



MESTRADO EM SISTEMAS E COMPUTAÇÃO

MARCELO MOREIRA WEST

**DADOS ABERTOS GOVERNAMENTAIS NA PRÁTICA: DESAFIOS NA SUA
DESCOBERTA, ACESSO E USO**

Salvador
2022

MARCELO MOREIRA WEST

**DADOS ABERTOS GOVERNAMENTAIS NA PRÁTICA: DESAFIOS NA SUA
DESCOBERTA, ACESSO E USO**

Dissertação apresentada ao Mestrado em Sistemas e Computação da Universidade Salvador – UNIFACS, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientador: Prof. Dr. Glauco de Figueiredo Carneiro.

Salvador
2022

Ficha Catalográfica elaborada pelo Sistema de Bibliotecas da Universidade Salvador
UNIFACS.

West, Marcelo Moreira.

Dados abertos governamentais na prática: desafios na sua descoberta, acesso e uso/
Marcelo Moreira West.- Salvador: UNIFACS, 2022.

75 f.: il.

Dissertação apresentada ao Programa de Pós-Graduação, Mestrado em Sistemas e
Computação da Universidade Salvador - UNIFACS, como requisito parcial para
obtenção do título de Mestre.

Orientador Prof. Dr. Glauco de Figueiredo Carneiro.

1. Ciência da computação. 2. Dados abertos. 3. Qualidade dos dados. I. Carneiro,
Glauco de Figueiredo, orient. II. Título.

CDD:004

MARCELO MOREIRA WEST

DADOS ABERTOS GOVERNAMENTAIS NA PRÁTICA: DESAFIOS NA SUA
DESCOBERTA, ACESSO E USO

Dissertação apresentada ao Mestrado em Sistemas e Computação da Universidade Salvador - UNIFACS, como requisito parcial para obtenção do título de mestre e aprovada pela seguinte banca examinadora:

Glauco de Figueiredo Carneiro – Orientador _____ Doutor
em Ciência da Computação - Universidade Federal da Bahia (UFBA)
Universidade Salvador - UNIFACS

Jorge Alberto Prado de Campos _____
Doutor em Spatial Information Science and Engineering - University of Maine, Orono, USA.
Universidade Salvador – UNIFACS

Juliano Lopes de Oliveira _____
Doutor em Ciência da Computação - Universidade Estadual de Campinas (UNICAMP)
Universidade Federal de Goiás (UFG)

Salvador, 21 de junho de 2022.

Queremos dedicar a Deus
mais esta vitória em nossas vidas.
Foi ele que nos deu sabedoria,
orientou cada um de nós
para fazer as melhores escolhas
e que está ao nosso lado nesse momento.

“Ensinar não é transferir conhecimento, mas criar as possibilidades para a sua própria produção ou a sua construção.” - Paulo Freire (2003).

RESUMO

Contexto: A implementação de dados abertos governamentais é cada vez mais frequente devido a cobrança da sociedade por transparência na gestão pública. A literatura tem mostrado que existem diversas técnicas disponíveis para avaliação da qualidade de dados abertos governamentais. *Problema:* A avaliação da qualidade de dados abertos governamentais não é trivial, particularmente quando realizada de forma manual, motivo que pode afetar a eficácia da avaliação. Dentre os desafios desta atividade, podem ser citados, dentre outros, o esforço para a identificação dos metadados para apoiar a avaliação. *Motivação:* A avaliação efetiva da qualidade de dados abertos governamentais viabiliza a identificação de oportunidades de melhorias, apoia a seleção de repositórios que atendem a requisitos mínimos de qualidade e, conseqüentemente, favorece a descoberta, acesso e uso destes tipos dados. *Objetivo:* Analisar a qualidade de dados publicados em portais governamentais brasileiros, com o propósito de identificar limitações e erros com respeito a descoberta, acesso e uso dos dados, do ponto de vista de pesquisadores e usuários, no contexto de programas sociais e transparência nos gastos de pessoal. *Questão de pesquisa:* A estrutura de dados abertos governamentais implementada por portais específicos disponibilizados pelo Governo Brasileiro atende os requisitos das dimensões de qualidade de dados? *Metodologia:* Um estudo exploratório foi realizado em duas partes. A primeira parte avaliou manualmente os dados do repositório do Fundo de Financiamento ao Estudante do Ensino Superior (FIES) em relação ao grau de alinhamento com as dimensões de qualidade de dados. A partir da primeira parte do estudo exploratório, foi proposta uma solução semiautomatizada para apoiar a avaliação dos dados e metadados e a conseqüente seleção de dados abertos governamentais. A segunda parte do estudo exploratório foi conduzida para avaliar a efetividade da solução proposta. *Resultados:* Foram analisados de forma semiautomática os metadados dos repositórios de dados abertos (FIES, Minha Casa Minha Vida e Portal da Transparência). A quantidade de recursos avaliados foi 107 no primeiro, 7 no segundo e no terceiro 357, respectivamente representam 22,72%, 1,49% e 75,80%. A quantidade total de recursos avaliados foi 471, ou seja, 100% dos recursos. Os resultados da análise mostram que no primeiro repositório foram encontrados apenas os metadados (url, format e description) inválidos. Enquanto no segundo repositório foi encontrado apenas o metadado (description) inválido e no terceiro foram encontrados os metadados (format, created, name e description) inválidos. Além disso, foram encontrados recursos em que a url identificada no metadado 'url' não pode ser acessada para validação e foram encontrados os metadados (format, description, created e name) sem o devido preenchimento, isto é, foram encontradas não conformidades com as dimensões de qualidade. Nota-se a efetividade da solução proposta para avaliação dos dados e metadados de portais de dados abertos governamentais. *Contribuições esperadas:* Apoiar a identificação e endereçamento de oportunidades de melhorias dos portais nacionais de dados abertos governamentais analisados. Acesso e uso da solução implementada no contexto desta dissertação para apoiar a avaliação da qualidade dos metadados, e, por extensão, apoiar também na identificação de repositórios de excelência. Estes últimos podem servir de referência para outros repositórios que precisam evoluir e, portanto, melhorar o alinhamento às dimensões de qualidade de dados.

Palavras-chave: Dados abertos, qualidade dos dados, dimensões de qualidade.

ABSTRACT

Context: The implementation of government open data is increasingly frequent due to society's demand for transparency in public management. The literature has shown that there are several techniques available to assess the quality of open government data. *Problem:* Assessing the quality of open government data is not trivial, particularly when performed manually, which is why may affect the effectiveness of the assessment. Among the challenges of this activity, we can mention, among others, the effort to identify the metadata to support the evaluation. *Motivation:* Effective assessment of the quality of open government data enables the identification of improvement opportunities, supports the selection of repositories that meet to minimum quality requirements and, consequently, favors the discovery, access and use of these data types. *Goal:* To analyze the quality of data published in Brazilian government portals, for the purpose of identifying limitations and errors with respect to discovery, access and use data, from the point of view of researchers and users, in the context of social programs and transparency in personnel expenses. *Research Question:* Does the government open data framework implemented by specific portals made available by the Brazilian government meet the requirements of the data quality dimensions? *Methodology:* An exploratory study was carried out in two parts. The first part manually evaluated the data from the Student Financing Fund repository of the Higher Education (FIES) in relation to the degree of alignment with the quality dimensions of data. From the first part of the exploratory study, a solution was proposed semi-automated to support data and metadata evaluation and subsequent selection of government open data. The second part of the exploratory study was conducted to evaluate the effectiveness of the proposed solution. *Results:* The metadata of the repositories of open data (FIES, Minha Casa Minha Vida and Transparency Portal). The amount of resources evaluated was 107 in the first, 7 in the second and 357 in the third, respectively represent 22.72%, 1.49% and 75.80%. The total amount of resources evaluated was 471, or ie 100% of the resources. The analysis results show that in the first repository only invalid metadata (url, format and description) were found. while in second repository found only the invalid metadata (description) and in the third Invalid metadata (format, created, name and description) was found. In addition Furthermore, resources were found where the url identified in the 'url' metadata cannot be accessed for validation and found the metadata (format, description, created and name) without the proper completion, that is, non-conformities with the quality dimensions. Note the effectiveness of the proposed solution for evaluating the data and metadata from government open data portals. *Expected Contributions:* Support the identification and addressing of opportunities for improvement of the analyzed national open government data portals. Access and use of the solution implemented in the context of this dissertation to support the assessment of metadata quality, and, by extension, also support the identification of excellent repositories. The latter can serve as a reference for other repositories that need to evolve and, therefore, improve alignment with data quality dimensions.

Keywords: Open data. Data quality. Quality dimensions.

LISTA DE FIGURAS

Figura 1 – Etapas da Pesquisa	17
Figura 2 – Linha do Tempo da Abertura dos Dados Governamentais no Brasil e no Exterior	20
Figura 3 – Arquitetura do Framework CKAN.....	25
Figura 4 – Arquitetura do Framework Socrata.....	26
Figura 5 – Etapas do Tutorial	34
Figura 6 – Etapas do Tutorial para Análise de Dados	35
Figura 7 – Etapas da Solução Proposta	42
Figura 8 – <i>Endpoint</i> - Dados do FIES	42
Figura 9 – <i>Endpoint</i> - Dados do Programa Minha Casa Minha Vida.....	43
Figura 10 – <i>Endpoint</i> - Dados do Portal da Transparência - Remuneração Servidores	43
Figura 11 – Arquitetura da Proposta de Análise Semiautomática dos Metadados.....	43
Figura 12 – Processo Manual e Processo Semiautomatizado.....	46
Figura 13 – Retorno da Consulta - API CKAN - Catálogo do FIES.....	47
Figura 14 – Retorno da Consulta - API CKAN - Catálogo Minha Casa Minha Vida.....	48
Figura 15 – Retorno da Consulta - API CKAN - Catálogo do Portal da Transparência.....	48
Figura 16 – Processo de Análise do Metadado URL.....	50
Figura 17 – Processo de Recuperação do Arquivo.....	50
Figura 18 – Processo de Análise do Metadado FORMAT	51
Figura 19 – Processo de Análise do Metadado CREATED	51
Figura 20 – Processo de Análise do Metadado NAME.....	52
Figura 21 – Processo de Análise do Metadado DESCRIPTION.....	52

LISTA DE QUADROS

Quadro 1 – Capítulos com Evidências e Respostas às Questões de Pesquisa.....	18
Quadro 2 – Fases do ProceDA	29
Quadro 3 – Objetivo do Estudo de Acordo com a Abordagem GQM	33
Quadro 4 – Etapas do Tutorial Realizadas pelos Participantes	34
Quadro 5 – Erros de Qualidade Frequentes em Dados Abertos	36
Quadro 6 – Distribuição da Quantidade de Erros por Tipo de Erro	38
Quadro 7 – Sugestões de Melhorias	39
Quadro 8 – Lista de <i>Endpoints</i> Consultados	47
Quadro 9 – Quantidade de Recursos Avaliados por Repositório	53
Quadro 10 – Percentual dos Metadados Válidos e Inválidos do Repositório – FIES	53
Quadro 11 – Percentual dos Metadados Válidos e Inválidos do Repositório - Minha Casa Minha Vida	54
Quadro 12 – Percentual dos Metadados Válidos e Inválidos do Repositório - Portal da Transparência.....	54
Quadro 13 – Limitações Encontradas após a Análise Semiautomática.....	55

LISTA DE ABREVIATURAS E SIGLAS

API	Interface de programação de aplicações
CGU	Controladoria Geral da União
CKAN	Comprehensive Knowledge Archive Network
CSV	Comma-separated values
EUA	Estados Unidos da América
FIES	Financiamento ao Estudante do Ensino Superior
FTP	File Transfer Protocol
GQM	Goal Question Metric
HTTP	Hypertext Transfer Protocol
ITNG	International Conference on Information Technology
JSON	JavaScript Object Notation
LAI	Lei de Acesso à Informação
MCMV	Programa Minha Casa Minha Vida
OCR	Optical Character Recognition
OE	Objetivo Específico
OECD	Organização para a Cooperação
OG	Objetivo Geral
OGD	Open Government Data
PDA	Plano de Dados Abertos
ProceDA	Processo para Dados Abertos
QP	Questão de Pesquisa
QPE	Questão de Pesquisa Específica
SIG	Sistema Integrado de Gestão
TCU	Tribunal de Contas da União
TICS	Tecnologias de Comunicação e Informação
UFRN	Universidade Federal do Rio Grande do Norte
URL	Uniform Resource Locator
XML	Extensible Markup Language

SUMÁRIO

1	INTRODUÇÃO.....	13
1.1	CONTEXTO.....	13
1.2	ABORDAGEM DE PESQUISA	14
1.2.1	Problema de Pesquisa.....	14
1.2.2	Motivação	14
1.2.3	Objetivos da Pesquisa.....	14
1.3	QUESTÕES DE PESQUISA	15
1.4	ETAPAS DA PESQUISA	17
1.4.1	Fase de Planejamento	17
1.4.2	Artigo Produzido	19
1.5	ESTRUTURA DESTA DISSERTAÇÃO	19
2	FUNDAMENTAÇÃO TEÓRICA.....	20
2.1	DADOS ABERTOS GOVERNAMENTAIS	20
2.2	DIMENSÕES DE QUALIDADE DE DADOS	21
2.3	DESAFIOS ENFRENTADOS PELOS GOVERNOS PARA ADOÇÃO DE DADOS ABERTOS GOVERNAMENTAIS	23
2.4	PRINCIPAIS FRAMEWORKS UTILIZADOS PARA FACILITAR A DESCOBERTA, ACESSO E USO DE DADOS ABERTOS GOVERNAMENTAIS	24
2.4.1	Framework CKAN	24
2.4.2	Framework SOCRATA.....	25
2.5	TRABALHOS RELACIONADOS	27
2.6	O CASO DA ABERTURA DOS DADOS DA UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE (UFRN)	28
2.7	RESPOSTA DA QUESTÃO DE PESQUISA ESPECÍFICA 1 (QPE1).....	30
3	ESTUDO EXPLORATÓRIO - PARTE 1	31
3.1	INTRODUÇÃO.....	31
3.2	ANÁLISE DAS DIMENSÕES DE QUALIDADE DE DADOS	32
3.3	METODOLOGIA DA PRIMEIRA PARTE DO ESTUDO EXPLORATÓRIO ..	32
3.3.1	Planejamento.....	33
3.3.2	Execução	35
3.4	RESULTADOS E DISCUSSÃO DE RESULTADOS DA PRIMEIRA PARTE DO ESTUDO EXPLORATÓRIO	36

3.5	Resposta da Questão de Pesquisa Específica QPE2	39
3.6	CONCLUSÃO	39
4	SOLUÇÃO PROPOSTA	41
4.1	JUSTIFICATIVA DE ESCOLHA DO FRAMEWORK CKAN	41
4.2	IMPLEMENTAÇÃO DA SOLUÇÃO PROPOSTA.....	41
4.3	ETAPAS DA SOLUÇÃO PROPOSTA	41
4.3.1	Detalhes da implementação	42
4.4	CONCLUSÃO	44
5	ESTUDO EXPLORATÓRIO - PARTE 2	45
5.1	INTRODUÇÃO	45
5.1.1	Metadados Utilizados no Estudo	47
5.2	PLANEJAMENTO DO ESTUDO	49
5.3	EXECUÇÃO DO ESTUDO CASO	49
5.3.1	Análises Realizadas pelo Algoritmos	49
5.3.2	Resultados.....	53
5.3.3	Análise dos Resultados	53
5.3.4	Limitação identificada no uso da solução.....	56
5.4	RESPOSTA DA QUESTÃO DE PESQUISA (QPE3)	56
5.5	CONCLUSÃO	56
6	CONCLUSÃO E PERSPECTIVAS FUTURAS.....	57
6.1	RESPOSTA DA QUESTÃO DE PESQUISA (QP).....	57
6.2	CONSIDERAÇÕES FINAIS	57
6.3	CONTRIBUIÇÕES	58
6.4	TRABALHOS FUTUROS	59
	REFERÊNCIAS.....	60
	APÊNDICE A – Distribuição da Quantidade de Erros por Atributo	64
	APÊNDICE B – Questionário	66
	APENDICE C – Resultado da Análise do Repositório Minha Casa Minha Vida.....	71
	ANEXO A – Dicionário de Dados	74

1 INTRODUÇÃO

Este Capítulo apresenta o contexto no qual essa dissertação foi conduzida, os problemas abordados, os objetivos estabelecidos, a metodologia de pesquisa adotada, as contribuições esperadas, além da estrutura dos demais capítulos.

1.1 CONTEXTO

A abertura dos dados governamentais possibilita maior controle social, apoia a democracia e a cidadania e agrega melhorias para administração pública (HARRISON *et al.*, 2012). Por outro lado, disponibilizar informações governamentais de qualidade, relevantes e confiáveis, contribui com a contínua aprovação do governante pelo cidadão (FERRANTI *et al.*, 2009).

Nesse contexto, o crescimento da transparência em nível mundial, especificamente ao acesso à informação, segue duas tendências: a primeira se refere a mudanças no governo para a ampliação da abertura dos dados e a segunda está associada ao constante desenvolvimento de Tecnologias de Informação que possibilitam às instituições, organizações e indivíduos encontrar, compartilhar, combinar e reutilizar dados governamentais (DAWES, 2010).

Segundo a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) o uso estratégico de dados, aproxima governo da população e cria oportunidades de melhoria da qualidade dos serviços públicos e da efetividade dos gastos públicos por meio de proteções éticas e ainda contribui com a manipulação mais segura e eficiente dos dados (OECD, 2019).

A abertura dos dados governamentais estimula os negócios e a inovação, assegura a transparência e a prestação de contas, empodera os cidadãos para fazerem escolhas de produtos e serviços. Além disso, através do seu uso frequente e amplo, ocorre melhoria da qualidade dos dados (WALSH, 2016). O acesso do cidadão à base de dados referentes aos atos realizados pela Administração Pública permite que o mesmo identifique ações corruptas e corrobora com o combate à corrupção e ao desvio de finalidade no setor público (AQUINO JÚNIOR *et al.*, 2019).

A evolução das Tecnologias de Comunicação e Informação (TICS) tem ampliado a transparência, pois facilita o acesso aos dados e as informações da Administração Pública (PINHO, 2008). A qualidade das informações prestadas pelo governo está ligada à transparência. Os cidadãos devem ter acesso e entendimento facilitado, na disponibilização e

no acesso aos serviços públicos, no controle das contas públicas e na participação no planejamento e no desenvolvimento das políticas públicas (SILVA; MONTEIRO; REIS, 2020).

Esta dissertação tem foco na caracterização de erros e limitações em repositórios de dados abertos e na análise da qualidade dos dados. A referência para a análise do grau de alinhamento dos dados e metadados é a comparação dos requisitos de qualidade presentes nas dimensões da qualidade de dados. A partir do contexto apresentado, as seções seguintes apresentam uma visão panorâmica do que será discutido nesta dissertação.

1.2 ABORDAGEM DE PESQUISA

Nesta Subseção são elicitados os problemas, objetivos, metodologia e contribuições.

1.2.1 Problema de Pesquisa

Uma tarefa extremamente desafiadora é a avaliação da qualidade de dados abertos governamentais, particularmente quando realizada de forma manual, motivo que pode afetar a eficácia da avaliação. Além de que, pode induzir a erros e produzir resultados divergentes e exigir um esforço extremo para a identificação dos metadados a serem avaliados.

