20 Arquivo metadados no formato PDF.

DICIONÁRIO DE DADOS

DATA: 08/02/2024

ATUALIZAÇÃO: 08/02/2024

NOME

DESCRICAO

Descrição dos Campos

CAMPO	NOME	TIPO	DESCRICAO
CIDADE	CIDADE	VARCHAR	CIDADE
CURSO	CURSO	VARCHAR	CURSO
DATA	DATA	VARCHAR	DATA
DATA_1	DATA_1	VARCHAR	DATA_1
ESTADO	ESTADO	VARCHAR	ESTADO
IDADE	IDADE	INTEGER	IDADE
NOME	NOME	VARCHAR	NOME

Fonte: O Autor, 2023.

Capítulo 4

EXPERIMENTO I

Este capítulo descreve a experimento exploratório realizado para avaliar a capacidade do Sophiadata de aprimorar conjuntos de dados e respectivos metadados para torná-los mais acessíveis. O experimento visa replicar cenários de uso típicos que instituições enfrentariam ao organizar e compartilhar dados, fornecendo uma análise do impacto prático do *software*.

4.1 METODOLOGIA DO EXPERIMENTO

4.1.1 Configuração do experimento

O experimento foi realizado utilizando um *notebook* com processador M1 e 16 GB de memória RAM. Esta escolha de *hardware* visa assegurar um desempenho robusto e eficiente para a execução do *software* Sophiadata, bem como para o gerenciamento e processamento dos conjuntos de dados.

O *software* Sophiadata foi configurado para operar localmente no *notebook*. Para isso, utilizou-se o servidor *web* Apache Tomcat, que é uma implementação de *software* livre de um servidor *web* Java que fornece uma plataforma para a execução de aplicações Java, incluindo aquelas que utilizam tecnologias *JavaServer Pages* (JSP) e *servlets*. O Apache Tomcat foi escolhido por sua confiabilidade, eficiência e compatibilidade ampla com diversas aplicações Java.

Juntamente com o servidor Tomcat, um banco de dados PostgreSQL foi utilizado. O PostgreSQL é um sistema de gerenciamento de banco de dados relacional, conhecido por sua robustez e conformidade com padrões. Para garantir flexibilidade e isolamento do ambiente, o PostgreSQL foi hospedado em um *container* Docker. O uso de *containers* Docker é uma prática comum em experimentos e desenvolvimento de *software*, pois oferece um ambiente controlado e replicável, o que é essencial para a consistência dos experimentos.

Para o propósito deste experimento, os conjuntos de dados foram escolhidos a partir do Portal Brasileiro de Dados Abertos (Controladoria-Geral da União, 2024), que é uma referência nacional para a disponibilização de conjuntos de dados abertos. A seleção

dos arquivos CSV deste portal foi baseada em critérios como relevância, atualidade e diversidade, de modo a representar um espectro amplo de cenários reais que instituições enfrentariam ao organizar e compartilhar dados. Esta escolha também visa simular a aplicabilidade do Sophiadata em um contexto realístico, onde o *software* pode ser utilizado para melhorar a qualidade e a acessibilidade de dados públicos.

4.1.2 Procedimentos do experimento

O experimento inicia com o *download* e a preparação dos arquivos selecionados para serem submetidos ao *software* Sophiadata e o respectivo processamento. A primeira etapa dentro do *software* envolve a carga desses arquivos como uma nova fonte de dados, criação do metadados e do *dataset*.

Após a carga dos metadados e geração dos metadados iniciais há um processo de edição e aprimoramento, utilizando informações disponíveis no repositório do site de origem dos dados. Essa ação foi direcionada para maximizar a riqueza e a utilidade dos metadados, ampliando assim sua conformidade com o Modelo de 5 Estrelas e os princípios FAIR.

Após a aplicação das funcionalidades do Sophiadata, executamos o *download* dos arquivos CSV que havia sido processados pelo *software*. Este arquivo representa os dados pós-processamento. Além disso, também foi realizado o *download* dos metadados enriquecidos. Ambos, o conjunto de dados e os metadados foram preparados para uma análise comparativa detalhada.

4.1.2.1 Métricas de Avaliação Na avaliação do conjunto de dados, foram aplicadas duas métricas principais para garantir a qualidade e a acessibilidade dos dados: o sistema de 5 Estrelas de Dados Abertos de Tim Berners-Lee e os princípios FAIR.

Os princípios FAIR buscam tornar os dados mais localizáveis, acessíveis, interoperáveis e reutilizáveis, enfatizando a importância de metadados e formatos abertos. Paralelamente, o sistema de 5 Estrelas incentiva a publicação de dados em formatos que facilitam sua utilização e integração na web.

A aplicação conjunta destas métricas fornece uma avaliação abrangente da qualidade dos dados, contemplando desde a sua forma de publicação até a sua efetiva utilização em diferentes contextos. Utilizar estas métricas em conjunto assegura que o conjunto de dados não apenas atende a critérios técnicos de estruturação e interoperabilidade, mas também é gerenciado e compartilhado de maneira ética e eficiente.

4.1.3 Execução do experimento

4.1.3.1 Preparação dos Dados Os conjuntos de dados para o experimento foram selecionados a partir do Portal de Dados Abertos do governo que disponibiliza dados em um formato acessível e utilizável para o público. Em 20 de dezembro de 2023, data em que a consulta foi realizada, o portal continha um total de 12.591 conjuntos de dados, que, por sua vez, abarcavam 132.494 recursos distintos.

Na página inicial do portal, identificamos uma seção destacando seis conjuntos de dados classificados como “Em alta”, o que indica uma maior relevância ou procura por parte

dos usuários. Dentre estes, selecionamos o conjunto intitulado "Gestão de Pessoas (Executivo Federal) - Cargos Vagos e Vacâncias", que havia sido atualizado em 07 de dezembro de 2023 e disponível em https://dados.gov.br/dados/conjuntos-dados/gestao-de-pessoas-executivo-federal—cargos-vagos-e-vacancias . A escolha desse conjunto foi estratégica, visando não apenas a sua atualidade e relevância, mas também sua capacidade de representar os desafios comuns enfrentados pelas instituições na gestão e compartilhamento de dados.

Os dados oferece uma visão abrangente sobre os cargos públicos no âmbito do Poder Executivo Federal Civil do Brasil. Ele detalha informações pertinentes sobre a distribuição de cargos, incluindo a quantidade de cargos aprovados, distribuídos, ocupados e vagos. Ele contempla dados da Administração Direta, Autarquias e Fundações, com a exceção do Banco Central do Brasil e às Carreiras de Inteligência da ABIN.

O conjunto de dados escolhido foi o que compreende os dados de novembro de 2023, disponibilizado no formato *OpenDocument Spreadsheet* (ODS). Este formato foi adotado pelo produtor dos dados a partir de agosto de 2021. O conjunto de dados começou a incluir informações adicionais sobre vacâncias conforme estabelecido no artigo 33 da Lei 8.112/90, assim como identificações de cargos em processo de extinção.

