UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA

UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO

COORDENAÇÃO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

LORENA SANTOS PEREIRA

CARACTERIZAÇÃO DA COMUNIDADE QUE UTILIZA DADOS ABERTOS GOVERNAMENTAIS SOBRE A EDUCAÇÃO BRASILEIRA

CAMPINA GRANDE - PB

2022

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Caracterização da Comunidade que Utiliza Dados Abertos Governamentais Sobre a Educação Brasileira

Lorena Santos Pereira

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Análise de Dados

Nome do Orientador
João Arthur Brunet Monteiro

Campina Grande, Paraíba, Brasil

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO CIENCIAS DA COMPUTACAO
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

LORENA SANTOS PEREIRA

CARACTERIZAÇÃO DA COMUNIDADE QUE UTILIZA DADOS ABERTOS GOVERNAMENTAIS SOBRE A EDUCAÇÃO BRASILEIRA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 11/03/2022

Prof. Dr. JOÃO ARTHUR BRUNET MONTEIRO, UFCG, Orientador

Prof. Dr. FÁBIO JORGE ALMEIDA MORAIS, UFCG, Examinador Interno

Prof. Dr. FLAVIO VINICIUS DINIZ DE FIGUEIREDO, UFMG, Examinador Externo

Documento assinado eletronicamente por **JOAO ARTHUR BRUNET MONTEIRO**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 11/03/2022, às 17:45, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **FABIO JORGE ALMEIDA MORAIS**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 11/03/2022, às 21:19, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Flavio Vinicius Diniz de Figueiredo**, **Usuário Externo**, em 14/03/2022, às 08:24, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

P436c Pereira, Lorena Santos.
Caracterização da comunidade que utiliza dados abertos governamentais sobre a educação brasileira / Lorena Santos Pereira. – Campina Grande, 2022.
99 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2022.
"Orientação: Prof. Dr. João Arthur Brunet Monteiro".
Referências.

1. Dados Abertos Governamentais. 2. Dados Educacionais. 3. Caracterização. 4. Tecnologia Cívica. I. Monteiro, João Arthur Brunet. II. Título.

CDU 004.632(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA MARIA ANTONIA DE SOUSA CRB 15/398

Resumo

Ao longo dos anos, sociedade civil e entidades governamentais têm se esforçado para entender o ecossistema de tecnologias cívicas e a utilização de dados abertos na transparência governamental. Alguns trabalhos focam na avaliação da conformidade dos dados disponíveis com os padrões que precisam seguir, e outros se aprofundam no entendimento de um tema específico, como os Dados Abertos Governamentais - DAGs da área de educação. O Brasil publica dados educacionais que descrevem desde a condição de escolas da educação básica até indicadores de qualidade da educação superior, alguns temas possuem conjuntos de dados com referência de 1995 até 2020. Nesse cenário, é possível entender que existe um empenho para o acompanhamento e facilitação do uso desse recurso, mas ainda não está descrito quais são as características com relação às pessoas, organizações e projetos de software nesse ecossistema. Por isso, o objetivo dessa pesquisa foi caracterizar a comunidade que utiliza DAGs brasileiros relacionados à educação, entendendo a sua dinâmica a partir de informações sobre 217 repositórios de projetos de software armazenados no *GitHub* e da experiência de 38 participantes dessa comunidade. Com os dados coletados foi possível entender que a comunidade se desenvolve, em sua maior parte, a partir de projetos pessoais (85%), porém existem também faculdades, centros de pesquisa, canais de mídia/comunicação e instituições governamentais. Nesses projetos, o *Jupyter Notebook* se mostra uma tecnologia importante. Com relação aos respondentes, uma das motivações relatadas é utilizar o aprendizado com a análise desses dados para incidir no próprio contexto educacional, auxiliando no trabalho de professores e gestores. Já os problemas relatados pelos participantes, reforçam a ideia da necessidade de conhecimento técnico para a manipulação desses dados. Com isso, esperamos ajudar órgãos governamentais na prevenção das inconsistências relatadas e no planejamento de ações para promoção da participação cidadã de forma mais democrática e inclusiva, juntamente a outros integrantes dessa comunidade.