1.2.2 Motivação

A avaliação efetiva da qualidade de dados abertos governamentais viabiliza a identificação de oportunidades de melhorias, apoia a seleção de repositórios que atendem a requisitos mínimos de qualidade e, conseqüentemente, favorece a descoberta, acesso e uso dos dados.

1.2.3 Objetivos da Pesquisa

Objetivo Geral (OG): Analisar a qualidade de dados publicados em portais governamentais brasileiros, com o propósito de identificar limitações e erros com respeito a descoberta, acesso e uso dos dados, do ponto de vista de pesquisadores e usuários, no contexto de programas sociais e transparência nos gastos de pessoal.

- **Objetivo Específico 01 (OE01):** Identificar os desafios enfrentados pelos governos para disponibilizar os dados abertos.
- **Objetivo Específico 02 (OE02):** Caracterizar erros e problemas frequentes encontrados em dados abertos governamentais em repositórios brasileiros.
- **Objetivo Específico 03 (OE03):** Avaliar os dados do repositório do Fundo de Financiamento ao Estudante do Ensino Superior (FIES) em relação ao grau de alinhamento com as dimensões de qualidade de dados.
- **Objetivo Específico 04 (OE04):** Analisar a efetividade de uma solução para a avaliação semiautomatizada de dados e metadados de repositórios de dados abertos governamentais.

1.3 QUESTÕES DE PESQUISA

As questões de pesquisa principal e específicas desta dissertação foram derivadas dos objetivos geral e específicos apresentados na Subseção anterior.

A Questão de Pesquisa (QP) principal desta dissertação é apresentada a seguir:

- **Questão de Pesquisa (QP)** *A estrutura de dados abertos governamentais implementados por portais específicos disponibilizados pelo Governo Brasileiro atende os requisitos das dimensões de qualidade de dados?*

A Questão de Pesquisa (QP) desta dissertação visa construir um panorama atualizado do que já foi feito e das principais oportunidades de melhorias identificadas nas iniciativas de dados abertos governamentais em relação a conformidade com as Dimensões de Qualidade de Dados. Com esta questão de pesquisa, pretende-se alcançar o Objetivo Geral (OG) desta dissertação.

A partir da Questão de Pesquisa (QP) principal desta dissertação, foram derivadas questões de pesquisa específicas, conforme apresentado a seguir:

- **Questão de Pesquisa Específica 1 - QPE1:** *Quais os desafios enfrentados pelos governos para adoção de dados abertos governamentais?*

A Questão de Pesquisa Específica 1 (QPE1) visa identificar os principais obstáculos enfrentados na implantação de dados abertos governamentais. Serão verificados os obstáculos e as tecnologias que contribuem para mitigá-los. QPE1 é utilizada nesta dissertação para alcançar o Objetivo Específico 01 (OE01).

- **Questão de Pesquisa Específica 2 - QPE2:** *Até que ponto os dados governamentais abertos do Programa de Financiamento Estudantil do Ensino Superior (FIES) estão em conformidade com as dimensões de qualidade de dados?*

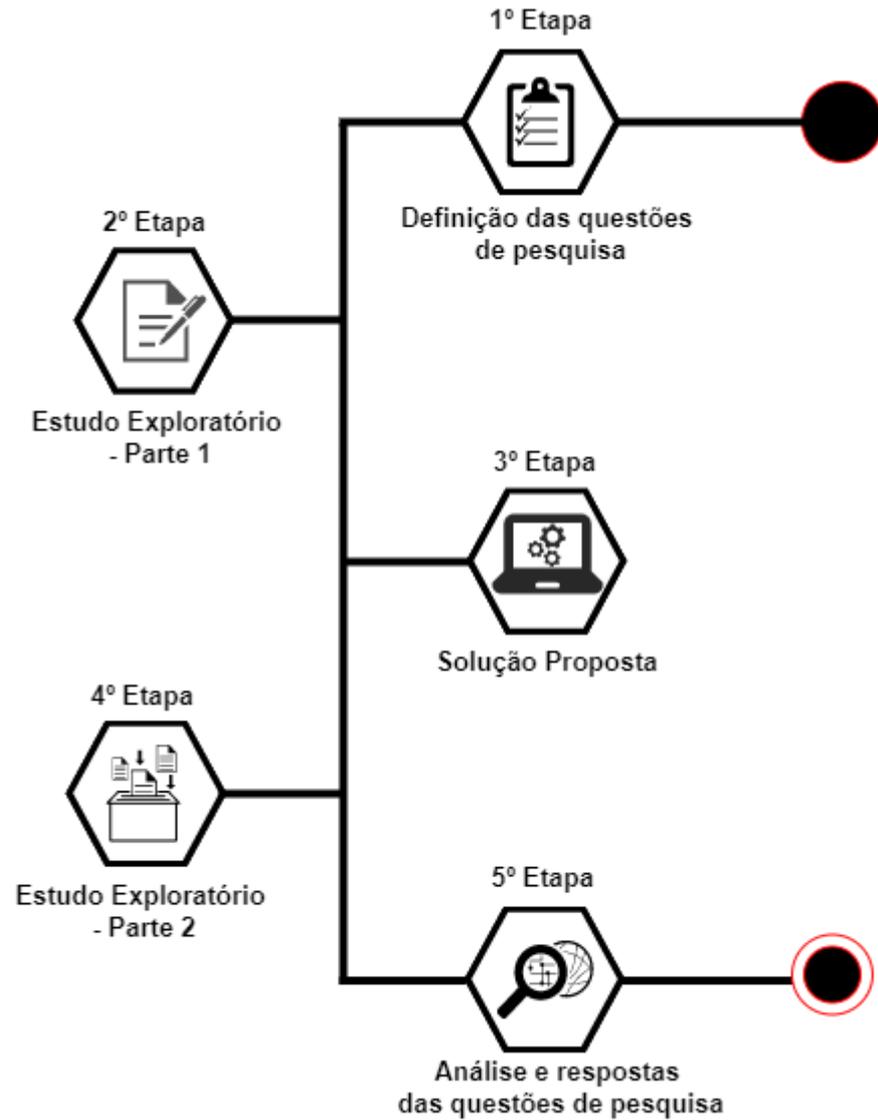
A Questão de Pesquisa Específica 2 (QPE2) visa caracterizar os erros e problemas nos dados abertos no contexto do Programa de Financiamento do Ensino Superior (FIES). QPE2 é utilizada nesta dissertação para alcançar os Objetivos Específicos 02 e 03 (OE02 e OE03).

- **Questão de Pesquisa Específica 3 - QPE3:** *Qual a efetividade de uma solução semiautomatizada para identificar limitações e erros no catálogo de dados do Portal Brasileiro de Dados abertos no contexto de programas sociais e gastos com pessoal em relação a descoberta, acesso e uso dos dados?*

A Questão de Pesquisa Específica 3 (QPE3) visa analisar a efetividade de uma solução semiautomatizada para apoiar a avaliação da qualidade de catálogos de metadados de repositórios de dados abertos governamentais com relação aos metadados que facilitam a descoberta, acesso e uso dos dados. QPE3 é utilizada nesta dissertação para alcançar o Objetivo Específico 04 (OE04).

1.4 ETAPAS DA PESQUISA

Figura 1 – Etapas da Pesquisa



Fonte: Autoria própria (2022).

1.4.1 Fase de Planejamento

A Figura 1 apresenta as etapas planejadas para esta dissertação de mestrado. A seguir tem-se a descrição de cada uma destas etapas.

- **Definição das Questões de Pesquisa:** Nesta etapa foram definidas as questões de pesquisa em alinhamento com o Objetivo Geral (OG) e os Objetivos Específicos

(OE). As Questões de Pesquisa são o ponto de partida para as próximas etapas e foram apresentadas neste Capítulo.

- **Estudo Exploratório (Parte 1):** O primeiro estudo exploratório focou na Questão de Pesquisa Específica 2 (QPE2). Neste estudo, foram considerados dados do repositório do Fundo de Financiamento ao Estudante do Ensino Superior (FIES). Este repositório foi avaliado em relação ao seu grau de alinhamento com as dimensões de qualidade de dados. O primeiro estudo exploratório é apresentado no Capítulo 3.
- **Solução Proposta: O Estudo Exploratório (Parte 1):** Apresentado no Capítulo 3 serviu de referência para a implementação de uma solução semiautomatizada para apoiar a avaliação de dados e metadados de repositórios de dados abertos em relação a facilidade de descoberta, acesso e uso dos dados, permitindo a escolha de repositórios segundo parâmetros previamente estabelecidos pelo usuário da solução. A Solução Proposta é apresentada no Capítulo 4.
- **Estudo Exploratório (Parte 2):** O segundo estudo exploratório avaliou a efetividade da solução proposta para avaliação de dados e metadados de repositórios de dados abertos em relação a facilidade de descoberta, acesso e uso dos dados. O segundo estudo exploratório é apresentado no Capítulo 5.
- **Respostas às Questões de Pesquisa:** Os resultados obtidos na fundamentação teórica e nas duas etapas do estudo exploratório serviram de referência para fundamentar e elaborar as respostas das questões de pesquisa específicas e também da questão de pesquisa principal desta dissertação. O Quadro 1 apresenta os capítulos que forneceram evidências (símbolo √) e nos quais foram efetivamente respondidas (símbolo ☒) as questões de pesquisa.

Quadro 1 – Capítulos com Evidências e Respostas às Questões de Pesquisa

Capítulos	QPE1	QPE2	QPE3	QP
2	√ ☒	√	√	√
3		√ ☒	√	√
4			√	√
5			√ ☒	√
6				☒

Fonte: Autoria própria (2022).

1.4.2 Artigo Produzido

Foi produzido o seguinte artigo resultante desta dissertação: *A Quality Dimension Analysis of Open Government Data in Undergraduate Public Funding in Brazil* de autoria de Marcelo Moreira West e Glauco Figueiredo Carneiro, publicado no ITNG 2022 - International Conference on Information Technology-New Generations (conferência internacional Qualis Capes A3) e disponível no portal da Springer Nature (ISBN: 978-3-030-97652-1) ¹.

1.5 ESTRUTURA DESTA DISSERTAÇÃO

Os demais Capítulos desta dissertação foram organizados da seguinte forma:

- **Fundamentação Teórica:** O Capítulo 2 apresenta os conceitos de dados abertos governamentais, as dimensões de qualidade de dados utilizadas no presente estudo, os desafios enfrentados por órgãos do governo brasileiro para disponibilizar os dados abertos, os principais frameworks utilizados por órgãos nacionais e internacionais para facilitar a descoberta, acesso e uso de dados abertos governamentais e os trabalhos relacionados.
- **Estudo Exploratório (Parte 1):** Os estudos exploratórios apresentados nesta dissertação estão divididos em duas partes. A Parte 1, apresentada no Capítulo 3, consiste em um estudo exploratório realizado de forma manual.
- **Solução Proposta:** O Capítulo 4 descreve a solução proposta para a avaliação semiautomatizada de metadados de repositórios de dados abertos governamentais.
- **Estudo Exploratório (Parte 2):** A Parte 2 do estudo exploratório descrito no Capítulo 5 relata a avaliação semiautomatizada da qualidade de metadados de repositórios de dados abertos governamentais.
- **Conclusão e Perspectivas Futuras:** Por fim, no Capítulo 6 são apresentadas as conclusões finais, assim como as contribuições e trabalhos futuros.

¹ https://link.springer.com/chapter/10.1007/978-3-030-97652-1_11

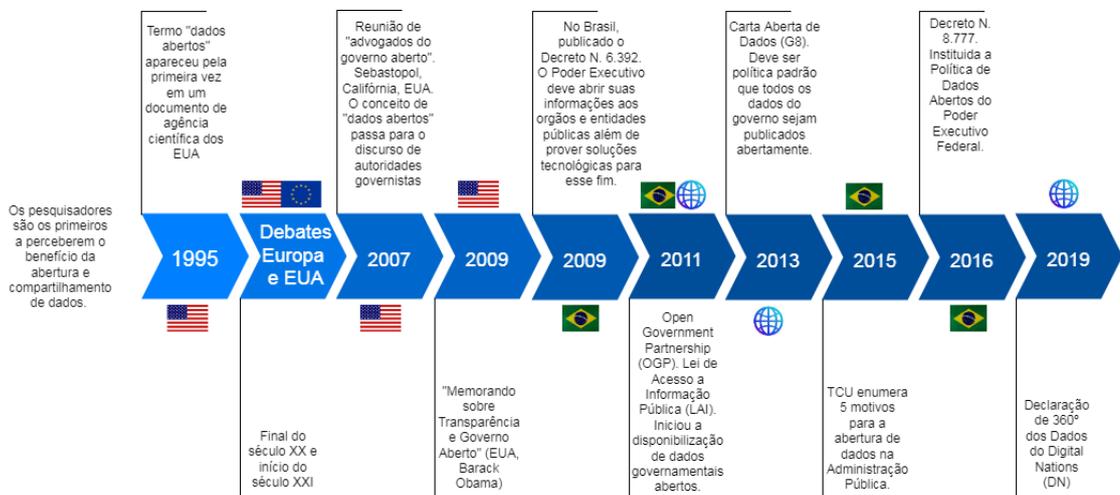
2 FUNDAMENTAÇÃO TEÓRICA

Este Capítulo apresenta os principais conceitos utilizados nos estudos exploratórios. São discutidos os conceitos gerais sobre Dados Abertos Governamentais, Dimensões de Qualidade de Dados, Desafios enfrentados pelos governos para a implementação de dados abertos e são apresentadas as plataformas disponíveis que facilitam a descoberta, acesso e uso de dados abertos.

2.1 DADOS ABERTOS GOVERNAMENTAIS

Tal como definido pelo World Wide Web Consortium (W3C) (2009), Dados Abertos Governamentais são dados públicos, produzidos, coletados ou mantidos por órgãos governamentais disponibilizados na web em formato aberto, estruturado, compreensível por máquina, com acesso livre para reutilização, modificação e redistribuição, exigindo no máximo a creditação da autoria e compartilhamento na mesma licença. A publicação, o acesso e o reuso dos dados abertos governamentais contribuem com o aumento da transparência, do controle social, da participação, da gestão e da colaboração governo-sociedade, além do fortalecimento da democracia e da capacidade estatal (POSSAMAI; SOUZA, 2020). A Figura 2 apresenta a linha do tempo da abertura dos dados governamentais, em destaque os momentos marcantes no Brasil e no exterior.

Figura 2 – Linha do Tempo da Abertura dos Dados Governamentais no Brasil e no Exterior



Fonte: Autoria própria (2022).

No Brasil, o Tribunal de Contas da União (TCU) elenca cinco motivos para a abertura de dados na Administração Pública: (1) Exigência da sociedade por mais transparência na gestão pública, (2) Permitir que sociedade contribua com os serviços inovadores aos cidadãos, (3) Aprimorar a qualidade dos dados governamentais, (4) Viabilizar novos negócios e (5) Atender as imposições da Lei (BRASIL, 2015). De acordo com o Artigo 8º da Lei nº 12.527 - Lei de Acesso à Informação (LAI) de 18 de novembro de 2011 (BRASIL, 2011), é obrigatório no Brasil, a publicação de informações de interesse coletivo ou geral em sítios oficiais dos órgãos e entidades públicas, onde deve-se permitir o acesso de forma automatizada por sistemas externos em formatos abertos; serem estruturados e legíveis por máquinas; possuírem documentação detalhada sobre os formatos utilizados na estruturação da informação; serem autênticos, íntegros e atualizados. Conforme o Decreto nº 8.777 de 11 de maio de 2016 (BRASIL, 2016), qualquer cidadão pode solicitar a abertura de bases de dados nos termos da LAI, desde que o dado não seja protegido pelas regras estabelecidas para o sigilo da informação. Ainda, segundo o decreto, os gestores são responsáveis por criar e atualizar o Plano de Dados Abertos (PDA), documento que orienta a implementação e a promoção da abertura de dados em cada órgão ou entidade da administração pública federal. O PDA estabelece padrões mínimos de qualidade que visem facilitar a manipulação e o reuso dos dados. Os dados abertos de diferentes instituições brasileiras são disponibilizados no Portal Brasileiro de Dados Abertos², criado para garantir a busca e o acesso de informações de relevância pública (SILVA; MONTEIRO; REIS, 2020).

2.2 DIMENSÕES DE QUALIDADE DE DADOS

Medir a qualidade dos dados permite classificar e identificar as informações que podem melhor atender aos interesses e necessidades dos usuários (DUTRA; BARBOSA, 2017). A interpretabilidade avalia se é possível interpretar corretamente o significado e as propriedades das fontes de dados através da documentação e dos metadados disponibilizados (BATINI; PERNICI, 2006). A compreensibilidade avalia até que ponto os dados são facilmente compreendidos. A credibilidade avalia o quanto os dados são considerados verdadeiros e confiáveis (PIPINO; LEE; WANG, 2002). A integridade avalia se estão faltando dados e se possuem amplitude e profundidade suficientes para o uso (VAZIRI; MOHSENZADEH, 2012). A acurácia avalia se os dados estão confiáveis e livres de erros (BATINI; PERNICI, 2006). A

² <https://dados.gov.br/>

representação concisa avalia até que ponto os dados são representados de forma compacta (VAZIRI; MOHSENZADEH, 2012). A consistência avalia se há compatibilidade entre os diferentes valores dos dados (SCANNAPIECO; CATARCI, 2002). A completude avalia o número de atributos preenchidos (REICHE; HÖFIG; SCHIEFERDECKER, 2014). A encontrabilidade mede o quanto os dados podem ser descobertos através dos metadados fornecidos e acessibilidade mede o quanto os dados podem ser recuperados através das URLs fornecidas (HASSENSTEIN; VANELLA, 2022).

De acordo com Oliveira, Rodrigues e Henriques (2004), boa parte dos dados apresentam erros como valores de atributos em falta, valores de atributos incorretos ou representações diferentes para os mesmos dados. Os valores de atributos em falta correspondem ao não preenchimento de atributos obrigatórios. Os valores incorretos consideram-se aqueles que não correspondem à realidade. As representações diferentes ocorrem quando os valores dos atributos possuem formatos variados. Os valores ilegais consideram-se aqueles fora do domínio de valores válidos. Para se alcançar conclusões acertadas é importante que os dados tenham um elevado nível de qualidade. Não há um consenso entre diversos autores, porém os conceitos multidimensionais da qualidade de dados têm como base, o entendimento intuitivo, a experiência industrial e a revisão da literatura (WAND; WANG, 1996). A seguir é apresentada a lista com os termos mais utilizados na literatura para nomear as dimensões de qualidade de dados:

- Acurácia
- Confiança
- Oportuna
- Relevância
- Completude
- Currência
- Importância
- Flexibilidade
- Precisão
- Formato
- Interpretabilidade
- Conteúdo
- Eficiência
- Imparcialidade
- Suficiência

- Usabilidade
- Utilidade
- Clareza
- Comparabilidade
- Concisão
- Informação
- Nível de Detalhe
- Quantidade
- Escopo
- Entendimento
- Consistência

2.3 DESAFIOS ENFRENTADOS PELOS GOVERNOS PARA ADOÇÃO DE DADOS ABERTOS GOVERNAMENTAIS

Nos últimos anos, governos do mundo inteiro têm buscado abrir os dados governamentais. No Brasil, desde a implementação da Lei nº 12.527 de 18 de novembro de 2011 – LAI, diversos órgãos públicos das esferas municipal, estadual e federal enfrentam desafios de ordem técnica e tecnológica e ainda o aumento das despesas com equipamentos e recursos humanos (BRASIL, 2011).