Importante ressaltar a disponibilidade de um dicionário de dados. No entanto o formato PDF em que o arquivo se apresenta é um desafio, especialmente em casos de textos e tabelas não facilmente acessíveis para seleção ou que se assemelham a imagens, não cumprindo integralmente com o critério de dados em formato estruturado e aberto. Para o experimento, o Dicionário de Dados em PDF foi utilizado para o enriquecimento dos metadados e também foi avaliado.

Os dados foram convertidos para o formato CSV. O arquivo ODS foi acessado utilizando o *software Libre Office*, onde se observou a presença de duas planilhas. Estas foram exportadas individualmente para o formato CSV padrão (UTF-8, virgula (,) como separador de campo e aspas duplas (" ") como limitador de texto), formato mais compatível e recomendado para dados abertos. Este processo de conversão foi necessário para que assegurar a execução do experimento.

4.1.3.2 Operacionalização No início do experimento, foi feita a importação dos arquivos CSV (foram convertidos do formato ODS) para o Sophiadata. Esta importação envolveu a criação de fontes de dados no *software*, cada uma correspondendo a um arquivo CSV específico. As informações de título e descrição para estas fontes foram baseadas nos dados encontrados no arquivo de metadados *Dicionario_CargosVagosVacancias.pdf*. As fontes de dados foram nomeadas da seguinte forma:

- **Cargos Vagos e Vacâncias por Órgão**: Inclui informações sobre cargos públicos no Poder Executivo Federal Civil, com detalhes sobre a quantidade de cargos aprovados, distribuídos, ocupados e vagos;

- **Cargos Vagos e Vacâncias por Órgão e Cargo**: Fornece uma visão detalhada, discriminando os cargos por órgãos e tipos de cargo.

A descrição para a fonte de dados foi adaptada a partir das informações encontradas no site (http://dados.gov.br).

Durante o processo de carga dos dados, o Sophiadata realizou a identificação automática dos tipos de dados presentes e gerou uma análise estatística descritiva para cada arquivo. Essa etapa também envolveu o uso dos metadados para enriquecer as informações disponíveis no Sophiadata, aumentando a utilidade e a compreensão sobre os dados. A solução direciona e facilita a criação de um metadado mais rico e M4M.

Os nomes das colunas e os dados passaram por um processo de transformação que englobou várias etapas para assegurar a consistência e a padronização. Inicialmente, todos os caracteres foram convertidos para maiúsculo, estabelecendo um padrão uniforme em toda a base de dados. Esta etapa assegurou que a variação entre caracteres maiúsculos e minúsculos não afetasse a identificação e comparação dos dados.

Em seguida, foi feita a remoção de todos os acentos, uma medida que simplifica o processamento dos dados por sistemas que podem não suportar caracteres especiais. Adicionalmente, espaços iniciais e finais foram eliminados, bem como espaços duplos entre as palavras, visando a manutenção da uniformidade e precisão tanto nos nomes das colunas quanto nos valores dos dados.

Uma alteração significativa foi a substituição de espaços por *underlines*, facilitando assim a referência aos campos de dados em ambientes de programação e banco de dados que não aceitam espaços em nomes de variáveis. Esta prática evita erros de sintaxe em muitas linguagens de programação.

Além disso, cada coluna foi identificada de maneira única, evitando ambiguidades e sobreposições que poderiam comprometer a integridade das análises. Paralelamente, o tipo de dado de cada coluna foi identificado, e estatísticas descritivas foram calculadas, fornecendo uma visão abrangente do conjunto de dados. Estas últimas etapas são essenciais para entender as características gerais dos dados, facilitando análises futuras e decisões baseadas em evidências.

Após a conclusão da carga e do enriquecimento dos dados, os conjuntos de dados e metadados foram baixados nos formatos oferecidos pelo Sophiadata. Para os dados, os formatos disponíveis incluíam CSV, XLSX e JSON; para os metadados, as opções eram CSV, PDF e JSON. Este passo demonstrou a variedade de formatos de arquivos, tornando possível atender às diferentes necessidades dos usuários.

4.1.3.3 Seleção dos Dados Inicialmente, focamos na seleção e no *download* dos dados processados. Isso incluiu os conjuntos de dados e metadados que passaram pelo tratamento do Sophiadata. Os arquivos CSV processados, juntamente com os metadados enriquecidos, foram baixados e verificados para garantir sua integridade. A precisão na coleta desses dados é crucial para assegurar a confiabilidade dos resultados da análise. Posteriormente, focamos na organização dos dados coletados. Eles foram categorizados em dois grupos principais: os dados originais, que representam o estado dos dados antes do processamento pelo Sophiadata, e os dados processados, refletindo o estado após o processamento. Tal organização facilita a comparação direta e eficiente entre os conjuntos de dados.

4.1.4 Análise dos resultados da experimento

Para análise dos resultados, podemos comparar os arquivos originais, CargosVagosVacancias_202311.ods, que contém os dados, e o arquivo Dicionario_CargosVagosVacancias.pdf, que contém os metadados. Segundo o Modelo de 5 Estrelas de Dados Abertos, o arquivo de dados seria classificado como duas estrelas, pois atende ao requisito de estar disponível na Internet sob uma licença aberta e por conta dos dados serem legíveis por máquina. No entanto, o arquivo que apresenta o modelo está no formato PDF, que já não é considerado legível por máquina. O Sophiadata, por sua vez, permite que os arquivos de dados e metadados sejam gerados em diferentes formatos, como CSV e JSON, legíveis por máquina. Isso atende o nível de três estrelas do Modelo de 5 Estrelas de Dados Abertos (Tabela 4.1).

Tabela 4.1 Tabela de Avaliação segundo Modelo de 5 Estrelas de Dados Abertos

Recurso	**Estrelas**
CargosVagosVacancias_202311.ods	⋆ ⋆
My-CargosVagosVacanciaPorOrgão.csv	⋆ ⋆ ⋆
My-CargosVagosVacanciaPorOrgãoCargo.csv	⋆ ⋆ ⋆

Fonte: Autor

A segunda abordagem da avaliação foi executada utilizando o conjunto de perguntas da ferramenta *SATIFYD* (https://fairaware.dans.knaw.nl/), que foi adaptado, traduzido e aplicado nos arquivos originais e nos arquivos tratados no Sophiadata. O objetivo foi identificar e comparar o quanto *FAIRness* estavam os arquivos. As perguntas aplicadas foram:

F1 Você forneceu metadados suficientes (informações) sobre seus dados para que outros possam encontrar, entender e reutilizar seus dados?