Abstract

Over the years, civil society and government entities have struggled to understand the civic technologies ecosystem and the use of open data in government transparency. Some works focus on assessing the conformity of available data with the standards they need to follow, and others deepen the understanding of a specific topic, such as the Open Government Data - (Dados Abertos Governamentais) DAGs in the education area. Brazil publishes educational data about the condition of basic education schools up to higher education quality indicators, some themes have data sets from 1995 to 2020. In this scenario, it is possible to understand that there is an effort to monitor and facilitate the use of this resource, but it is not yet described what the characteristics in relation to people, organizations and software projects in this ecosystem are. Therefore, the purpose of this research was to characterize the community that uses Brazilian DAGs related to education, understanding its dynamics from information on 217 software projects repositories stored on GitHub and the experience of 38 participants of this community. With the collected data, it was possible to understand that the community develops, for the most part, personal projects (85%), but there are also faculties, research centers, media/communication channels and government institutions. In these projects, Jupyter Notebook proves to be a important technology. With regard to the respondents, one of the motivations reported is to put the knowledge from the analysis of these data to use on the educational context itself, helping the work of teachers and managers. The problems reported by the participants reinforce the idea of the need for technical knowledge to manipulate these data. Hereby, we hope to help government agencies to prevent reported inconsistencies and plan actions to promote citizen participation in a more democratic and inclusive manner, along with other members of this community.

Agradecimentos

Eu agradeço sobretudo à espiritualidade que me permitiu vida o suficiente para imaginar perseguir, alcançar, viver e finalizar esse sonho. Que me permitiu ultrapassar, todas as vezes, o pensamento enraizado na minha carne de *"para quê você está se esforçando tanto se nunca vai ser boa o suficiente?"*. E principalmente, por ter me permitido essa vida, esse sopro recorrente de vida, através das **pessoas** no meu caminho. Muito obrigada do fundo do meu coração, de todo o meu espírito.

Desde de Matheus que me enviou o anúncio da vaga no Laboratório Analytics, João Arthur e Nazareno que me responderam o e-mail sobre como procurar emprego na cidade, Vini que me ajudou nos tramites para finalização do emprego na época, Priscila e sua família, que traziam tanta alegria e acalanto através de suas gargalhadas, pão com queijo de manteiga bem quentinho e café. E outras tantas pessoas incríveis que encontrei nessa aventura chamada Campina Grande, mas sem esquecer daquelas com quem partilho a vida desde as aventuras passadas. Muito obrigada a todes pelo acolhimento, oportunidade, apoio e ensinamentos durante todo esse período.

Agradeço a minha família, que tiveram que ouvir tantas vezes de mim “preciso trabalhar no mestrado”, e mesmo sem entender muito bem, me apoiavam e diziam - "Oxe! Esse negócio não acaba não?". Então pessoal... “foi sobre isso...".

A João Arthur pela paciência, generosidade e ensinamentos para construção e finalização desse trabalho, em todo o processo de orientação.

A todos os educadores que tive ao longo da vida por contribuir imensamente para a minha formação acadêmica, profissional e pessoal.

A todas as pessoas que acreditam e trabalham pela educação brasileira, que me possibilitaram mais uma formação pública, gratuita, de extrema qualidade e de imensurável valor. Muito obrigada à COPIN, ao Laboratório Analytics, ao CEEI, UFCG, UNEB, CAPES e CNPq. Que possamos fortalecer e ampliar os acessos, principalmente para as pessoas cerceadas pelas estruturas da nossa sociedade.

"cheguei da caçada e fui ao curral
saber se meu boi laranja tinha chegado
encontrei Boiadeiro amuntado
com uma corda na mão
- cadê meu boi, Boiadeiro?
pergunta seu Damião.
- Seu boi me viu, correu
correu não pude pegar...
Se meu patrão quer seu boi
vá lá no mato buscar!
Ô boiadeiro, ô boiadeiro, ô boiadeiro
cadê meu boi, boiadeiro?"
(Cantiga ensinada por Maçu e que a gente cantava na maior gaiatisse
- A benção às nossas mais velhas. Elas tiveram a audácia de nos trazer até aqui)

Conteúdo

Lista de Símbolos

OGP - *Open Government Partnership - Parceria para Governo Aberto*

API - *Application Programming Interface - Interface de Programação de Aplicativos*

DAG - *Dados Abertos Governamentais*

INEP - *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*

MEC - *Ministério da Educação*

ENEM - *Exame Nacional do Ensino Médio*

IDEB - *Índice de Desenvolvimento da Educação Básica*

CGEE - *Centro de Gestão e Estudos Estratégicos*

CNPq - *Conselho Nacional de Desenvolvimento Científico e Tecnológico*

FNDE - *Fundo Nacional de Desenvolvimento da Educação*

SEDUC - *Secretaria Estadual da Educação*

SIMEC - *Sistema Integrado de Monitoramento, Execução e Controle*

ENADE - *Exame Nacional de Desempenho dos Estudantes*

CPC - *Conceito Preliminar de Curso*

IDD - *Indicador de Diferença entre os Desempenhos*

IGC - *Índice Geral de Cursos*

PME - *Programa Mais Educação*

SISTEC - *Sistema Nacional de Informações da Educação Profissional e Tecnológica*

SIOPE - *Sistema de Informações sobre Orçamentos Públicos em Educação*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Dado aberto é todo dado que, sem restrição de uso, pode ser manipulado por qualquer pessoa através de meios automatizados. A aplicação desse conceito na transparência pública tem sido almejada em muitas vertentes. Orçamento federal, utilização de benefícios provenientes de cargos públicos e indicadores socioeconômicos, são exemplos de assuntos abordados tanto em projetos de pesquisa, a exemplo de Beghin, Zigoni et al. 2014 e Oliveira et al. 2016, como em iniciativas da sociedade civil, no caso dos projetos Serenata de Amor[1](Serenata de Amor 2016) e Datapedia[2](Datapedia 2015).