De acordo com Denis e Goëta (2014), antes da abertura dos dados, os governos devem proceder com os seguintes passos: 1º descobrir quais as bases de dados possuem, 2º extrair as bases dos sistemas legados e, por último, tratar as bases para torná-las genéricas o bastante para serem utilizáveis por qualquer pessoa. O primeiro passo foi nomeado pelos autores como exploração, o segundo extração e o último *'rawification'*.

Por outro lado, Almada Silva e Dominguez (2017), destacam que existem muitos desafios no âmbito da administração pública para a implementação dos dados abertos. Como por exemplo, o estilo da gestão pública em relação às rotinas organizacionais, tecnologias utilizadas e normas impostas pelas leis vigentes. Listam ainda os seguintes desafios: do ponto de vista infraestrutural e técnico, alguns órgãos não conseguem tornar os seus sistemas de informação mais dinâmicos e passíveis de intercâmbio com os meios de comunicação. Quanto ao fator cognitivo, há desafios vinculados às competências dos gestores, técnicos, desenvolvedores e tomadores de decisão de escolha das tecnologias a serem empregadas. Por fim, no âmbito institucional, ainda há uma cultura do sigilo em órgãos públicos, o que dificulta

a abertura dos dados. O problema encontra-se na cultura organizacional, pois existem órgãos que estão estruturados de natureza hierárquica e burocrática em excesso (PASQUIER; VILLENEUVE, 2007).

2.4 PRINCIPAIS FRAMEWORKS UTILIZADOS PARA FACILITAR A DESCOBERTA, ACESSO E USO DE DADOS ABERTOS GOVERNAMENTAIS

Os dados tornaram-se fundamentais para as organizações bem como para a ciência. Nesta conjuntura, são realizadas constantemente discussões sobre dados abertos. Especialmente discutem-se as políticas e recursos utilizados para a abertura das informações (CONEGLIAN; SANTAREM SEGUNDO, 2016). A organização Open Knowledge Foundation³ ocupa significativo papel nestas discussões e encabeça o desenvolvimento e a manutenção do framework Comprehensive Knowledge Archive Network (CKAN).

2.4.1 Framework CKAN

O CKAN é uma plataforma open source utilizada para criação de portais de dados abertos. São funcionalidades básicas da plataforma CKAN: apoiar a publicação, o compartilhamento, a descoberta e o uso dos dados. O objetivo do CKAN é facilitar a criação, o acesso e a disseminação de conhecimento (WINN *et al.*, 2013). O CKAN⁴ disponibiliza os dados em *datasets* (conjunto de dados), onde cada um armazena informações de um determinado contexto, como por exemplo, as informações governamentais de projetos sociais.

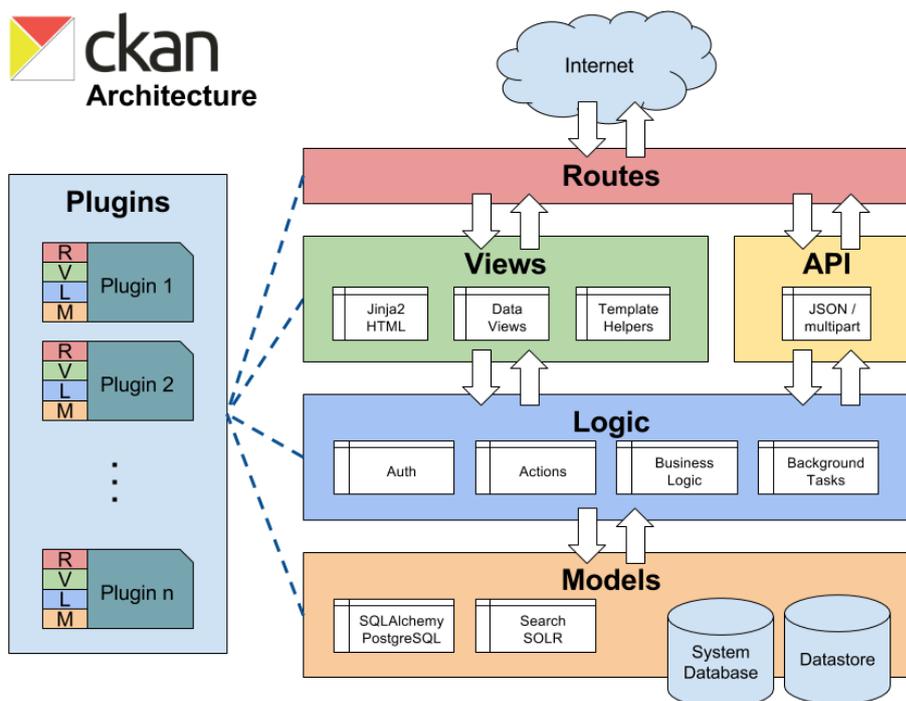
As informações contidas nos *datasets* são os metadados que descrevem as informações dos dados (por exemplo, data de criação, autor, título, formato entre outros) e os conjuntos de recursos (que são propriamente os dados ou links de conteúdo da web).

A Figura 3 apresenta a arquitetura do framework CKAN.

³ <https://ok.org.br/sobre/>

⁴ <http://docs.ckan.org/en/latest/>

Figura 3 – Arquitetura do Framework CKAN



Fonte: CKAN code architecture⁵

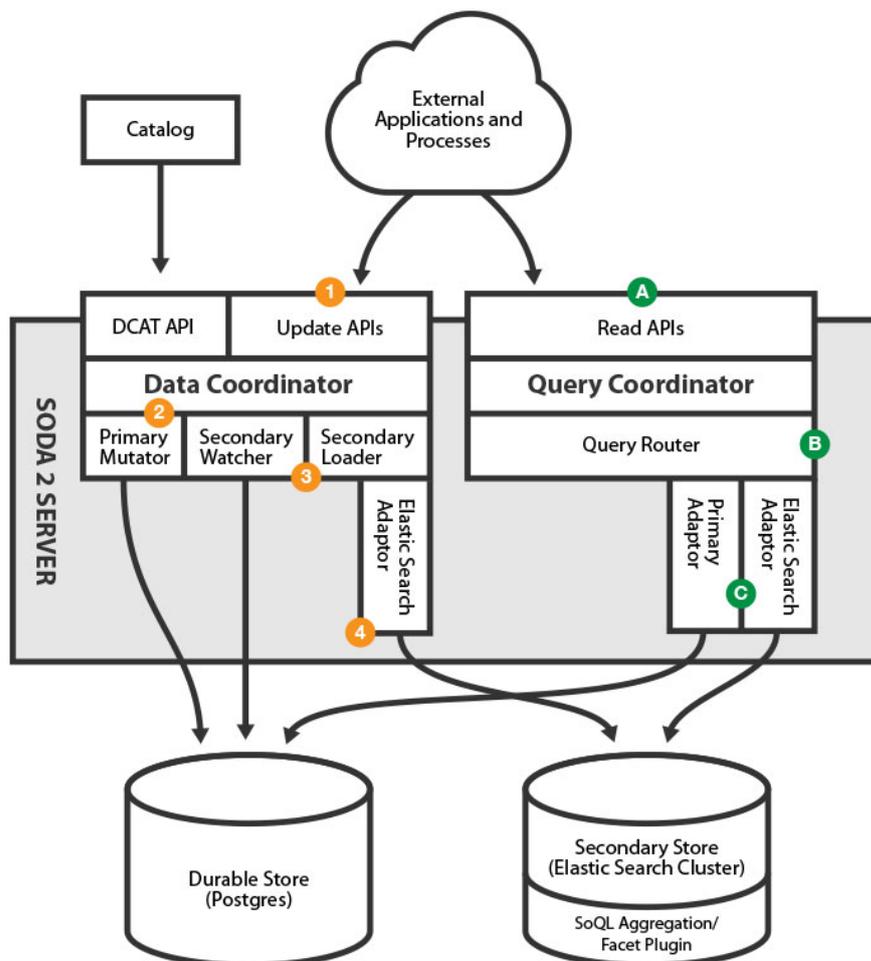
2.4.2 Framework SOCRATA

O Socrata⁶ é uma plataforma baseada em nuvem e na ideia de software como serviço, seu principal diferencial é na construção de visualizações mais elaboradas, com formatação condicional, gráficos e mapas. É utilizado no portal oficial do Kenya, no portal do Banco Mundial, em cidades e estados dos EUA como Chicago, Nova York, Austin, Maryland e Illinois (OLIVEIRA; LÓSCIO, 2014). Diversas tecnologias estão presentes no Socrata, como o *Javascript* e o *Ruby* para o *frontend*, *Java* e *Scala* para o *backend* e tecnologias de armazenamento como o *Postgres* e *Cassandra*. A arquitetura do framework Socrata é apresentada na Figura 4, que demonstra que os caminhos de leitura e escrita dos dados são divididos. Conforme a documentação do Socrata, tem como objetivo fornecer acesso mais rápido para quem apenas utiliza os dados.

⁵ <https://docs.ckan.org/en/2.9/contributing/architecture.html>

⁶ <https://www.socrata.com/products/open-data/>

Figura 4 – Arquitetura do Framework Socrata



Fonte: Socrata open data server⁷ (2022).

Os conjuntos de metadados dos repositórios de dados abertos governamentais podem apoiar a avaliação da qualidade dos metadados e dados. Muitos são os relatos encontrados na literatura quanto os desafios enfrentados na avaliação da qualidade de metadados e dados de portais de dados abertos governamentais (KUBLER *et al.*, 2018).

A avaliação da qualidade de metadados e dados não é uma tarefa tão fácil, especialmente pela falta de padronização (ZUIDERWIJK; JANSSEN, 2014).

⁷ <http://open-source.socrata.com/architecture/>

2.5 TRABALHOS RELACIONADOS

As referências de trabalhos selecionados para esta Subseção abordam o tema de descoberta, acesso e uso de dados abertos.

O primeiro trabalho intitulado: *Dadosjusbr.online: uma ferramenta para democratização das informações de pagamentos para magistrados* (AGOSTINI *et al.*, 2019), descreve uma solução que facilita a análise e o processamento de dados abertos governamentais. Seu foco é a formatação e a reorganização das informações por meio de mecanismos de extração e transformação. Nesta abordagem, a solução disponibiliza os dados no formato *Tabular Datapackage* composto por dois arquivos. O primeiro arquivo contém os dados tabulares no formato *comma-separated-values* (CSV) e o segundo arquivo denominado *'datapackage.json'* contém os metadados no formato *JavaScript Object Notation* (JSON).

O segundo trabalho intitulado: *Estudo de caso “operação serenata de amor”: análises de big data no combate à festa dos gastos público* (PINTO FILHO; FONTES, 2018), apresenta um projeto que utiliza a mineração e a interpretação de dados abertos governamentais. O cerne é a captura e análise de dados através de algoritmos. Nesta abordagem, a ferramenta conta com *datasets* e rotinas para extração de dados em arquivos nos formatos *comma-separated-values* (CSV), *Extensible Markup Language* (XML) e imagens nas extensões jpeg, jpg, png, bmp, gif, tif, psd, pdf, svg e outros. Para extração dos dados das imagens é utilizado o *Optical Character Recognition* (OCR) e consultas pela *API Cloud Vision da Google*. Também faz parte do projeto o 'robô' Rosie, desenvolvido em linguagem de programação *Python* incluindo a biblioteca *scikit-learn* e modelos de aprendizagem supervisionado (classificação) que aprende a partir de exemplos rotulados das saídas de um teste (variável de resposta) que devem ser produzidos por uma entrada (recurso) e ainda utiliza o modelo *k-means* de aprendizagem não supervisionado, que possuem dados não rotulados, ou seja, o algoritmo deve tentar entender por conta própria.

O terceiro trabalho intitulado: *Dados abertos governamentais no contexto de políticas públicas de saúde e sistemas prisionais: realidade ou utopia?* (CARDOSO; CARNEIRO; MENEZES, 2020), detalha uma ferramenta de avaliação semiautomatizada da qualidade de dados e metadados de repositórios de dados abertos nacionais e internacionais. O propósito é medir o grau de alinhamento dos dados com os princípios de dados abertos. Nesta perspectiva, a ferramenta é apoiada pelas APIs CKAN e Socrata na automação dos processos de consulta, extração e análise dos dados e metadados. Além disso, são armazenados metadados e respostas de um conjunto de questões mapeadas sobre os princípios *Open Government Data* (OGD) no *ElasticSearch*, mecanismo de busca e análise de dados distribuído, gratuito e aberto para todos

os tipos de dados, incluindo textuais, numéricos, geoespaciais, estruturados e não estruturados. O processo inicia-se com a busca dos metadados em catálogos de dados abertos governamentais. As APIs retornam os resultados das buscas no formato *JavaScript Object Notation* (JSON) contendo os metadados e algoritmos avaliam a qualidade dos mesmos.

Ao compararmos os trabalhos relacionados, alguns pontos foram levantados. No primeiro, apesar de ser utilizada uma ferramenta para a análise e processamento, o desafio desta abordagem concentra-se no tempo gasto para a avaliação dos dados contidos em um arquivo único e a falta de padronização. No segundo, mesmo sendo aplicados algoritmos de aprendizagem de máquina, os dados não rotulados comprometem os resultados. Os dois primeiros trabalhos adotam abordagens que não utilizam APIs reconhecidas internacionalmente como a CKAN e Socrata para a publicação dos dados, ao comparar as soluções propostas destas abordagens com a proposta do terceiro trabalho, vimos que é viável utilizar essas APIs para a análise semiautomatizada dos dados e metadados em portais de dados abertos. Para comprovar a viabilidade do uso da API CKAN, foi realizado esse estudo exploratório e foram obtidos resultados que demonstraram que a API CKAN colabora com a redução do tempo de análise entre outros benefícios de utilizá-la.

2.6 O CASO DA ABERTURA DOS DADOS DA UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE (UFRN)

Com a promulgação da Lei de Acesso à Informação órgãos públicos nas esferas, municipal, estadual e da união se mobilizaram para abertura dos seus dados. Entre os órgãos que aderiram a abertura dos dados em meados de 2016, selecionamos o Caso da Universidade Federal do Rio Grande do Norte (UFRN) que deu o pontapé inicial da implementação da abertura dos dados, com elaboração do Plano de Dados Aberto (PDA).

O PDA da UFRN elencou a ordem de prioridade para disponibilização dos seus conjuntos de dados, tendo como critério o grau de relevância das informações para o cidadão, o alinhamento com os instrumentos de planejamento estratégico e a disponibilidade dos dados gerenciados pelos Sistemas Integrados de Gestão da UFRN (SIGs) (AQUINO JÚNIOR *et al.*, 2019).

A UFRN adotou como estratégia de dados abertos as sugestões de fases e tarefas do Processo para Dados Abertos (ProceDA), que foi criado a partir de um profundo estudo da literatura sobre a disponibilização de dados abertos por instituições públicas, o qual incluiu materiais de vários países, o que gerou uma lista de tarefas, boas práticas e definiu as fases da

publicação de dados, como consequência foram adaptados à realidade brasileira (AQUINO JÚNIOR *et al.*, 2019).

O Quadro 2 apresenta as fases da implantação do ProceDA na Universidade Federal do Rio Grande do Norte.

Quadro 2 – Fases do ProceDA

01	Estabelecimento de Equipe	estabelecimento de papéis e definição da equipe que irá atuar no projeto de dados abertos da instituição
02	Diagnóstico e Planejamento	avaliação geral da situação da instituição, através da medição de seu nível de preparo para disponibilização de dados e da realização de estudos de viabilidade do projeto; estabelecimento de um Plano de Dados Abertos; determinação de uma política de disponibilização de dados; estudo de legislação e outras questões gerais de projeto
03	Seleção de Dados	nesta etapa, ocorre a seleção dos conjuntos de dados que serão publicados no ciclo em questão. Deve-se observar uma série de questões para o estabelecimento de prioridades, como o valor associado aos conjuntos de dados, facilidade de publicação, benefícios e riscos da disponibilização de determinados dados, entre outros
04	Modelagem de Dados	nesta fase, ocorre a preparação dos dados para sua posterior publicação. A preparação inclui a limpeza desses dados, elaboração de metadados, tratamento de informações confidenciais e a escolha do melhor formato de publicação
05	Definição de Licença	deve-se definir os termos de uso dos dados, que devem ser, resumidamente, de formato aberto, não proprietários, estáveis e de amplo uso
06	Publicação de Dados	esta fase é composta por tarefas relacionadas à publicação, de fato, dos dados na Web
07	Promoção	são definidas estratégias de promoção dos dados que foram publicados, aumentando a chance de reuso pela comunidade
08	Manutenção e Otimização	avaliação da necessidade de atualizar conjuntos de dados previamente publicados, além da realização de monitoramento e ajuste das atividades realizadas durante todo o processo, buscando encontrar possíveis pontos de melhoria

Fonte: Adaptado de Aquino Júnior *et al.* (2019).

Para atender o oitavo item da lista, a UFRN adotou o framework CKAN para a publicação dos dados, o que foi motivado pela expressiva adoção por outros órgãos nacionais e internacionais. Desde a divulgação do Portal de Dados da UFRN, aumentou o número de acessos às suas páginas. Os *datasets* mais acessados são os que possuem informações de discentes, cursos, turmas, matrículas em componentes, docentes, servidores e gastos. Esses e outros dados catalogados no Portal são frequentemente utilizados em atividades de disciplinas e trabalhos de fim de cursos dos níveis de graduação e de pós-graduação da Instituição, como também em seus projetos de ensino, pesquisa e extensão. Desta forma, aumentou a

transparência e aproximou os cidadãos da instituição bem como cumpre-se o que determina a LAI (AQUINO JÚNIOR *et al.*, 2019).

2.7 RESPOSTA DA QUESTÃO DE PESQUISA ESPECÍFICA 1 (QPE1)

Questão de Pesquisa Específica 1 - QPE1: *Quais os desafios enfrentados pelos governos para adoção de dados abertos governamentais?*

Resposta: Conforme pode ser verificado nas argumentações apresentadas nas Subseções 2.3 e 2.6, os governos enfrentam inúmeras dificuldades para implantação de dados abertos governamentais, especialmente no que tange a recursos tecnológicos e mudança de cultura.

3 ESTUDO EXPLORATÓRIO - PARTE 1

3.1 INTRODUÇÃO

Este estudo exploratório está dividido em duas partes: a primeira parte apresentada neste Capítulo constitui um estudo exploratório realizado de forma manual para identificar erros e problemas nos dados do FIES publicados no Portal Brasileiro de Dados Abertos. Esta caracterização consistiu na avaliação do grau de alinhamento dos dados analisados com as Dimensões de Qualidade de Dados. A segunda parte do estudo exploratório será apresentada no Capítulo 5.

Dados governamentais abertos têm atraído cada vez mais atenção dos cidadãos devido à sua relevância para a compreensão de vários aspectos da vida nas cidades. Órgãos governamentais em todo o mundo fornecem dados que são do principal interesse dos cidadãos. A literatura recomenda a adoção de melhores práticas para o fornecimento de dados abertos aos cidadãos de forma a atingir os seguintes objetivos: promover a transparência, incentivar a governança participativa e identificar a responsabilização do governo nas tomadas de decisões. Fornecer dados governamentais abertos sem considerar dimensões de qualidade relevantes é uma restrição significativa para melhorar seu uso prático. Dados incompletos, imprecisos, inconsistentes e duplicados são apenas exemplos de não conformidades neste cenário.