- (a) Metadados ricos com o máximo de informações possíveis
- (b) Campos de metadados necessários e alguns campos adicionais
- (c) Metadados mínimos, mas apenas campos obrigatórios (título, criador, data de criação, descrição, público)
- (d) Sem metadados

F2 Você usou padrões como vocabulários controlados, taxonomias (tesauros) ou ontologias para descrever seu conjunto de dados?

- (a) Vocabulários Controlados
- (b) Taxonomias (Tesauros)
- (c) Ontologias
- (d) Não há padrões para o domínio.

F3 Você forneceu documentação adicional rica e detalhada?

(a) Arquivo *Readme*

(b) Versionamento

(c) Procedência

A4 Os metadados são publicamente acessíveis, mesmo que os dados não estejam mais disponíveis?

(a) Sim

(b) Não

A5 O seu conjunto de dados contém dados pessoais?

(a) Sim

(b) Não

A6 Qual das licenças de uso você escolheu para cumprir com os direitos de acesso anexados aos dados?

(a) Acesso aberto (CC0)

(b) Acesso restrito

(c) Embargo

(d) Outro acesso

(e) Não sei qual dessas licenças se aplica aos meus dados

I7 Os dados no seu conjunto de dados estão armazenados em formatos preferenciais (CSV, JSON ou *eXtensible Markup Language* (XML))?

(a) Todos os dados estão em formatos preferenciais

(b) Alguns dos dados estão em formatos preferenciais

(c) Nenhum dos dados está em formatos preferenciais

(d) Não sei quais são os formatos preferenciais

I8 Você faz link com outros (meta)dados e este (meta)dado é resolvível online?

(a) Sim, e estes estão acessíveis online

(b) Sim, mas não estão acessíveis online

(c) Não

I9 Você forneceu informações contextuais sobre seu conjunto de dados?

(a) Identificadores Persistentes

(b) Referência a outros conjuntos de dados

(c) Referência a publicações

(d) Sem metadados contextuais

R10 Que tipo de informação você forneceu sobre a procedência dos seus dados?

(a) Origem dos dados

(b) Citações para dados reutilizados

(c) Descrição do fluxo de trabalho para coleta de dados (legível por máquina)

(d) Histórico de processamento dos dados

(e) Histórico de versões dos dados

R11 Qual licença de uso fornecida você escolheu para o seu conjunto de dados?

(a) Acesso aberto (CC0)

(b) Acesso restrito

(c) Embargo

(d) Outro acesso

(e) Não sei qual dessas licenças se aplica aos meus dados

R12 O seu (meta)dado atende aos padrões do domínio?

(a) Padrões de domínio nos metadados

(b) Padrões de domínio mínimos e padrões genéricos mínimos

(c) Padrões genéricos

(d) Nenhum padrão utilizado

A avaliação foi tabulada na Tabela 4.2 e o resultado do índice de *FAIRness* saiu de 16% nos arquivos originais para 52% nos arquivos obtidos pelo Sophiadata. Isso indica que a ferramenta proporcionou um incremento na qualidade da publicação dos dados com baixo esforço por parte do publicador, oferecendo a ele um suporte para criação do *dataset* e os respectivos metadados em formatos abertos, o que já o coloca com 3 estrelas na escala de maturidade dos dados. Do ponto de vista do consumidor dos dados, o mesmo passou a ter acesso aos dados e metadados em formato legível e interoperável, acessíveis por *download* e através de API. A melhoria do índice de *fairness* esta diretamente ligada ao uso de formato aberto para os dados e metadados.

4.1.5 Limitações do experimento

Aqui está a reescrita do trecho com as melhorias sugeridas e a inclusão da limitação relacionada à exportação dos dados no nível de 3 estrelas:

—

Tabela 4.2 Tabela de Avaliação de *FAIRness* baseada no *SATIFYD*

Item	Conjunto Original	Conjunto Sophiadata
F1	(b)	(b)
F2	-	(a)(b)(c)
F3	-	(a), (b) e (c)
A4	(b)	(a)
A5	(b)	(b)
A6	(d)	(b)
A6	(a)	(a)
I7	(c)	(a)
I8	(c)	(a)
I9	-	(a) e (b)
R10	-	-
R11	(d)	(d)
R12	(d)	(c)
RESULTADO	16%	52%

Fonte: Autor

4.1.6 Limitações do experimento

O experimento buscou replicar cenários típicos enfrentados por instituições ao organizar e compartilhar dados, com o objetivo de validar o comportamento do Sophiadata em diferentes tipos e tamanhos de conjuntos de dados. Embora essa abordagem seja útil para compreender o impacto prático da ferramenta, ela apresenta algumas limitações.

Primeiramente, o experimento foi conduzido em um ambiente controlado, onde as condições e premissas foram simplificadas para facilitar a análise dos resultados. No entanto, essa abordagem pode não capturar completamente a complexidade e os desafios encontrados em ambientes operacionais reais, como volumes elevados de dados, variabilidade nos formatos e a necessidade de integração com sistemas legados.

Além disso, a ferramenta atualmente permite a exportação dos dados apenas até o nível de três estrelas no modelo de 5 estrelas para Dados Abertos, uma vez que os formatos disponibilizados (CSV, JSON, XLSX e PDF) garantem legibilidade por máquina, mas não incluem mecanismos de interligação entre conjuntos de dados por meio de URIs. Esse fator pode limitar sua aplicabilidade em cenários que exigem maior interoperabilidade e conexão entre fontes de dados distribuídas.

Essas limitações são inerentes a muitos experimentos controlados, onde as condições padronizadas podem não refletir inteiramente o dinamismo e a imprevisibilidade do mundo real. Como trabalhos futuros, sugere-se expandir a validação da ferramenta em ambientes institucionais diversos, testar sua aplicabilidade em cenários mais complexos e

explorar melhorias para a exportação de dados em níveis mais altos de interoperabilidade.

Capítulo 5

EXPERIMENTO II

Este capítulo descreve um estudo experimental exploratório realizado com 63 participantes, seguindo o processo *Goal-Question-Metric* (GQM) de Basili e Rombach (1994), para avaliar o Sophiadata. O **objetivo** foi analisar a ferramenta com a **finalidade** e verificar se a solução é **eficaz** em extrair metadados, enriquecer e documentar dados, sob a ótica dos pesquisadores da solução, no **contexto** da publicação de dados abertos governamentais. O Sophiadata deve aprimorar conjuntos de dados e seus metadados, facilitando seu uso e compartilhamento por instituições em cenários típicos de organização e documentação de dados.

Também adotamos para o experimento o *Technology Acceptance Model* (TAM), desenvolvido por Fred Davis em 1989, que busca entender a aceitação da tecnologia pelos usuários, considerando suas percepções de utilidade e facilidade de uso. A avaliação com base no TAM permite observar a interação dos usuários com o Sophiadata, além de avaliar sua intenção de uso e aceitação.