Entender que *dado aberto* deve estar disponível para utilização por qualquer pessoa não significa que ele deve cumprir apenas o requisito de ser um *dado público*. Enquanto os dados públicos são aqueles que não estão sujeitos a um controle de acesso (Tauberer 2014), os dados abertos devem ir além, através da garantia de que "qualquer pessoa pode acessar, usar, modificar e compartilhar livremente para qualquer finalidade (sujeito, no máximo, a requisitos que preservem a proveniência e a abertura)" (Open Knowledge Foundation s.d.). Quando esse conceito é transitado para o contexto de transparência pública, onde os dados são especificamente produzidos ou comissionados pelo governo ou por entidades por ele controladas, é quando se chega aos chamados Dados Abertos Governamentais - DAGs (Open Knowledge Foundation s.d.).

A forma simplificada que Eaves 2009 encontrou de contextualizar as implicações dos

[1]https://serenata.ai/ - Projeto que utiliza ciência de dados para analisar os gastos de deputados federais e senadores no que diz respeito aos reembolsos pela Cota para Exercício da Atividade Parlamentar (CEAP).

[2]https://datapedia.info/mapa - Plataforma que permite a visualização de informações socioeconômicas sobre 5.570 municípios brasileiros.

Dados Abertos Governamentais, enquanto participava da Conference for Parliamentarians: Transparency in the Digital Era (Conferência para Parlamentares: Transparência na Era Digital) em 2009, foi apresentar o que ele chamou de *"As Três Leis dos Dados Abertos Governamentais"*:

> "1) Se não puder ser encontrado ou indexado na Internet, não existe; 2) Se não estiver disponível em formato aberto e legível por máquina, não pode engajar os cidadãos; 3) Se uma estrutura legal não permite que ele seja reaproveitado, ele não dá aos cidadãos possibilidade de compartilhamento e mobilização sobre o assunto."(Eaves 2009) *(Tradução nossa)*[3]

Já os 8 Princípios dos Dados Abertos Governamentais visam esmiuçar todos os requisitos necessárias para disposição desse recurso, eles defendem que os dados devem ser: completos, estar no nível mais granular possível, atuais, acessíveis, processáveis por máquina, com acesso sem necessidade de registro, disponíveis em formato não proprietário e sob licenças livres - e além disso, a conformidade com os princípios precisa ser passível de revisão (Tauberer 2014).

A comunidade em torno dos DAGs é formada por instituições do governo e pela sociedade civil, seu engajamento pode ser visto através de eventos que visam criar soluções tecnológicas explorando dados abertos governamentais como *Hackathons*[4] e *Hackfests*[5]. Além disso, temos a construção de portais de transparência ao redor do mundo, a exemplo de países como Singapura[6], Reino Unido[7], França[8], Estados Unidos[9] e Brasil[10], e a formação da Parceria para Governo Aberto (Open Government Partnership - OGP)[11]

[3]Texto original: 1) If it can't be spidered or indexed, it doesn't exist; 2) If it isn't available in open and machine readable format, it can't engage; 3) If a legal framework doesn't allow it to be repurposed, it doesn't empower.

[4]Hackathon Contas Públicas - Maratona de programação sobre compras públicas organizada pela Secretaria da Fazenda e Planejamento do Estado de São Paulo.

[5]http://hackfest.com.br/ - Evento com o objetivo de criar projetos para combater a corrupção através do uso de tecnologia. É realizado há mais de 3 anos com edições em diversos estados brasileiros, fazendo em 2019 a sua primeira edição internacional, na Colômbia.

[6]https://data.gov.sg

[7]https://data.gov.uk

[8]www.data.gouv.fr

[9]https://www.data.gov

[10]http://dados.gov.br

[11]https://www.opengovpartnership.org

A OGP foi criada em 2011 e conta com a participação de 78 países e 76 jurisdições locais (Open Government Partnership 2022). Seus participantes compartilham o objetivo de implantar uma administração pública atuante em ações de transparência, combate à corrupção, incentivo à participação social e desenvolvimento de novas tecnologias (Ministério da Transparência e Controladoria-Geral da União 2018). De acordo com o calendário da OGP, cada país participante é responsável pela organização de um plano de ação que deve ser implementado em um determinado período. O Brasil, que é co-fundador e participa desde 2011, lançou em dezembro de 2021 o 5º Plano de Ação Nacional em Governo Aberto, que conta com 12 compromissos, dentre eles: Meio Ambiente, Floresta e Dados Abertos; Direitos Humanos e Dados Abertos; Cadeias Agropecuárias e Dados Abertos; e Participação Social para Melhoria dos Dados Eleitorais Abertos. O plano foi construído a partir de oficinas de cocriação realizadas "em 72 encontros virtuais com o envolvimento de 141 pessoas, representantes de 79 instituições, sendo 41 organizações da sociedade civil e 38 órgãos e entidades da Administração Pública" (Controladoria-Geral da União 2021).