Estudos têm descrito o aumento do uso de portais de dados governamentais abertos em diferentes regiões do mundo. Por exemplo, houve um aumento representativo nos conjuntos de dados publicados no portal do governo australiano na ordem de 900% entre 2013 e 2015. No mesmo período, o portal do governo da Nova Zelândia atingiu a marca de 3.800 conjuntos de dados, Reino Unido 23.000, Estados Unidos 194.000 e Canadá 240.000 (SADIQ; INDULSKA, 2017).

Dados governamentais abertos podem contribuir para o desenvolvimento de cidadãos conscientes e ativos (SILVA; MONTEIRO; REIS, 2020). No entanto, a falta de qualidade dos dados é um desafio para atingir esse objetivo. Por exemplo, há um desafio em lidar com a não padronização da formatação dos dados, e a não conformidade na disponibilização de metadados (DING; PERISTERAS; HAUSENBLAS, 2012).

Esta parte do estudo exploratório tem como objetivo analisar de forma manual a qualidade dos dados para identificar limitações e erros quanto ao uso de dados governamentais abertos por pesquisadores e usuários no contexto do financiamento público de graduação em instituições brasileiras de ensino. Esta parte do estudo exploratório está estruturada da seguinte

forma: a seção 3.1 apresenta uma breve introdução, a seção 3.2 apresenta aspectos importantes da análise das dimensões de qualidade de dados e tipos de erros identificados em dados abertos governamentais, a seção 3.3 apresenta os procedimentos metodológicos referentes ao planejamento e execução do estudo, a seção 3.4 apresenta os resultados e as discussões dos mesmos, a seção 3.5 apresenta a resposta da QPE2 e a seção 3.6 apresenta a conclusão desta parte do estudo exploratório.

3.2 ANÁLISE DAS DIMENSÕES DE QUALIDADE DE DADOS

A análise das dimensões de qualidade de dados possibilita classificar e identificar os dados para atingir os objetivos dos usuários (DUTRA; BARBOSA, 2017). As dimensões de qualidade completude, interpretabilidade, compreensibilidade, credibilidade, acurácia e consistências avaliadas nesta parte do estudo exploratório contribuem para a seleção de um conjunto de dados com maior nível de qualidade. Os desafios para lidar com dados governamentais abertos influenciam o processo de tomada de decisão devido a não conformidades dos dados com as dimensões de qualidade.

Conforme Oliveira, Rodrigues e Henriques (2004) muitos dos dados governamentais abertos apresentam erros como valores de atributos ausentes, valores de atributos incorretos ou representações diferentes no mesmo conjunto de dados. Os valores de atributo ausentes correspondem ao não preenchimento dos atributos obrigatórios. Os valores incorretos são considerados aqueles que não correspondem à situação. As diferentes representações ocorrem quando os valores dos atributos possuem formatos diferentes. Os valores ilegais são considerados aqueles fora do domínio de valores válidos.

3.3 METODOLOGIA DA PRIMEIRA PARTE DO ESTUDO EXPLORATÓRIO

Esta subseção descreve as ferramentas adotadas na condução da primeira parte do estudo exploratório. Em seguida, são apresentadas as etapas de execução. O Quadro 3 apresenta o objetivo do estudo de acordo com a abordagem Goal Question Metric (GQM) (BASILI; ROMBACH, 1988).

Quadro 3 – Objetivo do Estudo de Acordo com a Abordagem GQM

<i>Analisar</i>	a qualidade de dados
<i>com o propósito de</i>	identificar limitações e erros
<i>em relação ao</i>	uso de dados abertos governamentais
<i>do ponto de vista de</i>	pesquisadores e usuários
<i>no contexto do</i>	Financiamento Estudantil Brasileiro (FIES)

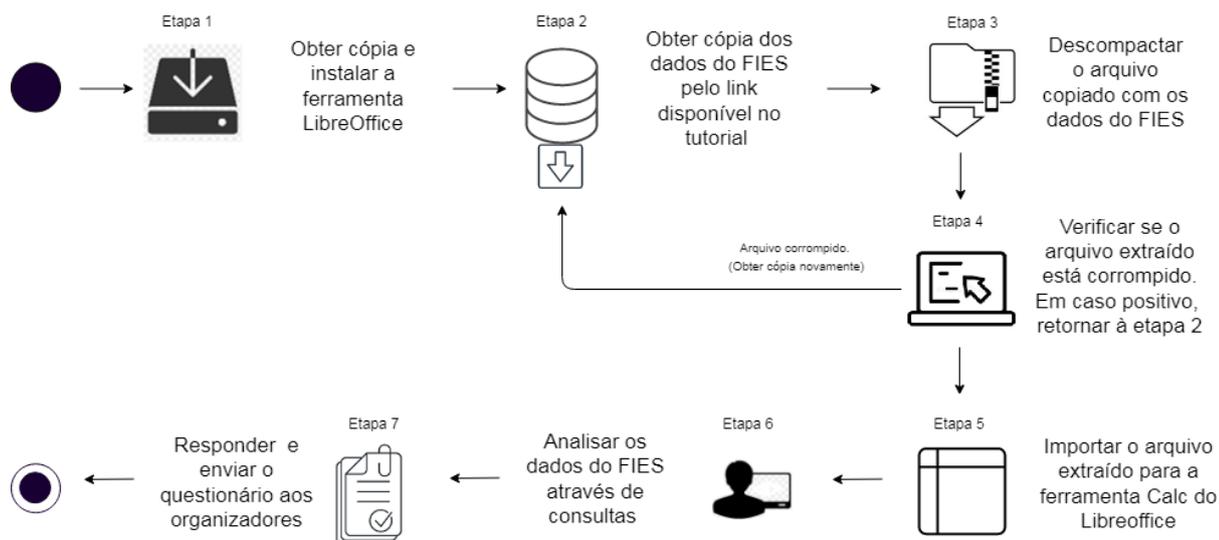
Fonte: Autoria própria (2022).

Para realizar as atividades, os participantes utilizaram o Google Meet, um serviço que permite a interação individual ou coletiva por meio de videoconferências (SANT'ANNA; SANT'ANNA 2020). Eles também usaram o *Google Classroom*, um componente do pacote de aplicativos *Google for Education* composto pelo gerenciador de e-mail (gmail), repositório de arquivos (drive) e editor de texto, planilhas e apresentações (docs) (DINIZ *et al.*, 2018). Os participantes analisaram os dados fornecidos pelo Fundo de Financiamento Estudantil (FIES), disponível no Portal Brasileiro de Dados Abertos no formato "CSV". O motivo de usar os dados do segundo semestre de 2019 foi por serem os dados mais recentes disponíveis no portal.

3.3.1 Planejamento

A análise das dimensões de qualidade de dados do FIES seguiu as sete etapas apresentadas na Figura 5. A primeira etapa consistiu na instalação da ferramenta *Libreoffice*. A segunda etapa correspondeu ao download do arquivo de dados abertos do FIES em formato compactado disponível na plataforma *Google Classroom*. Na terceira etapa, os participantes descompactaram o arquivo e, na quarta etapa, validaram o conteúdo do arquivo descompactado, comparando seu tamanho com o tamanho do arquivo disponível no Portal Brasileiro de Dados Abertos. Na quinta etapa, os participantes abriram o arquivo descompactado na ferramenta *Calc* no *Libreoffice*. Na sexta etapa os participantes realizaram a análise dos dados e na sétima etapa preencheram um questionário. O Quadro 4 apresenta as etapas do tutorial.

Figura 5 – Etapas do Tutorial



Fonte: Autoria própria (2022).

Quadro 4 – Etapas do Tutorial Realizadas pelos Participantes

Etapa	Descrição
1	Instalação da ferramenta Libreoffice
2	Download do arquivo de dados abertos do FIES em formato compactado
3	Descompactação do arquivo de dados obtido na etapa anterior
4	Validação do conteúdo do arquivo descompactado
5	Importação do arquivo de dados do FIES para a ferramenta Libreoffice Calc
6	Análise dos dados do FIES de acordo com as dimensões de qualidade
7	Preenchimento do questionário

Fonte: Autoria própria (2022).

Seleção de Participantes - Parte I: Selecionamos os participantes com base nos seguintes critérios: não possuir formação na área de Tecnologia da Informação, possuir habilidades no uso de computadores, softwares, acesso à internet, possuir conta de e-mail na plataforma google e utilizar a plataforma operacional sistema (*Windows ou Linux*).

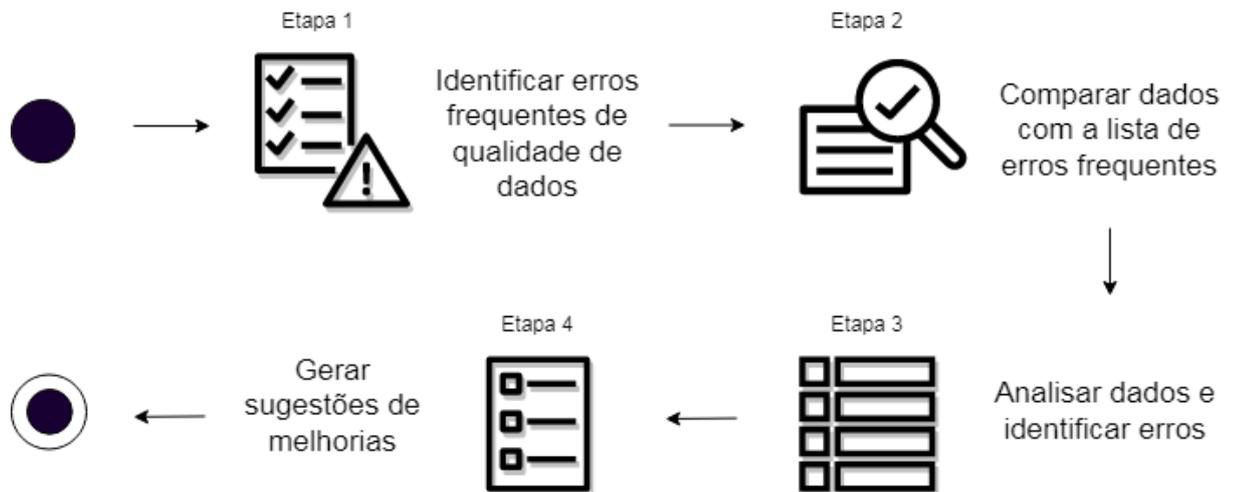
Seleção de Dimensão: Os critérios para seleção das dimensões de qualidade de dados foram: dimensões relacionadas ao uso de dados e tomada de decisão. Foram escolhidas as seguintes dimensões: representação concisa, interpretabilidade, compreensibilidade, credibilidade, integridade, completude e acurácia.

3.3.2 Execução

A Figura 6 apresenta as etapas realizadas pelos participantes. Os participantes executaram a primeira parte do estudo exploratório em duas seções. Os organizadores apresentaram as atividades propostas para baixar, instalar e configurar o ambiente na primeira seção. Na sequência, os participantes realizaram as etapas. Os participantes puderam fazer perguntas aos organizadores sobre as atividades.

Os participantes realizaram sete etapas para analisar as dimensões da qualidade dos dados: Na primeira, obtiveram uma cópia e instalaram a ferramenta *Libreoffice*. Na segunda, obtiveram uma cópia dos dados abertos do FIES em um arquivo compactado. Na terceira, eles descompactaram o arquivo copiado na etapa anterior. Na quarta, validaram o arquivo extraído.

Figura 6 – Etapas do Tutorial para Análise de Dados



Fonte: Autoria própria (2022).

As seguintes ferramentas foram utilizadas pelos participantes: sistema operacional Windows 10, navegador Google Chrome Versão 95.0.4638.54 64 bits e editor de planilhas LibreOffice Community Versão 7.2.0.4 (x86), WhatsApp Versão 4.1., descompactador de arquivo Winrar Versão 6.0 64 bits e gerenciador de correio.

3.4 RESULTADOS E DISCUSSÃO DE RESULTADOS DA PRIMEIRA PARTE DO ESTUDO EXPLORATÓRIO

Esta seção discute os resultados da análise da qualidade dos dados do FIES e também identifica oportunidades para melhorias adicionais.

A análise da qualidade dos dados do FIES revelou não conformidades conforme apresentado no Apêndice A.

A Figura 6 apresenta as quatro etapas da análise. A primeira etapa consistiu em revisar a literatura para identificar erros frequentes representados no Quadro 5. A segunda etapa teve como objetivo identificar os erros relatados no Quadro 5. O Apêndice A mostra o número de erros encontrados nos atributos dos dados do FIES analisando as respostas dos participantes. As evidências encontradas foram atributos preenchidos com erros ortográficos, preenchidos com valores ilegais ou não preenchidos. Na terceira etapa, os participantes registraram os erros detectados. Em contrapartida, a quarta etapa teve como objetivo propor sugestões para melhorar a qualidade dos dados representados no Quadro 7.

Quadro 5 – Erros de Qualidade Frequentes em Dados Abertos

(continua)

Id do Erro	Tipo do Erro	Nome do Erro	Descrição do Erro	Exemplo do Erro	Dimensão da Qualidade Associada
E1	Valores Incorretos	Dados usados para um fim diferente do pretendido	Atributo email preenchido com o nome	e-mail="Maria da Silva Souza"	Consistência
E2	Valores Incorretos	Dados do tipo texto preenchidos com tipo numérico	Atributo nome preenchido com o número do CPF	nome="999.999.999-99"	Consistência
E3	Valores Incorretos	Dados do tipo numérico preenchido com tipo texto	Atributo telefone preenchido com nome	telefone="Maria da Silva Souza"	Consistência
E4	Valores Incorretos	Dados do tipo data preenchidos com data maior que maior ou igual a 100 anos	Atributo data do nascimento preenchido com data do próximo século	datanascimento = 2121 - 01 - 01	Acurácia

Quadro 5 – Erros de Qualidade Frequentes em Dados Abertos

(conclusão)

E5	Valores Incorretos	Dados preenchidos com erros ortográficos	Atributo município preenchido sem acento, mas deve ser acentuado	município="Sao Paulo"	Acurácia
E6	Valores Incorretos	Dados preenchidos com mais de um atributo	Atributo nome preenchido com nome e número do processo	nome="Maria da Silva Souza 99393346-20"	Representação Concisa
E7	Atributos com valores ausentes	Dados não preenchidos	Atributo nome da empresa em branco	nome_empresa=	Compleitude
E8	Valores Incorretos	Dados preenchidos com caracteres indecifráveis	Atributo município preenchido com símbolos especiais da tabela ASCII	município="S o Paulo"	Interpretabilidade, Compreensibilidade e Credibilidade
E9	Valores Ilegais	Dados preenchidos com valores ilegais	Atributo valor mensal preenchido com valor negativo	mensalidade=-820,41	Consistência e Credibilidade

Fonte: Autoria própria (2022).

A análise dos dados e a identificação de erros consideraram as seguintes dimensões da qualidade dos dados: interpretabilidade, compreensibilidade, representação concisa, credibilidade, integridade e acurácia. É fundamental destacar que os dados publicados em portais governamentais abertos devem atender aos princípios das dimensões selecionadas, pois tratam de que os usuários devem interpretar os dados com facilidade. Além disso, dados completos, livres de erros e confiáveis contribuem para um processo de tomada de decisão eficaz.

Os organizadores analisaram as respostas fornecidas pelos participantes durante a identificação de erros nos dados. Os participantes responderam ao questionário fornecido no Apêndice B apoiado nas informações fornecidas pelo Anexo A. O Quadro 6 apresenta a lista de erros encontrados pelos participantes. Foram identificados os seguintes erros:

Valor de atributo ausente: Foram encontrados nos atributos Abreviatura da Unidade da Federação, Nome do Município, Número do Cadastro Nacional da Pessoa Jurídica da Mantenedora, Nome da Razão Social da Mantenedora, Código do Município da Mantenedora e Nome do Município dado em branco, o que demonstra o descumprimento da dimensão completude, de acordo com as questões 1 e 18.

Valores incorretos: Nos atributos Tipo de Fiança Contratada, Código do Município do Patrocinador, Código do Município da Instituição de Ensino Superior e Código do Município da Residência Estudantil, os dados foram representados em caracteres indecifráveis e os atributos Razão Social da Mantenedora, Nome da Instituição de Ensino Superior (NO_IES), Nome dos Dados da Instituição de Ensino Superior (NO_IES_EXT_ALUNOS) com acento incorreto, o que demonstra não conformidade com as dimensões: interpretabilidade, compreensibilidade, credibilidade e acurácia, conforme as questões 14 e 20.

Valores Ilegais: Foram encontrados valores inválidos no atributo Valor Mensal, o que demonstra não conformidade com as dimensões de qualidade de dados consistência e credibilidade.

Foram encontrados os seguintes atributos sem descrição no dicionário de dados: Código do Tipo de Curso, Descrição do Tipo de Curso, Código da Frequência do Curso, Descrição da Frequência do Curso, Ano do Processo, Mês do Processo e Sigla do Agente Financeiro.

Os participantes acharam difícil compreender e interpretar o atributo referente ao ano do comprovante de pagamento e não conseguiram identificar se o aluno é indígena por meio do atributo Descrição Raça/Cor, conforme questões 2 e 6.

Quadro 6 – Distribuição da Quantidade de Erros por Tipo de Erro

Tipo de erro	Número de erros	Frequência de erros n(%)
Valores incorretos	385949	99,297
Valores de atributos ausentes	2697	0,694
Valores ilegais	36	0,009
Total	388682	100

Fonte: Autoria própria (2022).

Com base nos resultados obtidos neste estudo, foram apresentadas as sugestões de melhorias na Quadro 7. A primeira foca em evitar registros de dados em branco. A segunda discute os benefícios da prevenção de dados indecifráveis e a terceira aponta as vantagens de utilizar dados para facilitar a compreensão e interpretação do conteúdo.

Quadro 7 – Sugestões de Melhorias

Nº	Sugestões
1	Tornar os atributos encontrados em branco obrigatórios
2	Utilizar a configuração de caracteres "Europa Ocidental (Windows 1252/Latin1)"
3	Incluir descrições de atributos ausentes no dicionário de dados

Fonte: Autoria própria (2022).

3.5 RESPOSTA DA QUESTÃO DE PESQUISA ESPECÍFICA QPE2

Questão de Pesquisa Específica - QPE2: *Até que ponto os dados governamentais abertos do Programa de Financiamento Estudantil do Ensino Superior (FIES) estão em conformidade com as dimensões de qualidade de dados?*

Resposta: A Subseção 3.4, em especial o Quadro 6 e o Apêndice A, apresentam em detalhes a distribuição de erros identificados nos dados abertos do repositório FIES e a quantidade de erros por atributos, quando comparados com as dimensões de qualidade de dados. Conforme já mencionado anteriormente neste capítulo, verifica-se que os repositórios nacionais avaliados ainda apresentam oportunidades de melhorias significativas, apontando necessidade de investir na qualidade dos dados para atender os requisitos das dimensões de qualidade de dados.

3.6 CONCLUSÃO

No início da execução do estudo exploratório havia dúvidas sobre as limitações enfrentadas pelos usuários de dados publicados em portais abertos, por isso, tornou-se imprescindível estudar a análise das dimensões da qualidade dos dados para conhecer essas limitações e saber como superá-las. Portanto, o estudo exploratório teve como objetivo analisar a qualidade dos dados para identificar limitações e erros no uso de dados governamentais abertos na perspectiva de pesquisadores e usuários no contexto do FIES. Verifica-se que o objetivo geral foi cumprido, pois com a realização do tutorial foi possível encontrar problemas de qualidade nos dados e propor soluções. A primeira meta a ser cumprida foi a preparação do ambiente para a realização da análise da qualidade dos dados, na qual os participantes atingiram com sucesso esse objetivo, pois realizaram a instalação do software de suporte, obtiveram uma cópia dos dados do FIES e os importaram para a ferramenta usada na análise. O segundo objetivo a ser alcançado foi a análise das dimensões da qualidade dos dados. Os participantes

atingiram com sucesso esse objetivo, pois o estudo foi realizado de forma total e as evidências foram tabuladas. A terceira meta a ser cumprida foi a elaboração de uma lista de sugestões de melhorias. Foram selecionados usuários que não possuíam formação na área de Tecnologia da Informação. As dimensões de qualidade em que o trabalho se concentrou foram: interpretabilidade, compreensibilidade, representação concisa, credibilidade, integridade e acurácia que são relacionadas ao uso dos dados.