Neste experimento, a validação foi realizada observando a percepção de participantes do processo de execução das tarefas e preenchimento de um questionário ao final. O estudo experimental busca demonstrar como o *software* facilita a documentação e o compartilhamento de dados em um ambiente operacional, permitindo avaliar a capacidade de enriquecimento de metadados, classificação conforme a LGPD e geração de documentação de dados para apoio na governança.

5.1 METODOLOGIA DA ESTUDO EXPERIMENTAL

5.1.1 Ambiente de estudo experimental

A experimento foi realizada utilizando computadores com diferentes configurações rodando o sistema operacional *windowns* em sua maioria.

O *software* Sophiadata foi configurado para operar localmente. Para isso, a ferramenta possui um servidor *web* Apache Tomcat embarcado. Juntamente com o servidor Tomcat, um banco de dados H2 também embarcado foi utilizado para para garantir flexibilidade e isolamento do ambiente. Isso permite uma montagem rápida do ambiente para execução

do experimento, pois oferece um ambiente controlado e replicável, o que é essencial para a consistência do estudo experimental.

5.1.2 Procedimentos do experimento

Para reproduzir o uso do sistema em um cenário mais próximo da realidade, o experimento consistiu nas seguintes etapas:

1. Importação de Fonte de Dados em Formato CSV:

   Os usuários realizaram a importação de um arquivo CSV com dados artificiais relacionados à avaliação de desempenho, executando o processo de ingestão de dados.

2. Criação de *Dataset*: Após a importação, foi criado um *dataset* a partir dos dados, e o sistema foi avaliado em termos de facilidade de uso e clareza da interface.

3. Enriquecimento de Metadados e Classificação LGPD:

   Os usuários enriqueceram os metadados adicionando descrições detalhadas para cada coluna e categorizando-os de acordo com as diretrizes da LGPD, especialmente para dados sensíveis e de identificação pessoal.

4. Geração de Documentação:

   Com o *dataset* completo e enriquecido, o sistema foi utilizado para gerar um arquivo PDF de documentação, que foi analisado quanto à clareza, organização e capacidade de comunicar os metadados essenciais.

5. Questionário Pós-Tarefas:

   Ao final, os participantes responderam a um questionário baseado no TAM, avaliando percepções de utilidade e facilidade de uso, além de sua intenção de uso e recomendações para aprimoramento do sistema.

5.1.2.1 Métricas de Avaliação No experimento realizado, o TAM e o GQM foram integrados para proporcionar uma avaliação abrangente do sistema Sophiadata. O TAM foi utilizado para analisar a aceitação tecnológica dos usuários, considerando percepções de utilidade, facilidade de uso, intenção de uso e recomendação do sistema. Esse modelo ajudou a captar as avaliações subjetivas dos participantes, especialmente nas questões relacionadas à percepção de utilidade, facilidade de uso e intenção de adoção da ferramenta.

Por outro lado, o GQM serviu como estrutura para definir métricas objetivas que refletem o desempenho do sistema em termos de eficiência na execução de tarefas e conformidade com diretrizes, como a LGPD. Essa abordagem proporcionou uma avaliação prática das funcionalidades principais do Sophiadata, incluindo a importação de dados, enriquecimento de metadados e geração de documentação. A integração entre TAM e GQM permitiu capturar tanto as percepções subjetivas quanto os aspectos técnicos, oferecendo uma visão complementar que considera a experiência do usuário e os resultados quantitativos alcançados.

- Eficácia de Importação de Dados e documentação dos dados:

  Tempo necessário para importar o arquivo CSV, converter em *dataset*, enriquecer 3 colunas, classificar segundo a LGPD e verificar a documentação dos dados.

  Objetivos alcançados - Um resultado bem-sucedido com um sistema. Obtém-se o valor SIM caso o objetivo da seja plenamente atingido, e NÃO em caso contrário;

  Tempo de tarefa - Tempo para conclusão completa e precisa de uma tarefa. Obtém-se o valor médio em minutos, para conclusão da tarefa com êxito;

  Independe de proficiência - Avalia se pode ser usado por pessoas que não possuem conhecimentos, habilidades ou experiências específicas, e nestes casos, obtém-se o valor SIM. Ou se deve ser potencialmente utilizável por pessoas com habilidades específicas, e nestes casos, obtém-se valor NÃO.

- Percepção de Utilidade (PU):

  Avaliação do impacto percebido do sistema na melhoria de documentação e governança de dados.

- Percepção de Facilidade de Uso (PEOU):

  Grau de facilidade na realização das tarefas de importação, enriquecimento e geração de documentação.

- Conformidade com LGPD:

  Capacidade do sistema de permitir a categorização adequada dos dados conforme as diretrizes de privacidade e segurança de dados.

- Satisfação e Intenção de Uso:

  Avaliação geral do sistema e intenção dos participantes de utilizar o Sophiadata em cenários reais.

Por fim as questões definidas para atender às métricas de avaliação do sistema foram:

- **Q1**: O sistema permite documentar metadados de maneira eficiente?
- **Q2**: A funcionalidade de enriquecimento de metadados é útil?
- **Q3**: O sistema facilita a classificação dos dados em relação à LGPD na gestão de dados?
- **Q4**: O sistema gera o arquivo de documentação do *dataset* com clareza, organização e detalhamento adequados?
- **Q5**: O processo de adicionar e configurar uma nova fonte de dados é simples?
- **Q6**: O sistema é intuitivo e fácil de navegar? A interface facilita o acesso rápido às principais funcionalidades?

- **Q7**: Eu pretendo adotar o sistema para documentar e gerenciar dados.
- **Q8**: Recomendaria o uso deste sistema para outros profissionais.

Cada questão tem como objetivo facilitar um ponto específico da análise em relação ao sistema, conforme descrito a seguir:

- **Q1** (*Eficiência na documentação de metadados*): Avalia a percepção sobre a eficiência do sistema na documentação de metadados. Valores altos indicam uma forte aceitação do sistema nesta funcionalidade.
- **Q2** (*Utilidade do enriquecimento de metadados*): Questiona a relevância prática dessa funcionalidade. Resultados baixos podem indicar falta de entendimento ou limitações percebidas na funcionalidade.
- **Q3** (*Classificação em relação à LGPD*): Aponta a percepção de quão bem o sistema ajuda na gestão de dados com conformidade à LGPD. Resultados baixos aqui seriam preocupantes, dada a importância dessa questão.
- **Q4** (*Clareza na documentação do dataset*): Questiona se a documentação gerada pelo sistema é clara e organizada. Resultados baixos indicariam problemas na comunicação dos dados.
- **Q5** (*Simplicidade na configuração de novas fontes*): Verifica a usabilidade no início de uso do sistema, uma métrica importante para a adoção por novos usuários.
- **Q6** (*Intuitividade e navegação do sistema*): Mede a experiência do usuário com a interface, um aspecto importante para a retenção e recomendação.
- **Q7** (*Intenção de adoção*): Reflete o comprometimento dos usuários em utilizar o sistema. Resultados baixos aqui podem dificultar a aceitação geral.
- **Q8** (*Recomendação para outros profissionais*): Mede um indicador direto da aceitação e popularidade do sistema.