Olhando alguns dos resultados da implementação do primeiro plano de ação brasileiro é possível ter dimensão da importância das atividades e esforços relacionados com os compromissos para governo aberto, principalmente no que diz respeito à prestação de contas públicas:

> "Entre as iniciativas implementadas no 1º Plano de Ação, destacam-se: o Sistema Federal de Acesso à Informação, que proporcionou ao Governo Federal o ambiente adequado para a implementação da Lei de Acesso à Informação (LAI), a reestruturação do Portal da Transparência, a criação da Infraestrutura Nacional de Dados Abertos (INDA) e do Portal Brasileiro de Dados Abertos."(Controladoria-Geral da União 2021).

Em *Guia do Governo Aberto para Céticos*, Open Government Partnership 2018 reúne resultados de como a aplicação de práticas que visam dar protagonismo a participação cidadã e a transparência da administração pública traz resultados em diversas áreas, como saúde, educação, infraestrutura, combate à corrupção e garantia da eficiência no gasto de recursos públicos.

A exemplo de Gonçalves 2014, que estudou os efeitos da aplicação de um modelo orça-

mentário participativo em municípios brasileiros entre os anos de 1990 e 2004. Essa forma de processo orçamentário permite aos cidadãos negociar diretamente com o governo sobre a alocação de recursos para o município e a prioridade desses investimentos. Como resultado a autora encontrou um padrão onde, diante do modelo participativo, os municípios tendem a investir mais em saúde e saneamento do que com o modelo não participativo - 2 a 3 pontos percentuais, o que equivale de 20 a 30% da média de participação orçamentária dessa categoria no início do recorte temporal da pesquisa (Gonçalves 2014). Assim como, uma redução na mortalidade infantil entre 1 e 2 bebês para cada 1.000 bebês residentes.

E Chen e Ganapati 2018 que fizeram uma metanálise de estudos publicados entre 1990 e 2017, que avaliaram o impacto da transparência pública na corrupção. Os autores descobriram que existe um impacto estatisticamente significativo, mesmo que pequeno, na redução da corrupção. Onde "um aumento de 100% nos esforços de transparência estaria, em média, correlacionado com a redução da corrupção governamental em 2,2% (com um intervalo de confiança de 95% da redução entre 1,9% e 2,6%)" (Chen e Ganapati 2018) *(tradução nossa)*.

Vemos então que o desenvolvimento do ecossistema de governo aberto e dos dados abertos governamentais, contribui em diversos aspectos de vida em sociedade: melhor aplicação de recursos públicos, melhor oferta de serviços públicos, mobilização social, participação e fiscalização cidadã. E que ainda existe potencial para evolução desses benefícios. Nesse sentido, provedores e consumidores de DAGs giram em torno desse recurso e convivem nesse ecossistema. Mas qual o conhecimento sistematizado que temos sobre essa essa comunidade?

Esforços têm sido feitos no entendimento dessa questão e avançaram o conhecimento sobre projetos dessa natureza, a nível mundial e nacional. Attard et al. 2015 fizeram uma revisão sistemática da literatura sobre o desenvolvimento de iniciativas com dados abertos governamentais. Como resultado trouxeram uma visão geral sobre conceitos, terminologias, desafios e possíveis soluções, do ponto de vista cultural e técnico.

Oliveira et al. 2016 fizeram uma análise mais específica, considerando portais de dados abertos governamentais no Brasil. Os autores afirmam que nesse cenário ainda é preciso resolver problemas como a implantação de um vocabulário comum para dados do mesmo contexto mas provenientes de órgãos diferentes. Outro problema relatado é a ausência de metadados capazes de gerar o entendimento completo dos dados disponíveis nesses portais.

Alguns trabalhos focam na avaliação da conformidade dos dados disponíveis com os princípios que eles deveriam seguir. Outros se aprofundam mais no entendimento de um tema específico desses dados, como os dados abertos abertos governamentais da área de educação.