4 SOLUÇÃO PROPOSTA

Este Capítulo apresenta a solução proposta para avaliar de forma semiautomatizada qualidade de repositórios de dados abertos governamentais através de seus metadados. O objetivo é apoiar pesquisadores e interessados na seleção de repositórios de dados abertos governamentais, tendo como referência o seu grau de alinhamento com as dimensões de qualidade de dados. A partir dos resultados fornecidos pela solução será possível selecionar repositórios segundo parâmetros previamente estabelecidos pelo usuário. O estudo Exploratório Parte 1 apresentado no Capítulo 3 serviu de referência e motivação para a implementação desta solução semiautomatizada.

4.1 JUSTIFICATIVA DE ESCOLHA DO FRAMEWORK CKAN

Escolhemos trabalhar com o Framework CKAN, devido o mesmo ser utilizado amplamente por instituições governamentais nacionais e internacionais para a abertura dos dados. Entre outras características que levam as instituições optarem pelo CKAN, destacamos que suporta metadados e viabiliza consulta, extração, análise e publicação de dados abertos, trata-se de uma ferramenta *opensource*, a internet é repleta de fóruns de desenvolvedores o utilizam e possui vasta documentação disponível. Levamos ainda em conta na escolha deste framework, o fato deste ter sido implementado no Portal Brasileiro de Dados Abertos. Por outro lado, identificamos que o trabalho de (CARDOSO; CARNEIRO; MENEZES, 2020), que serviu de ponto de partida para o desenvolvimento da presente proposta, implementa uma solução de análise semiautomática de metadados que utiliza as APIS do referido framework para a obtenção dos metadados.

4.2 IMPLEMENTAÇÃO DA SOLUÇÃO PROPOSTA

Nesta seção serão apresentadas as etapas da proposta: avalia-se a qualidade dos dados dos repositórios (FIES, MINHA CASA MINHA VIDA e PORTAL TRANSPARÊNCIA SERVIDORES) através de um *script python* empacotado no *Docker Container*.

4.3 ETAPAS DA SOLUÇÃO PROPOSTA

A Figura 7 apresenta as etapas da solução proposta nesta dissertação de mestrado.

Figura 7 – Etapas da Solução Proposta

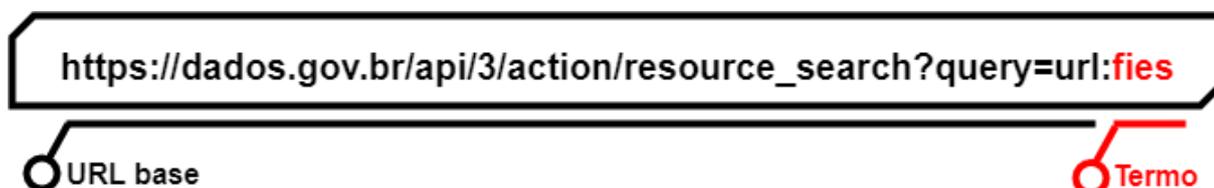


Fonte: Autoria própria (2022).

4.3.1 Detalhes da implementação

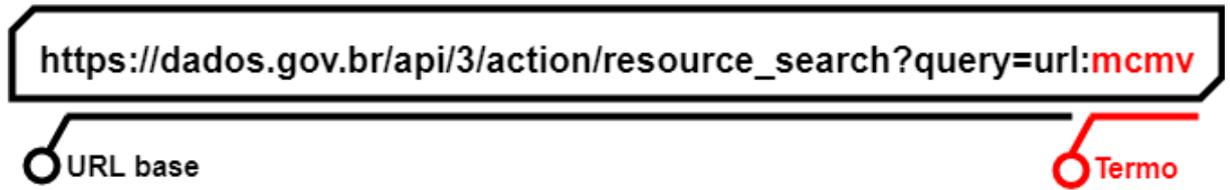
Para a realização da análise semiautomática da qualidade dos metadados, deve-se fornecer o termo de pesquisa. O termo é utilizado para recuperar os metadados do catálogo do repositório consultado. O algoritmo captura os metadados a partir da indicação da (url base + termo de pesquisa), representados nas Figuras 8, 9 e 10.

A seguir são apresentados os *endpoints* avaliados através dos algoritmos.

Figura 8 – *Endpoint* - Dados do FIES

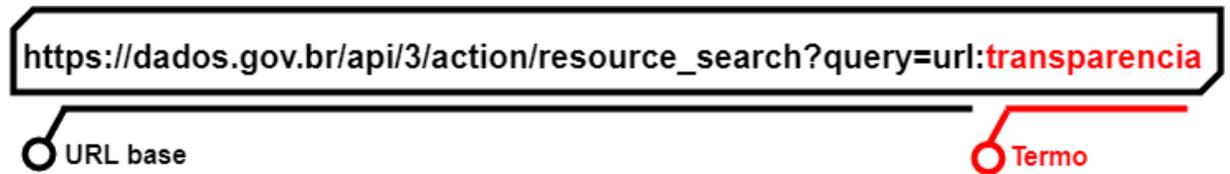
Fonte: Autoria própria (2020).

Figura 9 – *Endpoint* - Dados do Programa Minha Casa Minha Vida



Fonte: Autoria própria (2020).

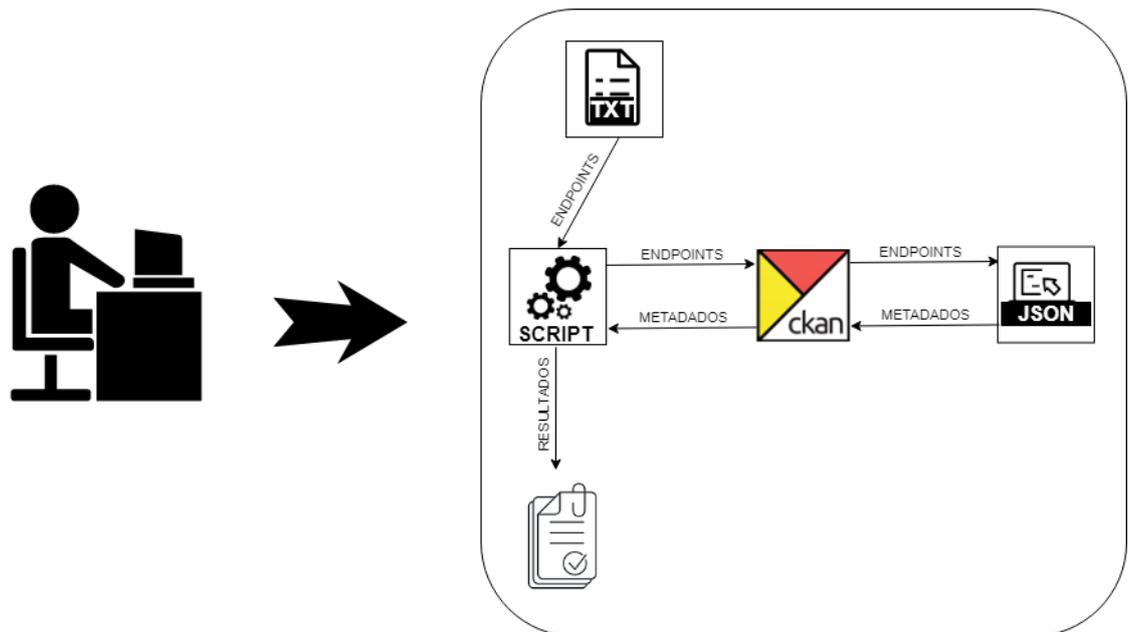
Figura 10 – *Endpoint* - Dados do Portal da Transparência - Remuneração Servidores



Fonte: Autoria própria (2020).

A Figura 11 apresenta a arquitetura da proposta de análise semiautomática dos metadados através do script.

Figura 11 – Arquitetura da Proposta de Análise Semiautomática dos Metadados



Fonte: Autoria própria (2022).

Com os resultados obtidos na coleta dos dados, inicia-se a análise dos metadados. Os algoritmos de análise percorrem os metadados *description*, *url*, *name*, *format* e *created* para validar o formato do arquivo, a descrição, o nome, a data de criação e a *url* respectivamente. Como o metadado *'url'* possibilita encontrar os dados, o mesmo é validado por meio de

requisição *http*. Para cada validação positiva, ou seja, a *url* permite a recuperação dos dados é informada a mensagem "*url* validada". Outro atributo avaliado é o formato do arquivo, que é validado quanto a compatibilidade com os formatos '*csv, xls, json, xml e pdf*' disponíveis na API CKAN e processáveis por máquinas. Os metadados avaliados se relacionam com as dimensões de qualidade completude, encontrabilidade e acessibilidade. O CKAN é uma plataforma open source utilizada para criação de portais de dados abertos. São funcionalidades básicas da plataforma CKAN: apoiar a publicação, o compartilhamento, a descoberta e o uso dos dados. O objetivo do CKAN é facilitar a criação, o acesso e a disseminação de conhecimento (WINN *et al.*, 2013).

4.4 CONCLUSÃO

Este Capítulo apresentou uma proposta de solução semiautomatizada para a avaliação de dados e metadados de portais de dados abertos governamentais a partir de limitações identificadas na abordagem manual discutida no Capítulo 3. Os detalhes de implementação da proposta foram discutidos, incluindo o objetivo de apoiar a seleção de dados abertos governamentais por parte de pesquisadores e interessados. Para esta finalidade, foram validados os metadados que facilitam a descoberta, acesso e uso dos dados. Também foram discutidas as limitações da presente solução proposta em função da disponibilização dos dados por meio do protocolo de acesso *File Transfer Protocol* (FTP) conforme discutido no Capítulo 5.

5 ESTUDO EXPLORATÓRIO - PARTE 2

5.1 INTRODUÇÃO

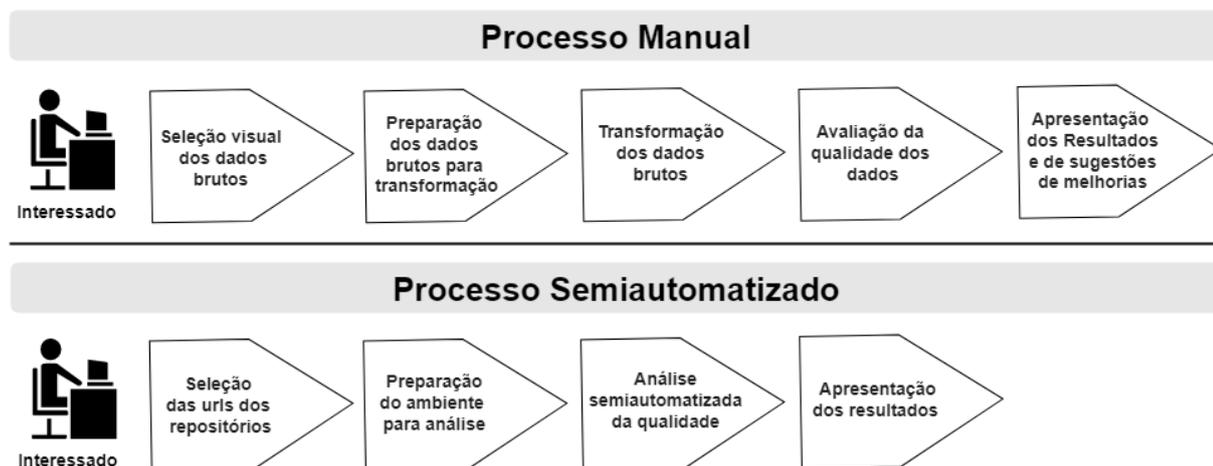
Nesta segunda parte, nos concentramos em realizar um estudo exploratório para analisar de forma semiautomatizada a qualidade de metadados em portais governamentais abertos. Para esta finalidade, pesquisamos na literatura diversos estudos e encontramos o estudo intitulado *Dados Abertos Governamentais no Contexto de Políticas Públicas de Saúde e Sistemas Prisionais: Realidade ou Utopia* que apresenta uma proposta de avaliação semiautomatizada de metadados que possibilita a medição do grau de alinhamento com as boas práticas de dados abertos governamentais (CARDOSO; CARNEIRO; MENEZES, 2020).

Em seguida, executamos um projeto piloto com base na solução proposta de Cardoso, Carneiro e Menezes (2020) utilizamos no projeto piloto, os termos de busca ‘Covid-19’ e ‘FIES’ para identificarmos se os algoritmos da solução proposta são capazes de validar os metadados vinculados às dimensões de qualidade Acessibilidade e Encontrabilidade.

Feito isto, constatamos a necessidade de implementarmos novos algoritmos para que a solução pudesse atender os objetivos deste estudo. Desta forma, desenvolvemos e incluímos novos algoritmos capazes de validar os metadados *url, format, created, name e description* que se relacionam com as dimensões de qualidade Acessibilidade e Encontrabilidade.

No presente estudo, adaptamos a solução encontrada na literatura e a utilizamos para avaliarmos a efetividade dos novos algoritmos no acesso via API CKAN aos dados dos repositórios FIES, Minha Casa Minha Vida e Portal da Transparência para medir o grau de alinhamento dos dados e metadados com as dimensões de qualidade e auxiliar com a descoberta, o acesso e uso de dados abertos governamentais. A Figura 12 apresenta os processos de avaliação da qualidade dos dados de forma manual e semiautomatizado.

Figura 12 – Processo Manual e Processo Semiautomatizado



Fonte: Autoria própria (2022).

Conforme apresentado na Figura 12 o processo manual é realizado pelo interessado através da navegação no portal para a seleção dos dados brutos, como descrito no Capítulo 4. Esta abordagem apresenta características que dificultam a análise da qualidade dos dados abertos governamentais. O primeiro ponto está na seleção e transformação dos dados brutos em formato que possibilita a análise manual, onde o interessado tem que obter os dados e efetuar a limpeza, transformação e importação dos mesmos para uma ferramenta própria para realização da análise. Esta abordagem manual dificulta a seleção, acesso, coleta e análise dos dados abertos governamentais por parte dos interessados.

No processo semiautomatizado desta solução, a busca é direcionada ao catálogo de metadados do Portal Brasileiro de Dados Abertos, que se refere ao contexto OGD. A partir dos resultados da pesquisa, são aplicados algoritmos sobre as estruturas de metadados disponibilizados neste catálogo, onde é avaliada a qualidade dos dados através dos metadados.

Após a avaliação é retornada uma lista com o resultado da validação dos metadados *'url'*, *'format'*, *'created'*, *'name'* e *'description'*. A solução proposta ainda conta com o uso de arquivo de texto para inserção dos endereços dos repositórios a serem avaliados e com o empacotamento da aplicação em um Docker Container, que otimiza o processo de análise. Esta abordagem facilita todo processo de seleção, coleta, e análise dos dados.

5.1.1 Metadados Utilizados no Estudo

Para acessar os metadados avaliados, foram realizadas através do Framework CKAN as seguintes consultas no Portal Brasileiro de Dados Abertos:

Quadro 8 – Lista de *Endpoints* Consultados

urls
https://dados.gov.br/api/3/action/resource_search?query=url:fies
https://dados.gov.br/api/3/action/resource_search?query=url:mcmv
https://dados.gov.br/api/3/action/resource_search?query=url:transparencia

Fonte: Autoria própria (2011)

As figuras a seguir com os retornos das consultas revelam os metadados e os valores armazenados nos mesmos. Foram utilizados os termos FIES, MCMV e transparencia para consultar os repositórios selecionados. Além disso, utilizou-se o navegador Google Chrome e a extensão JSON Formatter para apresentação dos dados no formato JSON. Os resultados das consultas são apresentados em sequência.

Figura 13 – Retorno da Consulta - API CKAN - Catálogo do FIES

Endpoint com metadados do FIES:

https://dados.gov.br/api/3/action/resource_search?query=url:fies

```

{
  "help": "https://dados.gov.br/api/3/action/help_show?name=resource_search",
  "success": true,
  "result": {
    "count": 107,
    "results": [
      { ... }, // 21 items
      { ... }, // 19 items
      { ... }, // 21 items
      { ... }, // 21 items
      { ... }, // 21 items
      {
        "mimetype": null,
        "cache_url": null,
        "state": "active",
        "hash": "",
        "description": "Valores dos cursos ofertados para o Fies 2/2019",
        "format": "CSV",
        "url": "http://dadosabertos.mec.gov.br/images/conteudo/fies/2019/fies_oferta_2019_2.csv",
        "created": "2019-12-26T23:16:39.059233",
        "cache_last_updated": null,
        "package_id": "677f7671-52da-48a4-ac3e-69cbfa1fedc4",
        "mimetype_inner": null,
        "last_modified": null,
        "position": 1,
        "revision_id": "b6de33b0-b38d-4598-8edd-12a93b47d223",
        "size": null,
        "url_type": null,
        "id": "c35b1dea-c8d8-4619-b960-6a52fddf8b0c",
        "resource_type": null,
        "name": "FIES Oferta 2/2019"
      }
    ]
  }
}

```

Fonte: Autoria própria (2022).

Figura 14 – Retorno da Consulta - API CKAN - Catálogo Minha Casa Minha Vida

Endpoint com metadados do Minha Casa Minha Vida:

https://dados.gov.br/api/3/action/resource_search?query=url:mcmv

```

{
  "help": "https://dados.gov.br/api/3/action/help_show?name=resource_search",
  "success": true,
  "result": {
    "count": {
      "count": 7,
      "results": [
        {
          "mimetype": "text/csv",
          "cache_url": null,
          "state": "active",
          "hash": "",
          "description": "",
          "format": "CSV",
          "url": "http://www.geoservicos.inde.gov.br/geoserver/wms?service=WFS&version=1.0.0&request=GetFeature&typeName=MPOG:MCHV&outputFormat=CSV",
          "created": "2018-06-12T05:08:08.632076",
          "cache_last_updated": null,
          "package_id": "2e1a5ade-d1c7-45da-b953-c6e8a65f9e83",
          "mimetype_inner": null,
          "last_modified": null,
          "position": 0,
          "revision_id": "5ea9baf0-dba2-4590-8a7a-510cad518072",
          "size": null,
          "url_type": null,
          "id": "f595714a-0356-40f2-aeae-bf7695a372e4",
          "resource_type": null,
          "name": "Programa Minha Casa Minha Vida"
        }
      ]
    }
  }
}

```

Fonte: Autoria própria (2022).