5.1.3 Execução do experimento

A execução seguiu as etapas definidas no experimento, e os dados coletados incluíram o tempo de execução para cada etapa, dificuldades relatadas pelos usuários e a taxa de sucesso em completar as tarefas de importação, enriquecimento e geração de documentação. Cada participante foi acompanhado para registrar a sequência de interação com o sistema e as percepções relatadas.

Os dados quantitativos (tempos de execução, respostas do questionário) foram coletados automaticamente pelo sistema e pelo Google Forms, enquanto os dados qualitativos (comentários e sugestões) foram documentados manualmente.

5.1.4 Análise dos resultados do estudo experimental

Os resultados foram analisados com base nas respostas do questionário e nos dados de uso registrados. As percepções de utilidade e facilidade de uso do sistema apresentaram valores predominantemente altos, indicando uma aceitação positiva da ferramenta. A classificação LGPD mostrou-se eficaz, com os usuários classificando corretamente os dados sensíveis e adicionando descrições que enriquecem o entendimento dos metadados.

Tabela 5.1 Métricas GQM Questão 1 Taxa de Conclusão. Tempo de Execução e experiencia

Métrica	Descrição	Unidade Medida	Resultado
M1	Objetivo Alcançado	SIM/NÃO	SIM
M2	Tempo Médio	20.67	30
M3	Independe de proficiência	SIM/NÃO	SIM

A Tabela 5.1 apresenta as métricas GQM aplicadas para avaliar a Questão 1 do experimento. Essa questão tem como foco verificar a eficiência do sistema em documentar metadados e a experiência geral dos participantes durante as etapas do processo. A seguir, cada métrica é detalhada:

1. M1 (Objetivo Alcançado): Esta métrica verifica se os participantes conseguiram concluir todas as etapas previstas (importação de dados, criação do *dataset*, enriquecimento de metadados e geração de documentação). O resultado "SIM" indica que o objetivo foi plenamente atingido por todos os participantes.

2. M2 (Tempo Médio): Mede o tempo necessário, em minutos, para a conclusão das tarefas. O tempo médio registrado foi de 20,67 minutos, com variações de desempenho entre os participantes, sendo 30 minutos o tempo máximo observado.

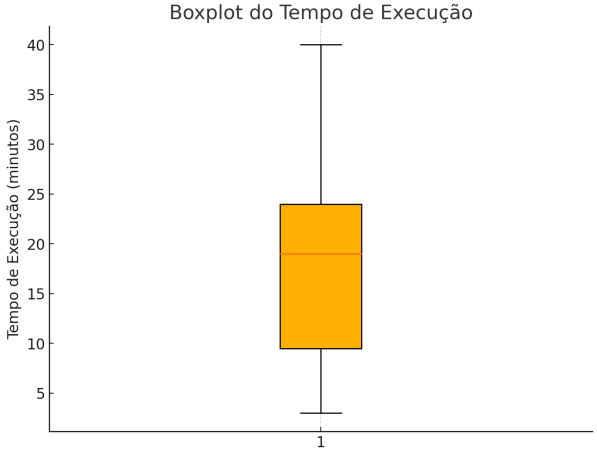
3. M3 (Independência de Proficiência): Avalia se o sistema pode ser utilizado sem a necessidade de conhecimentos técnicos avançados. O resultado "SIM" mostra que mesmo participantes com níveis variados de experiência conseguiram executar as tarefas propostas com sucesso.

A Figura 5.1 apresenta a distribuição dos tempos de execução das tarefas realizadas pelos participantes durante o experimento. As tarefas incluíram a importação de dados, criação de *datasets*, enriquecimento de metadados e geração de documentação.

Os resultados indicam que o tempo mediano para a conclusão das tarefas foi aproximadamente 20 minutos, conforme mostrado pela linha central dentro da caixa. A maior parte dos tempos registrados está concentrada dentro do intervalo entre o primeiro e o terceiro quartil, representando os 50% intermediários dos participantes, com variações típicas que refletem pequenas diferenças individuais no desempenho.

O gráfico também revela alguns *outliers*, correspondentes a tempos de execução significativamente mais altos. Esses valores podem estar associados a fatores como menor experiência dos participantes, dificuldades específicas encontradas durante as tarefas ou diferenças nas configurações de *hardware* utilizadas no experimento.

Figura 5.1 Tempo de Execução do Experimento.



Fonte: O Autor, 2024.

De forma geral, a maioria dos participantes concluiu as tarefas dentro de um intervalo de tempo de 30 minutos, com pouca dispersão, o que reflete a consistência do sistema em termos de desempenho. No entanto, os tempos extremos sugerem a necessidade de analisar em maior detalhe as dificuldades enfrentadas por esses usuários, para identificar possíveis melhorias no suporte às etapas mais complexas do processo.

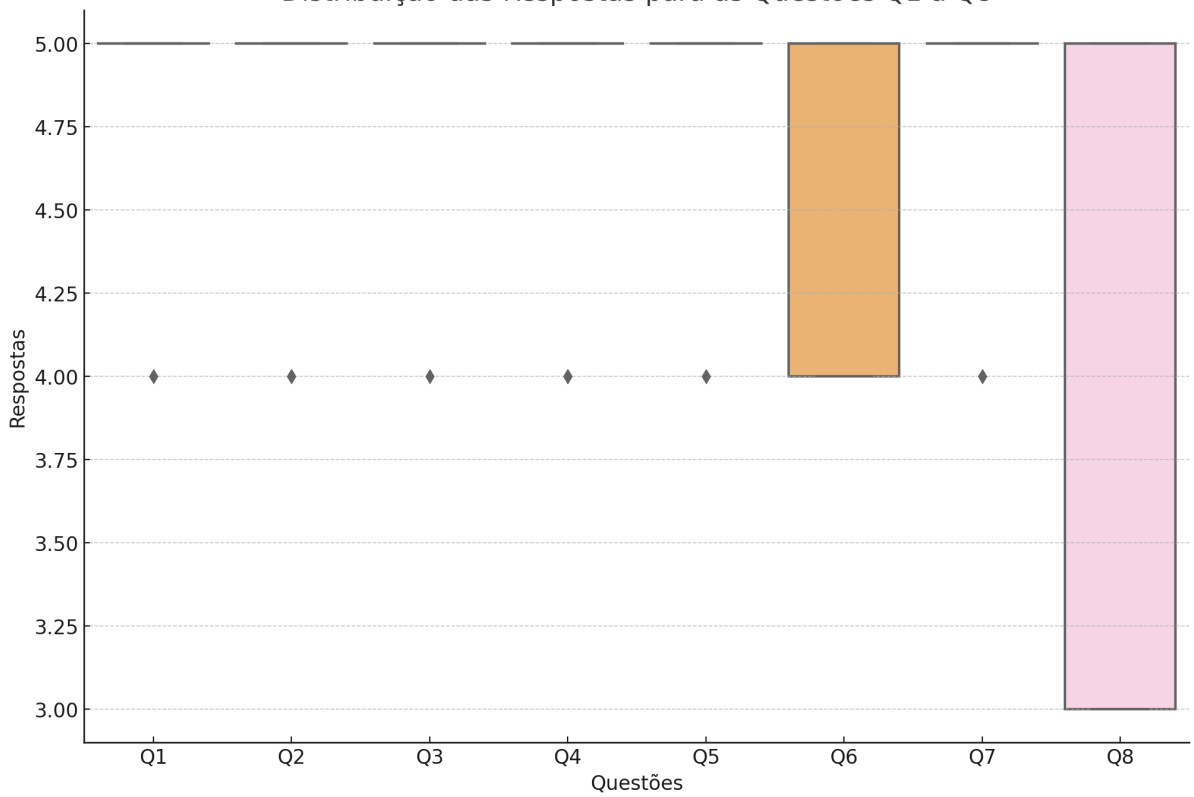
A interpretação dos dados da Figura 5.1 reforça que o Sophiadata apresentou um desempenho estável e eficiente para a maioria dos participantes, com algumas oportunidades de aprimoramento em termos de acessibilidade e usabilidade para casos específicos.