Schalkwyk, Willmers e Czerniewicz 2014, por exemplo, fizeram um estudo de caso para examinar a oferta, uso e possíveis impactos de dados abertos de universidades sul-africanas. Como resultado entenderam que existe a utilização por parte de planejadores universitários e pesquisadores do ensino superior, porém com pouca frequência, e que esses atores expressaram a necessidade de dados abertos mais ricos e granulares. Além disso, destacam o papel de intermediários de dados abertos, que são atores responsáveis por aumentar a acessibilidade e utilidade dos dados a partir da resolução de problemas em suas publicações originais.

"Quais dados abertos educacionais estão disponíveis para a sociedade?" é a pergunta que Penteado e Isotani 2017 procuram responder em seu trabalho, considerando o contexto brasileiro. Para isso, a partir de uma pesquisa no Portal Brasileiro de Dados Abertos filtrando pelos órgãos MEC - Ministério da Educação e INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, fizeram uma caracterização das fontes, periodicidade e assuntos cobertos pelos dados educacionais disponíveis. Os órgãos usados na filtragem foram criados em 1930 e 1937, respectivamente, e são considerados estruturais no contexto educacional brasileiro. A partir de 35 conjuntos de dados e 325 arquivos, os autores observaram o crescimento na publicação dos dados ao longo dos anos. Notaram também que o censo da educação básica, censo do ensino superior e o SAEB - Sistema de Avaliação da Educação Básica são conjunto de dados com bases desde 1995 e com frequência anual. Além deles, ENEM - Exame Nacional do Ensino Médio, FIES - Fundo de Financiamento ao Estudante do Ensino Superior, estrutura do ensino superior, estrutura do ensino básico, PROUNI - Programa Universidade para Todos, matrículas do ensino superior e ENADE - Exame Nacional de Desempenho de Estudantes são conjuntos com mais de 10 anos de publicações consecutivas (Penteado e Isotani 2017).

Já Santos, Ferreira e Miranda 2017, fizeram uma revisão da literatura sobre trabalhos que utilizaram dados abertos no contexto educacional, com destaque para os objetivos pedagógicos e origens das publicações. Os autores agruparam os trabalhos analisados nos seguintes objetivos: mineração de dados aplicados a dados abertos educacionais; análise qualita-

tiva; teórico (que descrevem métodos e desafios sobre a utilização dos dados); sistemas para apoio à aprendizagem colaborativa e visualização dos dados. Os autores também destacam o quanto ainda é um quantitativo pequeno de iniciativas e sobre como ainda existe um potencial para evolução da utilização desse recurso.

Até aqui, conseguimos entender que existem esforços para o acompanhamento e facilitação do uso de DAGs educacionais, que o governo brasileiro tem um compromisso com a facilitação desse recurso e que na educação brasileira existe uma diversidade considerável de assuntos representados por esses dados, assim como órgãos centrais que fazem a publicação dos mesmos. O que ainda não sabemos é quais são as características com relação às pessoas, organizações e projetos ao redor desse recurso. Não temos um mapeamento sobre forma de organização ao redor desses dados, quais projetos de software para além daqueles publicados academicamente são desenvolvidos, onde estão localizados, quais são as bases e tecnologias utilizadas ou se existem problemas particulares do tema educacional. Para que a partir desse entendimento, seja possível orientar esforços no fortalecimento e democratização dessa participação cidadã, que pode não ser tão acessível para cidadãos que não tem entendimento aprofundado na manipulação de dados, porém impacta tão diretamente na vida de todos.

Por isso, o objetivo dessa pesquisa foi caracterizar a comunidade que utiliza os dados abertos governamentais brasileiros relacionados à educação, entendendo a sua dinâmica a partir de informações sobre os repositórios dos projetos de software e da experiência de pessoas participantes dessa comunidade. Tratou-se de uma pesquisa exploratória dividida em duas partes:

1. Descrição das características dos projetos, como autores, rede de contribuição, fonte dos dados, regionalidade, tecnologias utilizadas, popularidade, temporalidade, objetivo dos projetos e trabalhos realizados;

2. Descrição das motivações na realização dos projetos, impactos percebidos, problemas encontrados, formas e canais de compartilhamento de soluções.

Para a primeira parte do trabalho, realizamos uma extração de dados sobre projetos armazenados na plataforma de versionamento e compartilhamento de arquivos, *GitHub*. Para a pesquisa dos repositórios foram usadas palavras chaves definidas a partir da análise dos temas e conjunto de dados disponíveis em sites de referência sobre a educação brasileira.

Após aplicar critérios de aceitação e exclusão, restaram 217 repositórios de projetos de software para categorizar manualmente com relação a informações não disponíveis de forma automática pela plataforma.

Para a segunda parte coletamos as respostas de 38 integrantes dessa comunidade, a partir de uma pesquisa de opinião que foi divulgada em redes sociais, grupos e perfis diretamente relacionados a DAGs. Consideramos uma amostra satisfatória entendendo a natureza exploratória deste trabalho e também devido ao alto nível de detalhe fornecido pelos respondentes para as questões abertas, as quais examinamos a partir da aplicação da metodologia de Análise Textual Discursiva Moraes e Galiazzi 2007.