Figura 15 – Retorno da Consulta - API CKAN - Catálogo do Portal da Transparência

Endpoint com metadados do Portal da Transparência:

https://dados.gov.br/api/3/action/resource_search?query=url:transparencia

```

{
  "help": "https://dados.gov.br/api/3/action/help_show?name=resource_search",
  "success": true,
  "result": {
    "count": {
      "count": 25,
      "results": [
        {
          "mimetype": null,
          "cache_url": null,
          "state": "active",
          "hash": "",
          "description": "Banco de download do Portal da Transparência do Governo Federal com informações, em formato aberto, sobre cargo, função, situação funcional e remuneração dos servidores civis e militares, bem como dos agentes públicos do Poder Executivo Federal.",
          "format": "xlsx",
          "url": "http://transparencia.gov.br/download-de-dados/servidores",
          "datastore_active": false,
          "created": "2017-09-28T15:11:47.750900",
          "cache_last_updated": null,
          "package_id": "4cf53feb-ffda-4897-b216-87270cb48800",
          "mimetype_inner": null,
          "last_modified": null,
          "position": 0,
          "revision_id": "5410072a-be3f-4429-9528-ffa2f1d124e2",
          "size": null,
          "url_type": null,
          "id": "44e042a-b001-4a1f-a08e-ade5debf21f",
          "resource_type": null,
          "name": "Servidores civis e militares do Executivo Federal"
        }
      ]
    }
  }
}

```

Fonte: Autoria própria (2022).

Nas Figuras 13, 14 e 15 é possível verificar o conteúdo dos metadados *'url'*, *'format'*, *'created'*, *'name'* e *'description'* que permitem a descoberta, acesso e uso dos dados.

5.2 PLANEJAMENTO DO ESTUDO

A seleção dos repositórios alvos foi predefinida com base na disponibilidade de consultas nos repositórios através do Framework CKAN, esta característica facilita o processo de coleta dos metadados, tendo em vista, que é possível obter uma lista de metadados por meio de algoritmos. Optamos implementar nos algoritmos consultas no padrão da API CKAN, pois trata-se do padrão adotado pelo governo brasileiro para a publicações de dados governamentais abertos.

Na preparação do ambiente da análise semiautomática criamos o arquivo “parametros.txt” e inserimos no mesmo os *endpoints* referentes aos repositórios analisados. O arquivo com os parâmetros foi lido através da aplicação e as consultas foram direcionadas aos *endpoints* listados no arquivo.

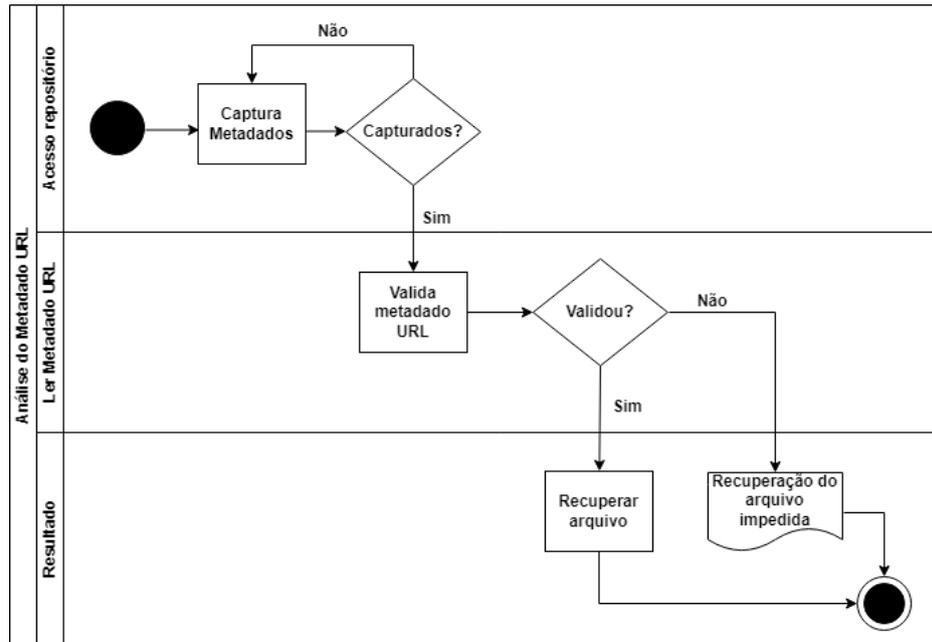
5.3 EXECUÇÃO DO ESTUDO CASO

5.3.1 Análises Realizadas pelo Algoritmos

Foi realizada a análise da qualidade dos metadados de forma semiautomatizada por meio da execução de algoritmos de validação dos seguintes metadados: *url*, *format*, *created*, *name* e *description*.

Para validar o metadado *url*, foi verificado se o metadado encontrava-se vazio e se inicia-se com *'http'*. Em caso positivo, foi passado o valor do metadado *url* para o algoritmo de recuperação do arquivo, do contrário apresentou-se a mensagem “recuperação do arquivo impedida”. A Figura 16 apresenta o processo de validação do metadado *url*.

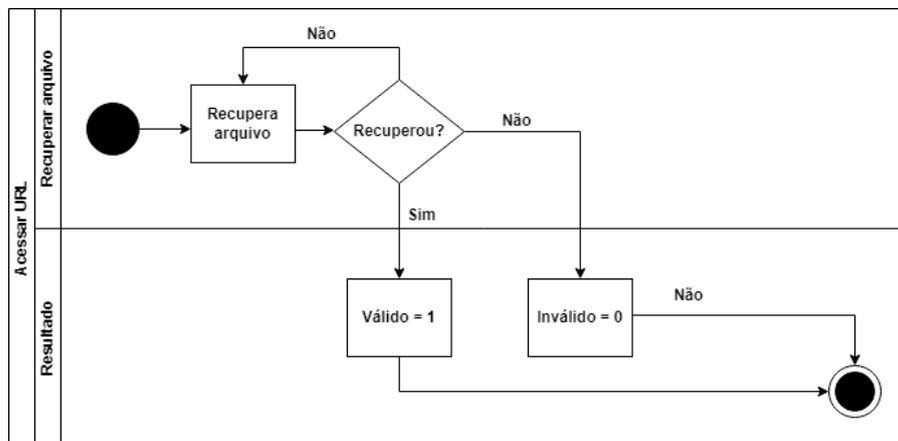
Figura 16 – Processo de Análise do Metadado URL



Fonte: Autoria própria (2022).

Para validar a recuperação do arquivo, foi realizada uma requisição do arquivo pela *url* constante no metadado. Apresentou-se as mensagens '*url* válida' e '*url* inválida' de acordo com a situação encontrada. A Figura 17 apresenta o processo de recuperação do arquivo.

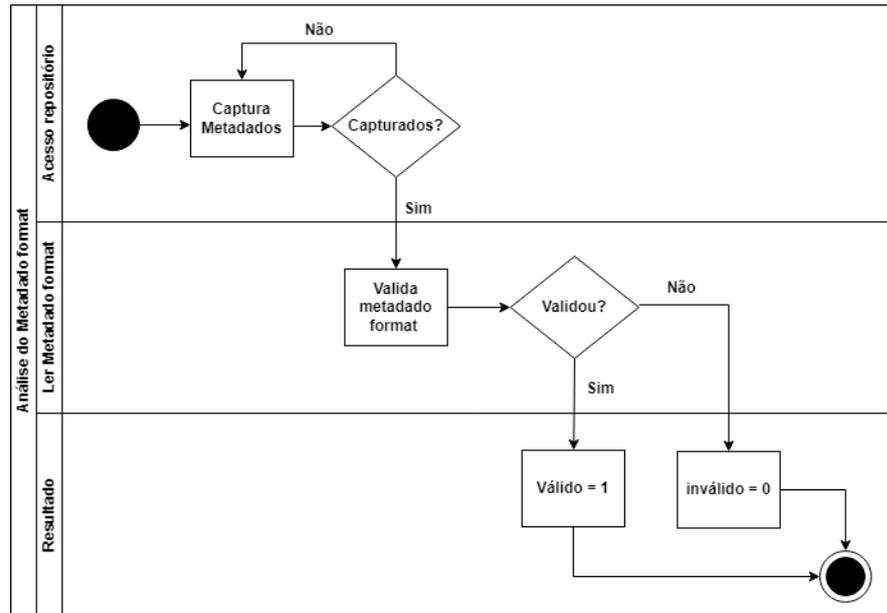
Figura 17 – Processo de Recuperação do Arquivo



Fonte: Autoria própria (2022).

Para validar o metadado *format*, foi verificado se o metadado encontrava-se vazio e se o formato é permitido pela API CKAN. Apresentou-se as mensagens '*formato válido*' e '*formato inválido*' para cada situação encontrada. A Figura 18 apresenta o processo de validação do metadado *format*.

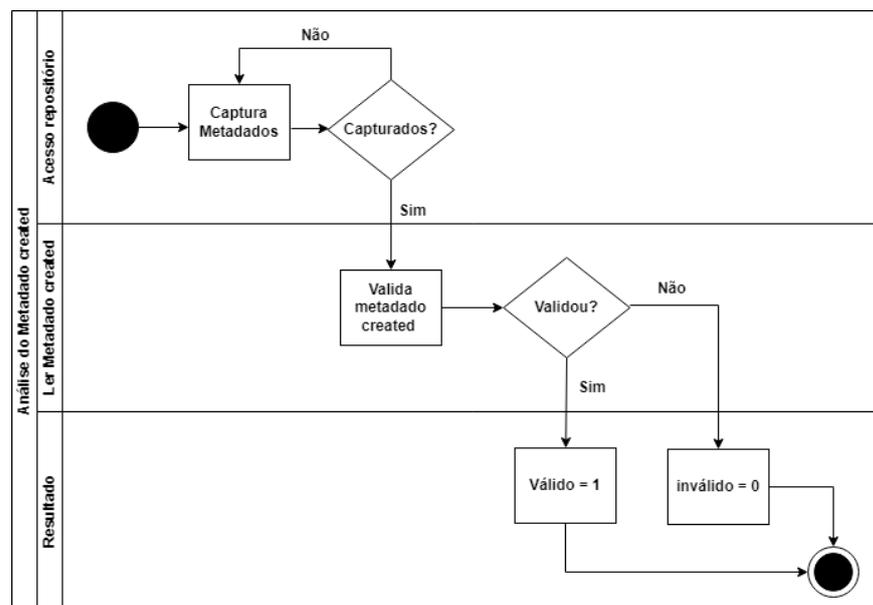
Figura 18 – Processo de Análise do Metadado FORMAT



Fonte: Autoria própria (2022).

Para validar o metadado *created*, foi verificado se o metadado encontrava-se vazio. Apresentou-se as mensagens 'data válida' e 'data inválida' de acordo com a situação encontrada. A Figura 19 apresenta o processo de validação do metadado *created*.

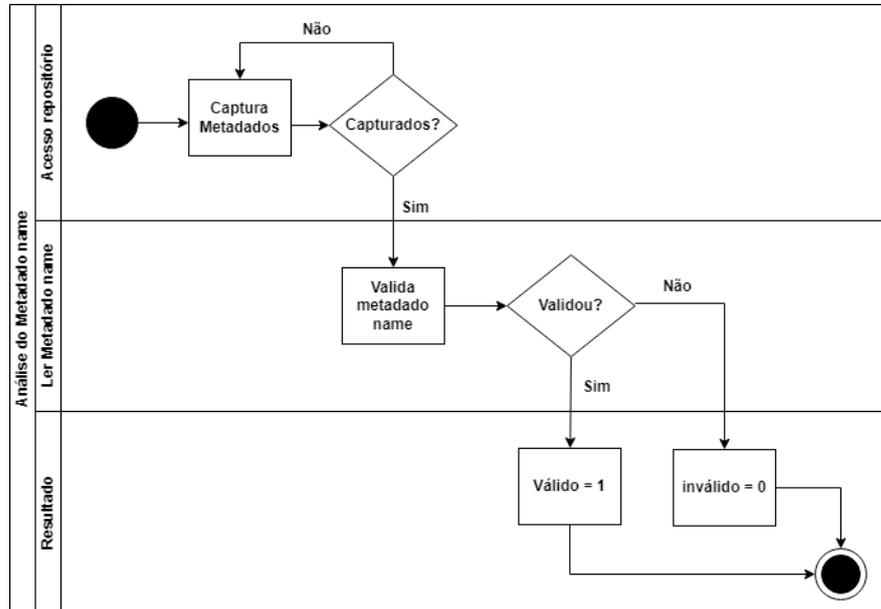
Figura 19 – Processo de Análise do Metadado CREATED



Fonte: Autoria própria (2022).

Para validar o metadado *name*, foi verificado se o metadado encontrava-se vazio. Apresentou-se as mensagens 'nome válido' e 'nome inválido' de acordo com a situação encontrada. A Figura 20 apresenta o processo de validação do metadado *name*.

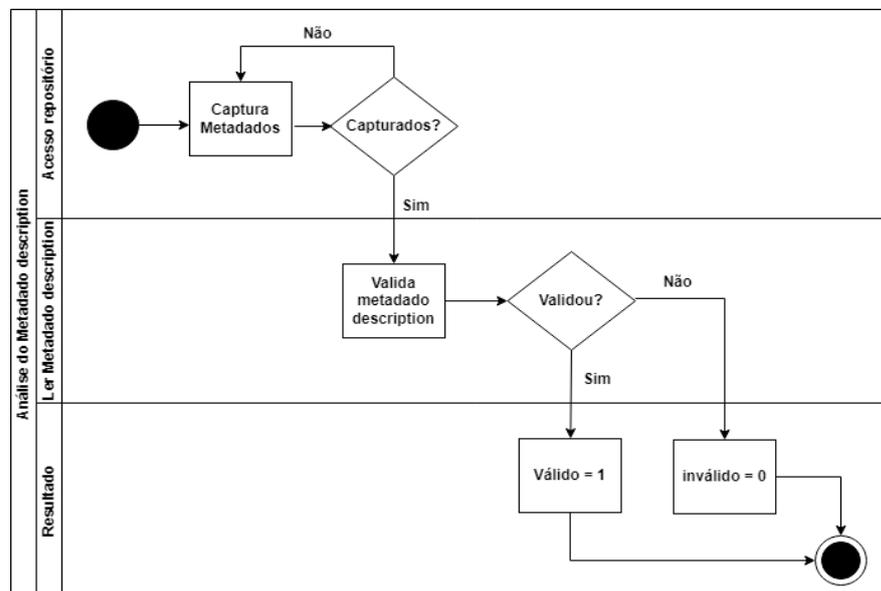
Figura 20 – Processo de Análise do Metadado NAME



Fonte: Autoria própria (2022).

Para validar o metadado *description*, foi verificado se o metadado encontrava-se vazio. Apresentou-se as mensagens 'descrição válida' e 'descrição inválida' de acordo com a situação encontrada. A Figura 21 apresenta o processo de validação do metadado *description*.

Figura 21 – Processo de Análise do Metadado DESCRIPTION



Fonte: Autoria própria (2022).

5.3.2 Resultados

Neste estudo foram considerados 3 repositórios, onde foram analisados 471 recursos pela solução proposta. Em cada um dos recursos foram validados 5 metadados, obtidos do Portal Brasileiro de Dados Abertos através de consultas via frameworks CKAN acionadas pelo script escrito em *Python*.

5.3.3 Análise dos Resultados

O Quadro 9 apresenta os repositórios selecionados e o quantitativo de recursos avaliados de cada um.

Quadro 9 – Quantidade de Recursos Avaliados por Repositório

Repositórios	Nº de Recursos Avaliados
FIES	107
Minha Casa Minha Vida	7
Portal da Transparência	357
TOTAL	471

Fonte: Autoria própria (2022).

Os Quadros 10, 11 e 12 apresentam os resultados da aplicação da solução proposta a este Estudo Exploratório Parte 2.

Quadro 10 – Percentual dos Metadados Válidos e Inválidos do Repositório – FIES

Metadados	Nº Válidos	Nº Inválidos	% Válidos	% Inválidos
<i>url</i>	33	74	30,84	69,16
<i>format</i>	106	1	99,07	0,93
<i>created</i>	107	0	100	0
<i>name</i>	107	0	100	0
<i>description</i>	106	1	99,07	0,93

Fonte: Autoria própria (2022).

O Quadro 10 apresenta o resultado da análise dos metadados do repositório FIES. Foram avaliados neste repositório 107 recursos, dos quais o metadado '*url*' possui 33 válidos e 74 inválidos (30,84% e 69,16%), o metadado '*format*' 106 válidos e 1 inválido (99,07% e 0,93%), os metadados '*created*' e '*name*' 107 válidos (100%) e o metadado '*description*' 106 válidos e 1 inválido (99,07% e 0,93%).

Quadro 11 – Percentual dos Metadados Válidos e Inválidos do Repositório - Minha Casa Minha Vida

Metadados	Nº Válidos	Nº Inválidos	% Válidos	% Inválidos
<i>url</i>	7	0	0	0
<i>format</i>	7	0	0	0
<i>created</i>	7	0	0	0
<i>name</i>	7	0	0	0
<i>description</i>	3	4	42,86	57,14

Fonte: Autoria própria (2022).

O Quadro 11 apresenta o resultado da análise dos metadados do repositório Minha Casa Minha Vida. Foram avaliados neste repositório 7 recursos, dos quais os metadados '*url*', '*format*', '*created*', '*name*' possuem 7 recursos válidos (100%) e o metadado '*description*' possui 3 válidos e 4 inválidos (42,86% e 57,14%).

Quadro 12 – Percentual dos Metadados Válidos e Inválidos do Repositório - Portal da Transparência

Metadados	Nº Válidos	Nº Inválidos	% Válidos	% Inválidos
<i>url</i>	357	0	100	0
<i>format</i>	333	24	93,28	6,72
<i>created</i>	353	4	98,88	1,12
<i>name</i>	356	1	99,72	0,28
<i>description</i>	317	40	88,80	11,20

Fonte: Fonte: Autoria própria (2022).

O Quadro 12 apresenta o resultado da análise dos metadados do repositório Portal da Transparência. Foram avaliados neste repositório 357 recursos, dos quais o metadado '*url*' possui 357 válidos (100%), o metadado '*format*' 333 válidos e 24 inválidos (93,28% e 6,72%), o metadado '*created*' 353 válidos e 4 inválidos (98,88% e 1,12%), o metadado '*name*' 356 válidos e 1 inválido (99,72% e 0,28%) e o metadado '*description*' 317 válidos e 40 inválidos (88,80% e 11,20%).

O Quadro 13 apresenta as limitações identificadas na análise semiautomática da qualidade dos metadados dos repositórios FIES, Minha Casa Minha Vida (MCMV) e Portal da Transparência.

Quadro 13 – Limitações Encontradas após a Análise Semiautomática

Limitações dos Metadados			
Repositório	Metadado	Quantidade	Descrição
FIES	<i>url</i>	74	não foi possível acessar as <i>urls</i> dos recursos
	<i>format</i>	1	não foi preenchido o formato do recurso
	<i>description</i>	1	não foi preenchida a descrição do recurso
MCMV	<i>description</i>	4	não foram preenchidas as descrições dos recursos
TRANSPARÊNCIA	<i>format</i>	24	não foram preenchidos os formatos
	<i>created</i>	1	não foi preenchida a data de criação do recurso
	<i>name</i>	1	não foi preenchido o nome do recurso
	<i>description</i>	42	não foram preenchidas as descrições dos recursos

Fonte: Fonte: Autoria própria (2022).

A análise semiautomática dos repositórios permite identificar não conformidade entre os recursos e as dimensões de qualidade. Observa-se que foram encontrados recursos com os metadados *format*, *description*, *created* e *name* não preenchidos, isto demonstra não conformidade desses recursos com a dimensão de qualidade '*completeness*'. Além disso, foram encontrados recursos com o metadado '*url*' que não permitiu o acesso, descoberta e uso dos dados, isto indica não conformidade desses recursos com as dimensões de qualidade '*accessibility*' e '*findability*'.