A análise das respostas às questões Q1 a Q8, representada na Figura 5.2, revela padrões interessantes sobre a percepção dos participantes em relação às funcionalidades e à experiência geral com o sistema. Em termos gerais, observa-se uniformidade nas respostas para a maioria das questões, mas também foram identificadas áreas que demandam maior atenção.

As questões Q1, Q2, Q3, Q4, Q5 e Q7 apresentam respostas predominantemente concentradas em valores altos (majoritariamente 5), sugerindo uma percepção consistente e positiva sobre essas funcionalidades. Esse padrão reflete que a maioria dos usuários compartilha opiniões similares quanto à eficiência do sistema na documentação de metadados (Q1), utilidade do enriquecimento (Q2), conformidade com a LGPD (Q3), clareza da documentação (Q4), simplicidade no processo de configuração de novas fontes (Q5) e intenção de adoção (Q7).

Por outro lado, maior variabilidade é observada nas questões Q6 e Q8. A questão Q6, que avalia a intuitividade do sistema, apresenta algumas respostas em níveis mais baixos, indicando que nem todos os participantes consideraram o sistema fácil de usar ou navegar. Já a questão Q8, que mede a recomendação do sistema para outros profissionais, apresenta uma dispersão ainda maior, com algumas respostas significativamente inferiores. Essa variabilidade pode refletir barreiras ou pontos críticos que dificultam a recomendação, como falta de funcionalidades adicionais ou complexidades específicas percebidas por alguns participantes.

Figura 5.2 Respostas para as Questões Q1 a Q8



Fonte: O Autor, 2024.

Esses resultados destacam a importância de priorizar melhorias nas funcionalidades avaliadas pelas questões Q6 e Q8, visto que ambas afetam diretamente a adoção e a recomendação do sistema. A percepção de intuitividade é um fator determinante para garantir uma experiência de uso satisfatória, enquanto a recomendação está diretamente ligada à aceitação geral da tecnologia por um público mais amplo.

Os participantes destacaram a clareza do processo de importação de dados e a facilidade de navegar entre as funcionalidades. A geração de documentação foi bem avaliada, sendo considerada clara e adequada para suportar a governança de dados, embora algumas sugestões para a melhoria na customização dos relatórios tenham sido mencionadas.

5.2 LIMITAÇÕES DO EXPERIMENTO

Embora o experimento tenha sido cuidadosamente planejado, existem alguns pontos que devem ser observados em relação aos resultados obtidos. Esses estão relacionados principalmente ao perfil dos participantes, ao contexto do experimento e à natureza das métricas avaliadas.

5.2.1 Validade Interna

- **Experiência prévia dos participantes**: O experimento foi realizado com alunos do curso de Bacharelado em Engenharia de Software, em sua maioria do 4º semestre. Esses alunos já cursaram disciplinas relacionadas a banco de dados e programação web, o que pode ter influenciado positivamente as avaliações das funcionalidades do sistema. Participantes com menos experiência podem apresentar percepções diferentes.

- **Uniformidade do perfil dos participantes**: O grupo de alunos pode não representar adequadamente um público mais diverso de profissionais da área, como analistas de dados, administradores de banco de dados ou desenvolvedores com níveis variados de experiência.

- **Influência do ambiente acadêmico**: Por serem alunos de um curso acadêmico, os participantes podem ter avaliado o sistema com base em expectativas relacionadas ao contexto educacional, e não necessariamente ao contexto real de uso profissional.

5.2.2 Validade Externa

- **Generalização dos resultados**: Como os participantes pertencem a um único curso e semestre específico, os resultados podem não ser generalizáveis para outros contextos, como empresas, organizações governamentais ou equipes multidisciplinares.

- **Aplicabilidade das métricas em outros cenários**: As questões avaliadas podem ser percebidas de maneira diferente em ambientes fora do âmbito acadêmico, especialmente em situações com maior pressão de uso em produção.

5.2.3 Validade de Construção

- **Formulação das questões**: Algumas questões podem ter sido interpretadas de forma ambígua pelos participantes, afetando a confiabilidade das respostas. Revisões mais detalhadas das questões poderiam minimizar este risco.
- **Métricas subjetivas**: As respostas dependem diretamente da percepção individual dos participantes, o que pode introduzir vieses relacionados à interpretação pessoal.

Esses pontos devem ser levados em consideração ao interpretar os resultados do experimento, e futuras iterações do estudo podem incluir estratégias para mitigar tais limitações à validade.

Capítulo 6

CONSIDERAÇÕES FINAIS

Considerando o cenário atual, que envolve o movimento de governo aberto e a gestão de dados, a qualidade dos dados governamentais abertos tem sido cada vez mais cobrada pela sociedade. No entanto, os dados disponibilizados frequentemente apresentam problemas de formato e qualidade, tanto nos dados em si quanto nos metadados. Esses desafios decorrem das dificuldades enfrentadas ao longo do processo, desde a extração até a publicação das informações.

Este trabalho apresenta uma solução integrada focada no gerenciamento e no compartilhamento de dados. Esta abordagem incluiu a simplificação dos processos de extração e tratamento de dados e metadados, com ênfase na melhoria da qualidade e transparência dos dados governamentais.