Os repositórios permitiram, por exemplo, o olhar sobre questões quantitativas e norteadoras, como forma de organização em torno desses projetos, sua regionalidade e seus principais objetivos. Entendemos que a comunidade se desenvolve, em sua maior parte, a partir de projetos pessoais (85%), porém existem também instituições da sociedade civil como faculdades, centros de pesquisa e canais de mídia/comunicação, como a Gênero e Número, e instituições governamentais, como a prefeitura de São Paulo. E que os projetos são realizados por pessoas que estão presentes em 16 estados brasileiros e em todas as regiões do país.

Os projetos realizados, majoritariamente, tiveram como objetivo informar determinado público sobre uma nuance dos dados em questão, se preocupando com uma forma agradável de comunicar aquele contexto. O segundo objetivo mais frequente foi o educacional, quando o repositório está relacionado com uma atividade de uma disciplina ou projeto de um curso, graduação ou pós graduação, ou seja, temos a utilização de DAGs educacionais para a formação de pessoas, seja com a aplicação de algoritmos preditivos ou construção de visualizações. E por fim, tivemos o objetivo de aumentar o acesso aos dados, expandindo assim o público que pode fazer uso deles, o que é imprescindível para a construção da cultura de participação cidadã e governo aberto.

Com relação aos problemas encontrados, pudemos ver os relatos já conhecidos na comunidade de dados abertos governamentais de forma geral, sobre falta de padronização e dados incompletos. O que chamou a atenção nessas experiências foi a inconsistência de informações que são de fontes diferentes mas deveriam descrever o mesmo assunto e a dificuldade na navegação pelo site que armazena o dado. Mas também devemos destacar o não enfren-

tamento de problemas por 9 dos respondentes (23%), e o destaque feito sobre o impacto positivo de tutoriais de utilização dos dados.

A partir dos resultados alcançados esperamos que o entendimento desse contexto possa proporcionar benefícios diretos para organizações governamentais e sociedade civil. As instituições públicas responsáveis pelos dados podem usar os problemas relatados para a correção e prevenção de inconsistências, inclusive em outros contextos. Assim como, investir em infraestrutura e arquitetura da informação dos sites, para evitar os problemas com dificuldade na navegação e falha pelo tempo de resposta ser muito demorado.

Entender o perfil de pessoas e suas experiências pode ajudar no planejamento de ações para o fortalecimento dessa comunidade. Sabendo que para lidar com os problemas relatados é preciso um conhecimento em extração e tratamento de dados e que essa não é uma formação básica nem acessível para a maioria da população, investir na capacitação e acesso de, pelo menos, trabalhadores da área de educação pode ser um fator primordial para conseguirmos uma participação cidadã mais efetiva, principalmente pensando nas estruturas educacionais com menores recursos e grande contingente de alunos.

Ao entender as formas de compartilhamento utilizadas, a comunidade pode se valer também da criação de um canal central e categorizado para o compartilhamento de soluções e troca de experiências, inclusive entre fornecedores e consumidores de DAGs educacionais.

Por fim, entendemos as realizações que a comunidade vem alcançando, suas dificuldades e um campo amplo para evolução e melhorias, para o qual acreditamos que a construção coletiva é fundamental. Por isso, e entendendo as relações com os trabalhos de Araújo 2017 e Santos, Ferreira e Miranda 2017, fazemos a recomendação de fortalecimento dessa comunidade e seus benefícios através do investimento em eventos que possam cultivar a ideia de colaboração entre governo, secretarias de educação e diversos outros atores da sociedade, como os *Hackathons* e *Hackfests*, porém voltando o olhar para as dinâmicas da rede pública municipal de ensino, que são responsáveis por 48,4% dos estudantes brasileiros (Cristaldo 2021).

Esta dissertação está organizada em 5 capítulos. Esse capítulo introduz o trabalho. O Capítulo 2 apresenta o referencial teórico para o entendimento desse trabalho. Na sequência, o Capítulo 3 apresenta os trabalhos correlatos. Já o Capítulo 4 descreve os detalhes do método de pesquisa aplicado. Os resultados obtidos são apresentados no Capítulo 5 e por fim, o Capítulo 6 apresenta as conclusões.