Os resultados da análise semiautomática dos repositórios de dados abertos (FIES, MCMV e PORTAL DA TRANSPARÊNCIA) nos revela que a quantidade de recursos avaliados foi 107 no primeiro, 7 no segundo e no 357 terceiro, respectivamente representam 22,72%, 1,49% e 75,80%. A quantidade total de recursos avaliados foi 471, ou seja, 100% dos recursos. Os resultados da análise do primeiro repositório mostram que foram encontrados apenas os metadados (*url*, *format* e *description*) inválidos. Enquanto os resultados da análise do segundo repositório mostram que foi encontrado apenas o metadado (*description*) inválido e os resultados da análise do terceiro repositório mostram que foram encontrados os metadados (*format*, *created*, *name* e *description*) inválidos. Além disso, foram encontrados recursos em que não foi possível acessar a *url* contida no metadado de mesmo nome e foram encontrados recursos onde os metadados (*format*, *description*, *created*, *name* e *description*) não foram preenchidos. Desta forma, nota-se que foram encontradas não conformidades com as dimensões de qualidade, o que comprova a eficiência da solução proposta para avaliação dos metadados.

5.3.4 Limitação identificada no uso da solução

Foi encontrada limitação na solução para a análise do metadado 'url', pois identificamos recursos com endereços do protocolo *File Transfer Protocol* (FTP) atribuídos ao mesmo. Ao validar esses endereços, não foi possível acessar os dados. Desta forma, o algoritmo de validação do metadado 'url' o validou parcialmente.

5.4 RESPOSTA DA QUESTÃO DE PESQUISA (QPE3)

Questão de Pesquisa - QPE3: *Qual a efetividade de uma solução semiautomatizada para identificar limitações e erros em metadados de repositórios de dados abertos governamentais brasileiros no contexto de programas sociais e gastos com pessoal em relação as dimensões de qualidade de dados?*

Resposta: A efetividade de uma solução semiautomatizada pode ser constatada nos Quadros 10, 11, 12 e 13, todas apresentadas na Subseção 5.3.3.

5.5 CONCLUSÃO

Esta segunda parte do estudo permitiu validar a efetividade no uso da solução proposta aplicada a este estudo exploratório, além de responder a Questão de Pesquisa Específica (QPE3). Foi possível identificar características associadas ao uso do framework CKAN, evidenciando por meio dos resultados analisados pela solução proposta, padrões de não conformidade com as dimensões de qualidade de dados. Entretanto, apenas a utilização do framework sem o devido preenchimento dos metadados disponibilizados, demonstra a importância da utilização correta do framework para se obter maiores percentuais no grau de alinhamento com as dimensões de qualidade e, conseqüentemente, maior qualidade nos dados disponibilizados.

6 CONCLUSÃO E PERSPECTIVAS FUTURAS

Este Capítulo apresenta as considerações finais, contribuições e perspectivas de trabalhos futuros.

6.1 RESPOSTA DA QUESTÃO DE PESQUISA (QP)

Questão de Pesquisa - QP: *A estrutura de dados abertos governamentais implementada pelo governo brasileiro atende os requisitos das dimensões de qualidade de dados?*

Resposta: Diante dos estudos, análises e resultados apresentados nesta dissertação, foram encontradas evidências de que os dados do Portal Brasileiro de Dados Abertos referentes aos repositórios FIES, Minha Casa Minha Vida e Portal da Transparência apresentam não conformidades de dados e metadados com as Dimensões de Qualidade de Dados. Portanto, atende de forma parcial, necessitando de adoção de estratégias que visem a avaliação constante da qualidade dos dados.

6.2 CONSIDERAÇÕES FINAIS

Este trabalho permitiu identificar os desafios para a implementação de dados abertos governamentais e apresentou duas alternativas para a avaliação da qualidade de dados abertos governamentais. Para esta finalidade, foi realizado um estudo exploratório dividido em duas partes. O estudo foi executado para responder as questões de pesquisa QP, QPE1, QPE2 e QPE3 estabelecidas nesta dissertação.

Na execução da primeira parte do estudo exploratório (Capítulo 3), foi possível identificar erros e limitações nos dados do FIES. Os erros encontrados neste repositório foram valores incorretos que apresentaram não conformidade com as dimensões interpretabilidade, compreensibilidade, credibilidade e acurácia, valores de atributos ausentes não conformidade com a dimensão completude, valores ilegais não conformidade com as dimensões consistência e credibilidade.

Na execução da segunda parte do estudo exploratório (Capítulo 5), foi possível identificar limitações em parte dos metadados dos repositórios FIES, Minha Casa Minha Vida e Portal da Transparência. Nesta etapa, foram avaliados os metadados *url*, *format*, *created*,

name e description. As limitações encontradas apontam que há não conformidade entre os dados e metadados e as dimensões de qualidade 'Completeness', Acessibilidade e Encontrabilidade, estas últimas estão relacionadas a descoberta, o acesso e uso dos dados. Diante deste cenário fica evidente a relevância do tema abordado nesta dissertação e como a utilização da API CKAN contribui para a disponibilização de dados com maior grau de qualidade.

O framework CKAN fornece uma infraestrutura que possibilita o acesso e automação do processo de coleta e análise dos dados e seus metadados. Este último, possui extrema relevância, pois é por meio da avaliação dos metadados que se torna possível a avaliação da qualidade dos dados e metadados. Em observância a este cenário, entende-se que os indicadores são positivos quando é utilizado o framework CKAN. Observou-se que este framework possibilita a avaliação do grau de alinhamento dos dados e metadados com as dimensões de qualidade. Entretanto, o fato de tão somente utilizá-lo não garante o total alinhamento com as dimensões de qualidade já citadas, tornando necessário esforço adicional para o referido alinhamento. Isto fica evidenciado nos resultados e análises apresentadas nesta dissertação. Como exemplo, diversos recursos analisados pela solução proposta apresentaram metadados que não puderam ser validados ou não foram preenchidos.

Isto é uma evidência de que embora a solução tecnológica disponibiliza a estrutura adequada, sem o preenchimento correto do conteúdo do metadado, determinados recursos terão sua qualidade impactada. Portanto, fica evidente a necessidade de uma estratégia efetiva para o uso do framework da forma correta para que efetivamente os dados sejam disponibilizados com maior qualidade. Por isso, existem muitas iniciativas que mesmo utilizando frameworks OGD, ainda apresentam baixo grau de alinhamento, além das iniciativas próprias que não apresentam requisitos mínimos de qualidade dos dados disponibilizados. A solução proposta nesta dissertação permitiu identificar os aspectos aqui mencionados, e ainda auxilia na superação dos desafios de implementação de dados abertos governamentais.

6.3 CONTRIBUIÇÕES

Apoiar na identificação e endereçamento de oportunidades de melhorias de portais nacionais de dados abertos governamentais com as áreas temáticas: educação, habitação e despesas com pessoal. Acesso e uso da solução implementada no contexto desta dissertação para apoiar a avaliação da qualidade dos metadados, e, por extensão, apoiar também na identificação de repositórios de excelência. Estes últimos podem servir de referência para outros

repositórios que precisam evoluir e, portanto, melhorar o alinhamento com as dimensões de qualidade de dados.

6.4 TRABALHOS FUTUROS

Considerando os resultados alcançados e a elevada importância do tema estudado, percebemos que é possível aprimorar esta solução, adicionando uma funcionalidade que recomende aos usuários uma lista de possíveis soluções que possam corrigir limitações encontradas nos metadados através dos algoritmos que analisam os dados e metadados.

REFERÊNCIAS

AGOSTINI, V. A. **Dadosjusbr.online**: uma ferramenta para democratização das informações de pagamentos para magistrados. 2019. 13 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Centro de Engenharia Elétrica e Informática - Universidade Federal de Campina Grande. Campina Grande, 2019. Disponível em: <http://dspace.sti.ufcg.edu.br:8080/jspui/bitstream/riufcg/20779/1/VINICIUS%20ALENCAR%20AGOSTINI%20-%20TCC%20CI%20NCIA%20DA%20COMPUTA%20C%27%20O%202019.pdf> Acesso em: 20 mar. 2020.

ALMADA SILVA, M. P.; DOMINGUEZ, M. E-transparência e dados abertos governamentais: o ambiente digital como lócus dos avanços do governo brasileiro nas políticas, leis e parcerias internacionais. *In: CONGRESSO DA ASSOCIAÇÃO BRASILEIRA DE PESQUISADORES EM COMUNICAÇÃO E POLÍTICA*, 7., Porto Alegre. **Anais [...]**Porto Alegre: Compolítica; UFRGS, 2017. Disponível: https://stateless-inctdd-website.storage.googleapis.com/2017/12/almada_dominguez_ettransparencia-e-dados-abertos-governamentais.pdf Acesso em: 14 abr. 2021.

AQUINO JÚNIOR, G. S. *et al.* Dados abertos para o fomento da transparência e inovação: o caso da UFRN. **ISYS-Brazilian Journal of Information Systems**, Porto Alegre, v. 12, n. 2, p. 39–59, 2019. Disponível em: <https://sol.sbc.org.br/journals/index.php/isys/article/view/388> Acesso em: 27 nov. 2021.

BASILI, V. R.; ROMBACH, H. D. Towards a comprehensive framework for reuse: a reuse-enabling software evolution environment. *In: ANNUAL SOFTWARE ENGINEERING WORKSHOP*. 13, 1988. **Proceedings [...]** Greenbelt, MD: NASA, 1988. Disponível em: <https://ntrs.nasa.gov/citations/19910001295> Acesso em: 26 fev. 2021.

BATINI, C.; PERNICI, B. Data quality management and evolution of information systems. *In: AVISON, D. et al. (ed.) The past and future of information systems: 1976–2006 and beyond*. Boston: Springer, 2006. p.51-62. Disponível em: https://link.springer.com/content/pdf/10.1007/978-0-387-34732-5_5.pdf Acesso em: 20 out. 2021.

BRASIL. **Decreto n.º 8.777, de 11 de maio de 2016**. Institui a Política de Dados Abertos do Poder Executivo Federal. Brasília, DF, 2016. Disponível em: http://www.planalto.gov.br/CCIVIL_03/_Ato2015-2018/2016/Decreto/D8777.htm Acesso em: 09 jul. 2021.

BRASIL. **Lei n.º 12.527, de 18 de novembro de 2011**. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Brasília, DF, 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/112527.htm Acesso em: 09 jul. 2021.

BRASIL. Controladoria Geral da União. **Acesso à informação pública**: uma introdução à lei nº 12.527, de 18 de novembro de 2011b. Brasília: CGU, 2011. Disponível em: <https://www.gov.br/acessoinformacao/pt-br/central-de-conteudo/publicacoes/arquivos/cartilhaacessoinformacao-1.pdf> Acesso em: 14 maio 2021.

BRASIL. Tribunal de Contas da União. **5 motivos para a abertura de dados na administração pública**. Brasília: TCU, 2015. Disponível em: <https://portal.tcu.gov.br/5-motivos-para-a-abertura-de-dados-na-administracao-publica.htm> Acesso em: 14 maio 2021.

CARDOSO, R. A. L.; CARNEIRO, G. d. F.; MENEZES, J. E. X. de. Dados abertos governamentais no contexto de políticas públicas de saúde e sistemas prisionais: realidade ou utopia? **Diálogos Possíveis**, Salvador, v. 19, n. 2, p. 65-80, 2020. Disponível em: <http://revistas.unisba.edu.br/index.php/dialogospossiveis/article/view/650> Acesso em: 26 maio 2022.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E. Profissional da informação no contexto dos dados abertos: o uso do CKAN para a disponibilização e a organização de dados. **Informação@ Profissões**, Londrina, v. 5, n. 2, p. 55–78, 2016. Disponível em: <https://brapci.inf.br/index.php/res/v/64591> Acesso em: 16 set. 2021.

CKAN. **Overview**. 2013. Disponível em: <http://docs.ckan.org/en/latest/>. Acesso em: 02 jun. 2021.

DAWES, S. S. Stewardship and usefulness: policy principles for information-based transparency. **Government Information Quarterly**, v. 27, n. 4, p. 377-383, oct. 2010.

DENIS, J.; GOËTA, S. **Exploration, extraction and ‘rawification’** the shaping of transparency in the back rooms of open data. 2014. Disponível em: <https://ssrn.com/abstract=2403069> Acesso em: 13 jul. 2021.

DING, L.; PERISTERAS, V.; HAUSENBLAS, M. Linked open government data. **IEEE Intelligent Systems**, v. 27, n. 3, p. 11–15, 2012.

DINIZ, R. H. N. *et al.* **Utilizando o Google Classroom como ferramenta educacional: percepções e potenciais**. Para de Minas, MG: FAPAM, 2018. Disponível em: <http://www.abed.org.br/congresso2018/anais/trabalhos/5896.pdf> Acesso em 4 abr. 2021.

DUTRA, F. G.; BARBOSA, R. R. Modelos e critérios para avaliação da qualidade de fontes de informação: uma revisão sistemática de literatura. **Informação & Sociedade: Estudos**, João Pessoa, v. 27, n. 2, p. 19-33, maio/ago. 2017. Disponível em: <https://periodicos.ufpb.br/ojs/index.php/ies/article/view/32676/18197> Acesso em: 27 abr. 2021.

FERRANTI, D. D. *et al.* **How to improve governance**: a new framework for analysis and action. Washington, DC: Brookings Institution Press, 2009.

HARRISON, T. M. *et al.* Open government and e-government: democratic challenges from a public value perspective. **Information Polity**, Amsterdam, v. 17, n. 2, p. 83–97, 2012.

HASSENSTEIN, M. J.; VANELLA, P. Data quality: concepts and problems. **Encyclopedia, MDPI**, v. 2, n. 1, p. 498–510, 2022. Disponível em: <https://www.mdpi.com/2673-8392/2/1/32/htm> Acesso em: 18 jul. 2022.

KUBLER, S. *et al.* Comparison of metadata quality in open data portals using the analytic hierarchy process. **Government Information Quarterly**, v. 35, n. 1, p. 13–29, 2018.

OECD. **The path to becoming a data-driven public sector**. Paris: OECD Publishing, 2019.

OLIVEIRA, L. E. R. A.; LÓSCIO, B. F. Uma abordagem para captura de informações sobre aplicações que fazem uso de dados abertos. **Revista Brasileira de Administração Científica**, Aquidabã, SE, v. 5, n. 2, p. 127–140, 2014. Disponível em: <https://sustenere.co/index.php/rbadm/article/view/SPC2179-684X.2014.002.0010/471> Acesso em: 28 out. 2020.

OLIVEIRA, P.; RODRIGUES, F.; HENRIQUES, P. **Limpeza de dados-uma visão geral**. 2004. Disponível em: <http://wiki.di.uminho.pt/twiki/pub/Research/Doutoramentos/SDDI2004/ArtigoOliveira.pdf> Acesso em: 25 set. 2020.

OPEN KNOWLEDGE BRASIL. **Um capítulo da Open Knowledge Foundation no Brasil**. São Paulo, [2021] Disponível em: <https://ok.org.br/sobre/> Acesso em: 14 nov. 2021

PASQUIER, M.; VILLENEUVE, J.-P. Organizational barriers to transparency: a typology and analysis of organizational behaviour tending to prevent or restrict access to information. **International Review of Administrative Sciences**, London, v. 73, n. 1, p. 147–162, 2007.

PINHO, J. A. G. Investigando portais de governo eletrônico de estados no Brasil: muita tecnologia, pouca democracia. **Revista de Administração Pública**, Rio de Janeiro, v. 42, p. 471–493, 2008. Disponível em: <https://www.scielo.br/j/rap/a/j8dGWgrkM8KVBwm5gzXhSJr/abstract/?lang=pt> Acesso em: 15 ago. 2021.

PINTO FILHO, J. C. R.; FONTES, C. V. **Estudo de caso “Operação Serenata De Amor”**: a análise de big data no combate à festa dos gastos públicos. *In*: CONGRESSO DE LA ASOCIACIÓN LATINOAMERICANA DE INVESTIGADORES DE LA COMUNICACIÓN. 14., 2018. Costa Rica. **Anais [...]** Costa Rica: ALAIC, 2018. Disponível em: <https://www.alaic.org/wp-content/uploads/2022/03/GT-10-ALAIC-2018.pdf> Acesso em: 16 set. 2020.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. **Communications of the ACM**, New York, v. 45, n. 4, p. 211–218, apr. 2002.

POSSAMAI, A. J.; SOUZA, V. G. de. Transparência e dados abertos governamentais: possibilidades e desafios a partir da lei de acesso à informação. **Administração Pública e Gestão Social**, Viçosa: MG, v. 12, n. 2, 2020. Disponível em: <https://periodicos.ufv.br/apgs/article/view/5872/5460> Acesso em: 20 out. 2021.

REICHE, K. J.; HÖFIG, E.; SCHIEFERDECKER, I. Assessment and visualization of metadata quality for open government data. *In: CONFERENCE FOR E-DEMOCRACY AND OPEN GOVERNMENT (CeDEM 2014)*, 2014. **Proceedings** [...] 2014. p. 335-346.

SADIQ, S.; INDULSKA, M. Open data: quality over quantity. **International Journal of Information Management**, v. 37, n. 3, p. 150–154, 2017.

SANT'ANNA, D. F. F. A.; SANT'ANNA, D. V. Google Meet como modalidade de ensino remoto: possibilidade de prática pedagógica. *In: CONGRESSO INTERNACIONAL DE EDUCAÇÃO E TECNOLOGIAS – CIET; ENCONTRO DE PESQUISADORES EM EDUCAÇÃO A DISTÂNCIA – EnPED*, 2020. São Carlos. **Anais** [...] São Carlos, 2020. Disponível em: <https://cietenped.ufscar.br/submissao/index.php/2020/article/view/1787>. Acesso em: 24 ago. 2022.

SCANNAPIECO, M.; CATARCI, T. Data quality under a computer science perspective. **Archivi & Computer**, San Miniato, v. 2, p. 1–15, 2002.

SILVA, A. A. P.; MONTEIRO, D. A. A.; REIS, A. O. Qualidade da informação dos dados governamentais abertos: análise do portal de dados abertos brasileiro. **Revista Gestão em Análise**, v. 9, n. 1, p. 31–47, 2020. Disponível em: <https://periodicos.unichristus.edu.br/gestao/article/view/2555> Acesso em: 30 set. 2021.

SOCRATA Open Data Server. [2021]. Disponível em: <http://open-source.socrata.com/> Acesso em: 02 abr. 2022.

VAZIRI, R.; MOHSENZADEH, M. A questionnaire-based data quality methodology. **International Journal of Database Management Systems**, v. 4, n. 2, p. 55, 2012.

WALSH, M. **House of commons science and technology committee inquiry into science communication**. London: Science Media Centre, 2016.

WAND, Y.; WANG, R. Y. Anchoring data quality dimensions in ontological foundations. **Communications of the ACM**, New York, v. 39, n. 11, p. 86–95, 1996.

WEST, M. M.; CARNEIRO, G. d. F. A quality dimension analysis of open government data in undergraduate public funding in Brazil. *In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY-NEW GENERATIONS. - ITNG 2022*, 19., 2022. **Proceedings** [...] [S.l.], 2022. p. 83–90.

WINN, J. *et al.* **Open data and the academy: an evaluation of CKAN for research data management**. Lincoln: University of Lincoln, 2013. Disponível em: <http://eprints.lincoln.ac.uk/id/eprint/9778/1/CKANEvaluation.pdf> Acesso em: 27 abr. 2021

WORLD WIDE WEB CONSORTIUM. **Xforms 1.1**. Cambridge: W3C, 2009. Disponível em: <https://www.w3.org/TR/xforms11/> Acesso em: 18 mar. 2022.