De maneira geral, o Sophiadata se posiciona como uma solução computacional web que auxilia na gestão, no tratamento e na publicação de dados abertos governamentais. Através das funcionalidades da ferramenta, foram mitigadas as dificuldades técnicas do processo de extração e tratamento de dados em órgãos governamentais. Entre as funcionalidades destacadas, podemos ressaltar a classificação dos dados segundo a LGPD e as funcionalidades de governança de dados, que permitem auxiliar a conformidade com as leis e regulamentos de privacidade.

A validação da ferramenta indica não apenas a simplificação do processo de tratamento de dados, mas também a melhoria na qualidade dos dados que devem ser publicados em portais governamentais, como os dados da saúde, educação e segurança pública, por exemplo. O uso da ferramenta contribuiu para otimizar o processo de validação de conformidade e a organização dos dados em um formato adequado para os usuários finais.

Segundo Batista (2021), para que os dados sejam eficientemente utilizados e compreensíveis tanto para humanos quanto para máquinas, é aconselhável que alcancem pelo menos o nível de três estrelas do sistema de 5 estrelas para Dados Abertos. Os resultados apresentados no capítulo 4 demonstram que o Sophiadata facilitou o compartilhamento de arquivos de dados e metadados, garantindo que os dados alcançassem o nível três estrelas do sistema de 5 estrelas para Dados Abertos, o que é um marco importante para a acessibilidade e reutilização dos dados abertos.

A solução desenvolvida oferece funcionalidades que facilitam a ingestão de dados, a aplicação de transformações nos dados e a extração de metadados. Além disso, permite a gestão eficiente desses metadados, possibilitando seu enriquecimento e publicação subsequente em diferentes formatos.

Ainda com base na análise dos resultados obtidos na simulação, permitiu inferir que a especificação e o desenvolvimento da ferramenta Sophiadata atenderam de maneira satisfatória às necessidades levantadas na pesquisa. A simulação da ferramenta foi o método empregado para uma avaliação mais eficaz da ferramenta e demonstrar sua eficiência e benefícios em contextos mais próximos da realidade.

Um dos resultados obtidos durante o processo de construção da ferramenta foi seu uso para pré-processar e organizar o *dataset* que foi utilizado no artigo de "Avanços no acesso de estudantes pretos e pardos à Educação Profissional, Científica e Tecnológica na rede federal utilizando dados da plataforma Nilo Peçanha" publicado na Revista Brasileira da Educação Profissional e Tecnológica (RBEPT) e disponível em https://doi.org/10.15628/rbept.2024.15199 (PEREIRA; NOVAIS, 2024).

Por fim, conforme detalhado neste trabalho, o desenvolvimento do *software* Sophiadata foi formalmente registrado junto ao INPI, sob o número de processo SEI 23278.009135/2024-61. O pedido foi recebido e protocolado com a numeração BR 512024004735-2, resultando na expedição do Certificado de Registro de Programa de Computador em 10/12/2024, conforme apresentado no Anexo I.

6.1 LIMITAÇÕES E TRABALHOS FUTUROS

Durante o processo de validação, o sistema também apresentou alguns pontos a serem melhorados. Os participantes da validação apresentaram possíveis melhorias de usabilidade.

Para continuação da trabalho, a expansão do número de experimentos, incorporando uma variedade de conjuntos de dados com diferentes características permitirá testar a ferramenta em cenários mais complexos e realistas, melhorando a generalização dos resultados. Além disso, seria benéfico explorar a integração do Sophiadata com ferramentas de visualização de dados e plataformas de publicação de dados, avaliando sua eficácia em um ambiente operacional mais amplo e variado. Esses aprimoramentos contribuirão para uma compreensão mais profunda da aplicabilidade e eficiência do Sophiadata, assegurando que ele atenda às necessidades de instituições diversas na gestão de dados e metadados.

Durante o processo de validação da ferramenta, o sistema apresentou alguns pontos a serem melhorados, especialmente no que diz respeito à usabilidade da plataforma. Os participantes da validação sugeriram melhorias que podem ser incorporadas na próxima versão do sistema. Essas sugestões indicam que, embora a ferramenta seja eficaz para os objetivos propostos, ainda há espaço para ajustes que a tornariam mais acessível, intuitiva e eficiente.

Ainda que a validação realizada tenha apresentado resultados positivos, o sistema necessita de uma nova etapa de validação. Para mitigar uma das limitações observadas que foi a amostra restrita de participantes na fase de validação. Embora o número de partici-

pantes tenha sido suficiente para fornecer uma primeira análise qualitativa da ferramenta, uma nova etapa de validação, com um volume maior de usuários, seria fundamental para ampliar a abrangência dos resultados. A coleta de mais informações sobre a experiência do usuário permitirá uma análise mais robusta e fornecerá subsídios para aprimorar o sistema. Além disso, o aprimoramento do questionário utilizado na validação, com a inclusão de questões mais específicas e relevantes, poderá gerar retornos mais detalhados e aplicáveis.

Como trabalhos futuros a aplicação de Inteligência Artificial (IA) pode ser uma solução importante para aumentar a eficiência e a precisão dos processos de validação e tratamento de dados. No futuro, pode-se integrar IA ao Sophiadata para automatizar a validação de dados, identificar padrões e inconsistências automaticamente, aprimorar o processo de classificação de dados sensíveis e garantir a conformidade com as leis de privacidade, como a LGPD, sem a necessidade de intervenção manual, aumentando a escalabilidade e a segurança do sistema.

A IA também poderia ser utilizada para otimizar o tratamento de dados em diferentes contextos, como identificar automaticamente os dados faltantes ou incoerentes, sugerir transformações e até corrigir problemas com base em dados históricos.

Esses aprimoramentos não só aumentariam a utilidade e a eficiência do Sophiadata, mas também garantiriam que a solução se tornasse um recurso ainda mais inteligente, capaz de se adaptar automaticamente às necessidades dinâmicas das instituições que o utilizam. A automação do tratamento e da validação dos dados com IA representa uma direção estratégica importante para o futuro do sistema, ampliando suas capacidades e assegurando uma evolução contínua da plataforma.

A Figura 5.2 evidencia visualmente esses padrões, com baixa variabilidade nas questões que obtiveram respostas mais consistentes e maior dispersão nas questões consideradas críticas. Essa análise detalhada orienta futuros aprimoramentos no sistema, visando atender às expectativas dos usuários de forma mais eficaz.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBUQUERQUE, G. T. R. de; ALVES, G. M. R.; MACIEL, A. M. A. Desenvolvimento de um modelo de ingestão de dados para automl. *Revista de Engenharia e Pesquisa Aplicada*, v. 7, n. 3, p. 19–28, 2022.

ARAKAKI, A. C. S.; ARAKAKI, F. A. Dados e metadados: conceitos e relações: concepts and relationships. *Ciência da Informação*, v. 49, n. 3, 2020.

BATISTA, G. V. *Dados Abertos Governamentais: Avaliação de Maturidade e Qualidade*. Dissertação (B.S. thesis), 2021.