Capítulo 2

Fundamentação Teórica

2.1 Governo Aberto

> "Governo aberto se refere a uma nova visão da Administração Pública, que promove projetos e ações voltados ao aumento da transparência, à luta contra a corrupção, ao incentivo à participação social e ao desenvolvimento de novas tecnologias que tornem os governos mais responsáveis por suas ações e preparados para atender às necessidades dos cidadãos." (Ministério da Transparência e Controladoria-Geral da União 2018)

Esse é um conceito que vem sendo difundido em 78 países, desde 2011, através da formação da Parceria para Governo Aberto (Open Government Partnership - OGP) (Open Government Partnership 2022). A participação nessa iniciativa é apoiada por um plano de ação desenvolvido e implementado por cada membro. Cada plano de ação tem um prazo de 2 anos. No Brasil estamos na 5ª edição[1], feita em 2021 pelo Ministério da Transparência e Controladoria-Geral da União juntamente a outros órgãos governamentais e organizações da sociedade civil. No documento do plano de ação de 2018 os princípios para governo aberto são apresentados por Ministério da Transparência e Controladoria-Geral da União 2018 como:

- Accountability (Prestação de contas e responsabilização): Refere-se ao estabeleci-

[1]Mais informações sobre a execução do plano de ação atual podem ser encontradas no site https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao

mento de normas e mecanismos para tornar as instituições governamentais mais responsáveis e transparentes sobre suas ações;

- Participação social: Envolvimento dos cidadãos nos processos de criação de políticas públicas e de espaços de troca de conhecimento com as instituições governamentais;
- Transparência: Dispor de informações sobre planos de ação, fontes de dados, atribuições e prestação de contas, estimulando o controle e a participação social;
- Tecnologia e Inovação: Entender e aplicar novas tecnologias que podem auxiliar governos abertos em suas atribuições.

O incremento da confiança nos governos, fortalecimento das instituições, combate a corrupção, promoção da inovação e da cidadania são alguns dos benefícios na adoção dessa nova forma de governar. As implicações positivas do governo aberto são acompanhadas em diversos países. Em *Guia do Governo Aberto para Céticos*, a OGP descreve resultados de ações governamentais e da sociedade civil que visam unir participação cidadã e a transparência da administração pública trazendo resultados em diversas áreas, como saúde, educação, infraestrutura, combate à corrupção e garantia da eficiência no gasto de recursos públicos (Open Government Partnership 2018).

Ao observar os resultados do primeiro ano de implementação do plano de ação brasileiro, por exemplo, são notórios os impactos das atividades e esforços relacionados com os compromissos para governo aberto. As ações trouxeram resultados, como o Sistema Federal de Acesso à Informação, a reestruturação do Portal da Transparência, a criação da Infraestrutura Nacional de Dados Abertos (INDA) e do Portal Brasileiro de Dados Abertos (Controladoria-Geral da União 2021).

Um outro marco no cenário brasileiro foi a Lei nº 12.527 de 18 de novembro 2011, que regulamentou o acesso a informação dos Poderes Executivo, Legislativo, Judiciário e do Ministério Público (Brasil 2011). A partir dela, conforme descreve a seção I da lei, qualquer pessoa interessada poderá apresentar um pedido de acesso a informação aos órgãos e entidades especificados no Art. 1º (Brasil 2011). A criação e aplicação dessa lei incentivou ainda mais a discussão a cerca de um outro componente importante do governo aberto: Os Dados Abertos Governamentais.

2.1.1 Dados Abertos Governamentais

Attard et al. 2015 faz uma distinção entre 3 termos essenciais ao abordar o tema Governo Aberto: Dados públicos, dados abertos e dados abertos governamentais. Onde podemos entender que (i) dados públicos são aqueles disponíveis gratuitamente sem limitação de acesso; (ii) dados abertos são aqueles que podem ser usados, reutilizados e redistribuídos por qualquer pessoa em formato não proprietário e legível por máquina (Dietrich et al. 2009) e (iii) dados abertos governamentais são os dados produzidos ou comissionados pelo governo, ou por entidades controladas por ele, que estão disponíveis de acordo a definição de dados abertos (Open Knowledge Foundation s.d.).

Para satisfazer essa definição, as organizações responsáveis devem seguir os chamados 8 Princípios dos Dados Abertos Governamentais, conforme elenca o Portal Brasileiro de Dados Abertos 2020. O portal descreve que os dados devem ser:

1. Completos: Estarem disponíveis de forma completa, sem limitações;
2. Primários: Disponíveis na maior granularidade possível;
3. Atuais: Disponíveis o mais rápido possível de forma a preservar o seu valor;
4. Acessíveis: Disponíveis para o maior público possível com os propósitos mais variados;
5. Processáveis por máquina: Disponíveis em formato que permitam a leitura automatizada;
6. Acesso não discriminatório: Disponíveis sem necessidade de registro ou autenticação;
7. Formatos não proprietários: Disponível em formato que não permita o acesso por uma entidade ou organização exclusiva;
8. Licenças livres: Disponíveis sem estarem sujeitos a restrição por direito autoral ou patente.