ZUIDERWIJK, A.; JANSSEN, M. The negative effects of open government data investigating the dark side of open data. *In: ANNUAL INTERNATIONAL CONFERENCE ON DIGITAL GOVERNMENT RESEARCH*, 15., 2014. **Proceedings** [...] 2014. p. 147–152.

APÊNDICE A - Distribuição da Quantidade de Erros por Atributo

(conclusão)

QT_SEMESTRE_FINANCIADO	0	0	0	0	0	0	6	0	0	6
CO_ADITAMENTO_EXT_ALUNOS	0	0	0	0	0	0	0	0	0	0
CO_CURSO	0	0	0	0	0	0	0	0	0	0
DS_CURSO	0	0	0	0	0	0	0	0	0	0
CO_TIPO_CURSO	0	0	0	0	0	0	33	0	0	33
DS_TIPO_CURSO	0	0	0	0	0	0	33	0	0	33
CO_PERIODICIDADE_CUR	0	0	0	0	0	0	51	0	0	51
DS_PERIODICIDADE_CUR	0	0	0	0	0	0	51	0	0	51
VL_SEMESTRE	0	0	0	0	0	0	0	0	0	0
NU_ANO_PROC	0	0	0	0	0	0	0	0	0	0
NU_MES_PROC	0	0	0	0	0	0	0	0	0	0
Total	0	0	0	0	385949	0	26970	0	36	388682

APÊNDICE B – Questionário

(continua)

01	Na sua opinião, os dados do registro que contém o código da mantenedora igual a 17024, o código do contrato do FIES igual a 2187276, o mês de referência do extrato de pagamento igual a 8 e o ano de referência do extrato de pagamento igual a 2019 estão corretos?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
02	Você considera que os dados do registro que contém o código da mantenedora igual a 1534, o código do contrato do fies igual a 274338, o mês de referência do extrato de pagamento igual a 5 e o ano de referência do extrato de pagamento igual a 2012 possui inconsistências?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
03	Você acredita que é possível identificar os nomes dos municípios nos dados do contrato número 2830492?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
04	Você julga que não há falha no registro que contém o município Canoas código da mantenedora igual a 314, o código do contrato 135498, o mês de referencia do extrato igual a 1 que impede a identificação do agente financeiro?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
05	Você considera que no município de Alagoinhas, código da mantenedora 2079 existem 3 alunos com deficiência?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
06	Você considera que no código do contrato igual a 2165590, número de inscrição igual a 3966632 é possível identificar que o candidato é indígena?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
07	É correto afirmar que no código do contrato igual a 2378896, código de inscrição igual a 2945086, cujo nome da instituição de ensino superior é igual a Faculdade da Amazônia Ocidental a descrição do curso é arquitetura?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
08	Na sua opinião os dados do código da instituição igual a 780, código do contrato igual a 2286193, código mantenedora igual a 14514, situação do ensino médio em escola pública igual a S permitem identificar se o aluno estudou o ensino médio na rede pública?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
09	Você considera que no código do contrato igual 2128687, código de inscrição igual a 3068960, mês de referência do extrato de pagamento, os dados do código do município da instituição de ensino e do código da cidade são redundantes?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
10	Você concorda que no município igual a Arapiraca, descrição de raça pardo temos três candidatos inscritos?	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
11	Você encontrou campo(s) utilizado(s) com propósito diferente do previsto? Exemplo de problema encontrado: campo e-mail preenchido com o nome. Em caso afirmativo, informe os nomes dos campos encontrados que foram utilizados com propósito diferente do previsto:	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
12	Você encontrou campo(s) alfabético(s) preenchido(s) com números? Exemplo de problema encontrado: campo nome preenchido com o cpf. Em caso afirmativo, informe os nomes dos campos alfabéticos que foram preenchidos com números:	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
13	Você encontrou campo(s) numérico(s) que foram preenchidos com texto? Exemplo de problema encontrado: campo telefone preenchido com o número e nome. Em caso afirmativo, informe os nomes dos campos numéricos que foram preenchidos com texto:	<input type="checkbox"/> Sim	<input type="checkbox"/> Não
14	Você encontrou campo(s) preenchido(s) com erros ortográficos? Exemplo de problema encontrado: campo no_razao_social_mantenedora está escrito de forma incorreta “educacao”. Em caso afirmativo, informe os nomes dos campos encontrados preenchidos com erros orto- gráficos:	<input type="checkbox"/> Sim	<input type="checkbox"/> Não

APENDICE B – Questionário

(conclusão)

15	Você encontrou campo(s) preenchido(s) com data com diferença maior que 100 anos em relação a data atual? Exemplos de problemas encontrados: campo dt_nascimento preenchido com 01/01/1900 ou 01/01/2100. Em caso afirmativo, informe os nomes dos campos encontrados preenchidos com data com diferença maior que 100 anos em relação a data atual:	(<input type="checkbox"/>) Sim	(<input type="checkbox"/>) Não
16	Você encontrou registros que representam o mesmo dado, porém com identificações diferentes? Exemplo de problema encontrado: nome preenchido em um registro de forma completa e em outro registro de forma abreviada. Em caso afirmativo, informe a quantidade de registros encontrados que representam o mesmo dado, porém com identificações diferentes:	(<input type="checkbox"/>) Sim	(<input type="checkbox"/>) Não
17	Você encontrou campo(s) preenchido(s) com mais de um atributo? Exemplo de problema encontrado: campo nome preenchido com o nome e o número do processo Em caso afirmativo, informe os nomes dos campos encontrados preenchidos com mais de um atributo:	(<input type="checkbox"/>) Sim	(<input type="checkbox"/>) Não
18	Você encontrou campo(s) não preenchido(s)? Exemplo de problema encontrado: campo no_razao_social_mantenedora não preenchido. Em caso afirmativo, informe os nomes dos campos encontrados sem preenchimento:	(<input type="checkbox"/>) Sim	(<input type="checkbox"/>) Não
19	Você encontrou campos(s) com nome(s) diferentes dos nomes do dicionário de dados? Em caso afirmativo, responda: Quais foram os campos que você encontrou com nomes diferentes dos nomes do dicionário de dados? O que você recomenda a ser feito para corrigir os nomes dos campos que foram encontrados com nomes diferentes dos nomes do dicionário de dados? O que você fez para corrigir os nomes dos campos que foram encontrados com nomes diferentes dos nomes do dicionário de dados?	(<input type="checkbox"/>) Sim	(<input type="checkbox"/>) Não
20	Você encontrou campos(s) com dados com caracteres indecifráveis? Em caso afirmativo, responda: Quais foram os campos que você encontrou com dados com caracteres indecifráveis? O que você recomenda a ser feito para corrigir os campos que você encontrou com dados com caracteres indecifráveis? O que você fez para corrigir os campos encontrados com dados com caracteres indecifráveis?	(<input type="checkbox"/>) Sim	(<input type="checkbox"/>) Não


```

66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
else:
    if (k2doc == 'url'):
        print('url:' + v2doc)
        inicio = v2doc[0:4]
        #print(inicio)
        if inicio == 'http':
            texturl = validarurl(v2doc)
            cont_url_valid = cont_url_valid + 1
            print(texturl)
        else:
            cont_imposs_valid = cont_imposs_valid + 1
            print('não foi possível validar, devido
            protocolo disponibilizado')
    else:
        if (k2doc == 'name'):
            if v2doc is not None:
                if v2doc == '':
                    cont_name_branco = cont_name_branco +
                    1
                    print('Nome em branco')
                else:
                    cont_name_valid = cont_name_valid + 1
                    print('name:' + v2doc)
            else:
                if (k2doc == 'format'):
                    if v2doc is not None:
                        if v2doc == '':
                            cont_format_branco =
                            cont_format_branco + 1
                            print('formato em branco')
                        else:
                            cont_format_valid =
                            cont_format_valid + 1
                            print('format:' + v2doc)
                    else:
                        if (k2doc == 'created'):
                            if v2doc is not None:
                                if v2doc == '':
                                    cont_date_branco =
                                    cont_date_branco + 1
                                    print('data em branco')
                                else:
                                    cont_date_valid =
                                    cont_date_valid + 1
                                    print('created:' + v2doc)
                else:
                    print('')
print('final da análise.....')
print('Quantidade de Recursos')
print('\n')
print(cont)
print('\n')
print('Quantidade de Url Válidas')
print('\n')
print(cont_url_valid)
print('\n')
print('Quantidade de Url Inválidas')
print('\n')
print(cont-cont_url_valid)
print('\n')
print('Quantidade de Datas Válidas')
print('\n')
print(cont_date_valid)
print('\n')
print('Quantidade de Datas Inválidas')
print('\n')
print(cont-cont_date_valid)
print('\n')
print('Quantidade de Datas em Branco')

```

```
132 print('\n')
133 print(cont_date_branco)
134 print('\n')
135 print('Quantidade de Formatos Válidas')
136 print('\n')
137 print(cont_format_valid)
138 print('\n')
139 print('Quantidade de Formatos Inválidas')
140 print('\n')
141 print(cont-cont_format_valid)
142 print('\n')
143 print('Quantidade de Formatos em Branco')
144 print('\n')
145 print(cont_format_branco)
146 print('\n')
147 print('Quantidade de Nomes Válidos')
148 print('\n')
149 print(cont_name_valid)
150 print('\n')
151 print('Quantidade de Nomes Inválidos')
152 print('\n')
153 print(cont-cont_name_valid)
154 print('\n')
155 print('Quantidade de Nomes em Branco')
156 print('\n')
157 print(cont_name_branco)
158 print('\n')
159 print('Quantidade de Descrições Válidas')
160 print('\n')
161 print(cont_description_valid)
162 print('\n')
163 print('Quantidade de Descrições Inválidas')
164 print('\n')
165 print(cont-cont_description_valid)
166 print('\n')
167 print('Quantidade de Descrições em Branco')
168 print('\n')
169 print(cont_description_branco)
170 print('\n')
171 print('Quantidade de Urls que não puderam ser validadas')
172 print('\n')
173 print(cont_imposs_valid)
174 print('\n')
```

APENDICE C – Resultado da Análise do Repositório Minha Casa Minha Vida

C:\Users\westm\PycharmProjects\pythonProjectCaptura\venv\Scripts\python.exe
 C:/Users/westm/PycharmProjects/pythonProjectCaptura/main.py
 inicio da anfilise.....

Nº Recurso: 1 Descrição em Branco format:CSV

url:http://www.geoservicos.inde.gov.br/geoserver/wms?service=WFS&version=1.0.0&request=GetFeature&typeName=MPOG:MCMV&outputFormat=C_SV
 url validada
 created:2018-06-12T05:08:08.632076
 name:Programa Minha Casa Minha Vida

Nº Recurso: ?

description:Apura despesas pagas, pelo critério valores pagos, relativas a subvenções econômicas do Programa Minha Casa, Minha Vida, programa habitacional inserido no Programa de Aceleração do Crescimento - PAC.

Os valores pagos correspondem aos documentos de pagamento emitidos no SIAFI no período e difere dos valores sacados da conta única dado que a compensação bancária de alguns documentos é feita no dia posterior.

format:XLSX

name:Subsídios e Subvenções da União - Valores Pagos - Minha Casa, Minha Vida.XLSX
 created:2015-12-16T03:08:54.885892
 url:<https://www.tesourotransparente.gov.br/ckan/dataset/68da87bf-9c20-4b90-8913-842c24d23b95/resource/d79dae65-aa71-4972-8bdb-60b7cd73f79f/download/SUBSIDIOS-E-SUBVENCOES-DA-UNIAO---VALORES-PAGOS-MCMV.xlsx>
 url validada

Nº Recurso: 3 Descrição em Branco format:GeoJSON

url:<http://www.geoservicos.inde.gov.br/geoserver/wms?service=WFS&version=1.0.0&request=GetFeature&typeName=MPOG:MCMV&outputFormat=JSON>
 url validada
 created:2018-06-12T05:08:08.632089
 name:Programa Minha Casa Minha Vida

Nº Recurso: 4

description:Apura despesas pagas, pelo critério valores pagos, relativas a subvenções econômicas do Programa Minha Casa, Minha Vida, programa habitacional inserido no Programa de Aceleração do Crescimento - PAC.

Os valores pagos correspondem aos documentos de pagamento emitidos no SIAFI no período e difere dos valores sacados da conta única dado que a compensação bancária de alguns documentos é feita no dia posterior.

format:CSV

name:Subsídios e Subvenções da União - Valores Pagos - Minha Casa, Minha Vida.CSV
 created:2022-03-01T20:27:48.537507

url:<https://www.tesourotransparente.gov.br/ckan/dataset/68da87bf-9c20-4b90-8913-842c24d23b95/resource/d1b7b1d1-eb0a-4a31-bf5b-71Se794e4sb3/download/SUBSIDIOS-E-SUBVENCOES-DA-UNIAO---VALORES-PAGOS---MCMV.csv>
url validada

Nº Recurso: 5 Descrição em Branco format:ZIP SHP

url:http://www.geoservicos.inde.gov.br/geoserver/wms?service=WFS&version=1.0.0&request=GetFeature&typeName=MPOG:MCMV&outputFormat=S_HAPE-ZIP
url validada

created:2018-06—12T05:08:08.632095
name:Programa Minha Casa Minha Vida

Nº Recurso: 6

description:Metadados relativo aos Subsidios e Subvenções da Uniao. format:PDF
name:Metadados - Subsidios e Subvenções da União -MCMV created:2015-12-13T21:02:39.032988

url:<https://www.tesourotransparente.gov.br/ckan/dataset/68da87bf-9c20-4b90-8913-842c24d23b95/resource/3c0b1b78-e184-4183-b227-6694f9d7acfa/download/Metadados---Subsidios-e-Subvencoes-da-Uniao---Valores-Pagos---PMCMV.pdf> url validada

Nº Recurso: 7 Descrição em Branco format:KML

url:<http://www.geoservicos.inde.gov.br/geoserver/wms?service=WMS&version=1.1.0&request=GetMap&layers=MPOG:MCMV&width=1024&height=768&bbox=-74,-34,-29,6&format=application/vnd.google-earth,kml>
url validada created:2018—06—12T05:08:08.632101
name:Programa Minha Casa Minha Vida final da analise.....

Quantidade de Recursos

7

Quantidade de Url Vfilidas

7

Quantidade de Url Invãlidas

0

Quantidade de Datas Validas

7

Quantidade de Datas Invalidas

0

Quantidade de Datas em Branco

0

Quantidade de Formatos Validas

7

Quantidade de Formatos Invalidas
0

Quantidade de Formatos em Branco
0

Quantidade de Nomes Validos
7

Quantidade de Nomes Inválidos
8

ANEXO A – Dicionário de Dados

(continua)

Campo	Descrição
SG_UF	Sigla da Unidade da Federação
NO_MUNICIPIO	Nome do Município
NU_CNPJ_MANTENEDORA	Número do Cadastro Nacional da Pessoa Jurídica (CNPJ) da mantenedora
NO_RAZAO_SOCIAL_MANTENEDORA	Nome da razão social da mantenedora
CO_MUNICIPIO_MANTENEDORA	Código do Município da mantenedora de acordo com o IBGE
CO_IES	Código da Instituição de Ensino Superior (IES) no sistema eletrônico e-MEC
NO_IES	Nome da Instituição de Ensino Superior - IES
CO_MUNICIPIO_IES	Código do Município da Instituição de Ensino Superior IES de acordo com o IBGE
CO_PROCESSO	Código de referência do processo de pagamento no Sistema Informatizado do FIES (SisFIES)
CO_CONTRATO_FIES	Código do contrato no Sistema Informatizado do FIES (SisFIES)
CO_ADITAMENTO	Código de identificação do aditamento do contrato de financiamento no SisFies.
CO_AGENTE_FINANCEIRO	Código do Agente Financeiro responsável pelo financiamento estudantil: (1) Banco do Brasil (2) Caixa Econômica Federal.
NO_AGENTE_FINANCEIRO	Nome do Agente Financeiro.
SG_AGENTE_FINANCEIRO	
NU_MES	Mês de referência do extrato de pagamento.
NU_SEMESTRE	Semestre de referência do extrato de pagamento.
NU_ANO	Ano de referência do extrato de pagamento.
VL_MENSALIDADE	Valor da mensalidade do curso financiado.
CO_INCRICAO	Código da inscrição do estudante no Sistema Informatizado do FIES (SisFIES).
TP_FIANCA	Tipo de fiança contratada: Normal / Convencional (N); Fundo de Garantia de Operações de Crédito Educativo – FGEDUC (F) e Solidária (S).
NU_ANO_EXERCICIO_INSC	Ano de exercício referente à inscrição do estudante no Sistema Informatizado do FIES (SisFIES).
NU_PERCENTUAL_PROUNI	Percentual da bolsa de estudo concernente ao Programa Universidade para Todos- ProUni.
NU_PERCENT_SOLICITADO_FINANC	Percentual de financiamento solicitado pelo estudante.
VL_REPASSE	Valor total do financiamento repassado à mantenedora.
DT_NASCIMENTO	Data de nascimento do estudante.
ST_DEFICIENCIA	O estudante possui necessidades especiais: Não (N) e Sim (S).
ST_ENSINO_MEDIO_ESCOLA_PUBLICA	O aluno cursou o ensino médio em escola pública: Não (N), Sim (S) e Parcialmente (P).
CO_CIDADE	Código do município de residência do estudante de acordo com o IBGE.
SG_SEXO	Sigla do sexo: (F) Feminino (M) Masculino.
DS_SEXO	Descrição de sexo.

ANEXO A – Dicionário de Dados

(conclusão)

CO_ESTADO_CIVIL	Código do estado civil: (1) Solteiro (2) Casado (3) Viúvo (4) Separado (5) Divorciado (6) União Estável.
DS_ESTADO_CIVIL	Descrição do estado civil.
SG_RACA_COR	Sigla Raça/Cor.
DS_RACA_COR	Descrição Raça/Cor (A) Amarelo (B) Branco (I) Índio (N) Negro (P) Pardo.
CO_INSCRICAO_EXT_ALUNOS	Código da inscrição do estudante no Sistema Informatizado do FIES (SisFIES).
CO_CONTRATO_FIES_EXT_ALUNOS	Código do contrato firmado junto ao Fundo de Financiamento Estudantil - FIES.
CO_IES_EXT_ALUNOS	Código da Instituição de Ensino Superior (IES) no sistema eletrônico e-MEC.
NO_IES_EXT_ALUNOS	Nome da Instituição de Ensino Superior - IES.
CO_CAMPUS	Código do campus no sistema eletrônico e-MEC.
NO_CAMPUS	Nome do campus.
VL_PERC_FINANCIAMENTO	Percentual do financiamento concedido ao estudante.
VL_MENSALIDADE_EXT_ALUNOS	Valor da mensalidade do curso financiado.
QT_SEMESTRE_FINANCIADO	Quantidade de semestre financiado.
CO_ADITAMENTO_EXT_ALUNOS	Código de identificação do aditamento do contrato de financiamento no Sistema Informatizado do FIES (SisFIES).
CO_CURSO	Código do curso no sistema eletrônico e-MEC.
DS_CURSO	Descrição do curso.
CO_TIPO_CURSO	
DS_TIPO_CURSO	
CO_PERIODICIDADE_CUR	
DS_PERIODICIDADE_CUR	
VL_SEMESTRE	Valor da semestralidade do curso financiado
NU_ANO_PROC	
NU_MES_PROC	

Fonte: Portal Brasileiro de Dados Abertos (2021).