BERNERS-LEE, T. *Linked Data*. 2009. Disponível em: ⟨https://www.w3.org/DesignIssues/LinkedData.html⟩. Acesso em: 25 de janeiro de 2024.

BHAGESHPUR, K. *Data is The New Oil and That's a Good Thing*. 2019. Disponível em: ⟨https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing⟩. Acesso em: 16 de Junho de 2023.

BONETTI, L. G. Metadados e os princípios fair: um estudo dos repositórios de dados de pesquisa em são paulo. Universidade Federal de São Carlos, 2023.

BRASIL. *Lei nº 12.527, de 18 de novembro de 2011*. 2011. ⟨http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm⟩. Acesso em: 10 out. 2023.

BRASIL. *Lei nº 13.709, de 14 de agosto de 2018*. 2018. ⟨http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm⟩. Acesso em: 10 out. 2023.

CHAPLIN, B. et al. Patent Application, *DATA CATALOG AND RETRIEVAL SYSTEM*. Minneapolis, MN: United States Patent and Trademark Office, 2024. Patent Application Publication. Pub. No.: US 2024/0028606 A1.

CONSORTIUM, W. W. W. et al. Data catalog vocabulary (dcat). World Wide Web Consortium, 2014.

Controladoria-Geral da União. *Portal Brasileiro de Dados Abertos*. 2024. Disponível em: ⟨https://dados.org.br⟩. Acesso em: 25 de janeiro de 2024.

CRISTÓVAM, J. S. da S.; HAHN, T. M. Administração pública orientada por dados: Governo aberto e infraestrutura nacional de dados abertos. *Revista de Direito Administrativo e Gestão Pública*, v. 6, n. 1, p. 1–24, 2020.

CZAJKOWSKI, K.; KESSELMAN, C.; SCHULER, R. Ermrest: A collaborative data catalog with fine grain access control. In: IEEE. *2017 Ieee 13th international conference on E-science (E-science)*. [S.l.], 2017. p. 510–517.

DataHub Project. *DataHub - An open-source metadata platform for the modern data stack*. 2024. Disponível em: ⟨https://datahubproject.io⟩. Acesso em: 10 de dezembro de 2023.

EVANS, A. M.; CAMPOS, A. Open government initiatives: Realizing principles of citizen participation. Lyndon Baines Johnson School of Public Affairs, 2009.

FORTINI, C.; AMARAL, G.; CAVALCANTI, C. M. L. Lgpd x lai: sintonia ou antagonismo. *Lei geral de proteção de dados no setor público. Belo Horizonte: Fórum*, p. 101–122, 2021.

GIL, A. C. *Métodos e técnicas de pesquisa social*. [S.l.]: 6. ed. Ediitora Atlas SA, 2008. 200 p.

GO-FAIR. Germany; The Netherlands; Paris: [s.n.], 2022. Disponível em: ⟨https://www.go-fair.org/fair-principles/⟩. Acesso em: 16 de Junho de 2023.

ILYAS, I. F.; CHU, X. *Data cleaning*. [S.l.]: Morgan & Claypool, 2019.

ISO/IEC. *Information technology – Specification and standardization of data elements – Part 1: Framework for data definitions*. [S.l.], 2023.

JAHNKE, N.; SPIEKERMANN, M.; RAMOUZEH, B. *Data Catalogs: Implementing Capabilities for Data Curation, Data Enablement and Regulatory Compliance*. [S.l.]: Fraunhofer Institut für Software-und Systemtechnik, 2022.

KORTE, T. et al. Data catalogs-integrated platforms for matching data supply and demand. Fraunhofer Verlag, 2019.

OECD. *Open Government in Latin America*. [s.n.], 2014. 260 p. Disponível em: ⟨https://www.oecd-ilibrary.org/content/publication/9789264223639-en⟩.

OECD. *The Path to Becoming a Data-Driven Public Sector*. Organisation for Economic Co-operation and Development, 2019. Disponível em: ⟨https://www.oecd.org/gov/the-path-to-becoming-a-data-driven-public-sector.pdf⟩.

OLESEN-BAGNEUX, O. *The Enterprise Data Catalog*. First. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media, Inc., 2023. ISBN 978-1-098-13531-7.

Open Knowledge Foundation. *CKAN - The open source data portal software*. 2024. Disponível em: ⟨https://ckan.org⟩. Acesso em: 21 de dezembro de 2023.

PEREIRA, M. J.; NOVAIS, R. L. Avanços no acesso de estudantes pretos e pardos à educação profissional, científica e tecnológica na rede federal utilizando dados da plataforma nilo peçanha. *Revista Brasileira da Educação Profissional e Tecnológica*, v. 3, p. e15199, nov. 2024. Disponível em: ⟨https://www2.ifrn.edu.br/ojs/index.php/RBEPT/article/view/15199⟩.

PINHO, M. D. C. Dados abertos governamentais: usuários e apropriações sociais no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 2021.

PRESSMAN, R. S.; MAXIM, B. R. *Engenharia de software-9*. [S.l.]: McGraw Hill Brasil, 2021.

SOUZA, H. L. D. de. *Visualize Your Region (VYR): Visualização de Dados Sob Perspectiva Regionalizada*. Dissertação (Dissertação de Mestrado) — Instituto Federal de Educação, Ciência e Tecnologia da Bahia, Salvador BA, agosto 2021.

STACH, C. Data is the new oil–sort of: A view on why this comparison is misleading and its implications for modern data administration. *Future Internet*, v. 15, n. 2, 2023. ISSN 1999-5903. Disponível em: ⟨https://www.mdpi.com/1999-5903/15/2/71⟩.

WAZLAWICK, R. S. *Metodologia de pesquisa para ciência da computação*. [S.l.]: Elsevier, 2009.

WILKINSON, M. D. et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, Nature Publishing Group, v. 3, n. 1, p. 1–9, 2016.

Capítulo

I

CERTIFICADO DE REGISTRO DE PROGRAMA DE COMPUTADOR

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512024004735-2**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 03/06/2023, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: Sophiadata

Data de criação: 03/06/2023

Titular(es): INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA BAHIA - IFBA

Autor(es): RENATO LIMA NOVAIS; PABLO VIEIRA FLORENTINO; MARIO JORGE PEREIRA

Linguagem: HTML; JAVA; JAVA SCRIPT; CSS

Campo de aplicação: IF-10

Tipo de programa: GI-01

Algoritmo hash: SHA-512

Resumo digital hash:
887c4eb5be847def5c794f92a1bd1aba52775f13bc696fc307c651735f9a7ca3ddbb1372bb7de7871adf03640d75fe14cd8c30574c474f948beac79d424274b4

Expedido em: 10/12/2024

Aprovado por:
Carlos Alexandre Fernandes Silva
Chefe da DIPTO