Esforços foram realizados no entendimento, disposição e incentivo da utilização dos dados abertos governamentais, mas sabemos que ainda existem dificuldades e que elas podem ser recorrentes na rotina das pessoas e organizações interessadas nesse recurso

(Oliveira et al. 2016), (Pinho 2018). Recentemente um relatório feito pelo projeto Parlametria[2] identificou 19 barreiras no acesso aos dados do congresso nacional. O relatório organiza as barreiras em: graves, críticas e pontos a melhorar. Além de identificar problemas, descreve o impacto e sugere soluções (Parlametria 2019). Algumas das barreiras relatadas foram:

- Dados de votações nas comissões legislativas não estão em formato aberto (Câmara);
- O texto dos discursos nas comissões legislativas não está em formato aberto (Câmara e Senado);
- A movimentação de cargos e licenças não está publicada em formato aberto (Câmara e Senado);
- Não há notas fiscais de despesas digitalizadas (Senado);
- Não há ferramenta de busca de dados abertos (Câmara e Senado).

Exemplos de soluções são: disponibilização dos dados através de APIs já existentes para os órgãos e adoção de ferramentas *open source* para gerenciamento de dados como o CKAN, que facilita a publicação e compartilhamento dos dados (CKAN 2018), inclusive é utilizada no Portal Brasileiro de Dados Abertos.

O Portal Brasileiro de Dados Abertos é uma ferramenta que foi disponibilizada pelo governo para centralizar dados de natureza pública, e como falado anteriormente foi fruto do primeiro plano de ação para governo aberto no pais. No portal encontramos diversos temas, como cultura, segurança, agricultura, habitação, saneamento, relações internacionais, saúde e educação - que foi o tema delimitador na caracterização feita por esse trabalho. No total, são mais de 10 mil conjuntos de dados publicados vinculados aos respectivos órgãos responsáveis pelo fornecimento da informação (Portal Brasileiro de Dados Abertos 2020).

A principal organização sobre dados educacionais no portal é o INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. O órgão foi criado em 1937 e passou a fazer parte do MEC - Ministério da Educação em 1997, tornando-se assim o órgão responsável por avaliações e exames educacionais, pesquisas estatísticas e indicadores, além da

[2]O projeto atua com inteligência de dados para ação cidadã e é uma realização das organizações Dado Capital, Open Knowledge Brasil e Laboratório Analytics com apoio do Instituto Betty & Jacob Lafer e Brasil.io - https://parlametria.org/home

gestão do conhecimento sobre estudos educacionais (Inep, Ministério da Educação 2021). A Tabela 2.1 apresenta alguns dos temas dos dados disponíveis no site do INEP.

Temas, siglas e acrônimos	**Descrição**
Censo Escolar	Censo Escolar da Educação Básica. Realizado em conjunto com as Secretarias de Educação estaduais e municipais, incluindo escolas públicas e privadas
ENCCEJA	Exame Nacional para Certificação de Competências de Jovens e Adultos
Censo da Educação Superior	Pesquisa sobre as instituições de educação superior, seus cursos, discentes e docentes
SAEB	Sistema de Avaliação da Educação Básica
ENEM	Exame Nacional do Ensino Médio
IDEB	Índice de Desenvolvimento da Educação Básica. Índice criado em 2007 e que é calculado a partir de dados do Censo Escolar e das médias de desempenho do SAEB
ENADE	Exame Nacional de Desempenho dos Estudantes

Tabela 2.1: Alguns temas de DAGs sobre a educação brasileira

No site do MEC também também está disponível alguns conjuntos de dados, como bolsas concedidas para o PROUNI - Programa Universidade para Todos e dados sobre o FIES - Fundo de Financiamento Estudantil.

2.2 Plataformas de Armazenamento e Versionamento de Arquivos

Git é um sistema de controle de versão gratuito e de código aberto (Git s.d.). Esse é um sistema utilizado pela comunidade desenvolvedora de software pela necessidade de gerenciar, integrar e reverter seus arquivos de trabalho, principalmente em um contexto de colaboração entre diversos programadores, trabalhando em diversas funcionalidades.

O *GitHub* é uma plataforma online que hospeda repositórios *Git* e permite a interação entre usuários de todo o mundo. Segundo o (Git 2014) *GitHub* é o maior serviço de hospedagem de repositórios *Git*. Na descrição atual da plataforma, são mais de 73 milhões de usuários e mais de 200 milhões de repositórios. *GitLab* e *Bitbucket* também são alternativas para os desenvolvedores, atualmente o primeiro possui 30 milhões de usuários

(GitLab 2022) e o segundo comemorava em 2019 ter atingido a marca de 10 milhões de usuários (Bitbucket 2019), até a data da escrita desta dissertação não constava na seção Sobre do site do Bitbucket a informação atualizada sobre seus usuários.

O *GitHub* tem aumentado a sua popularidade ano após ano, e só em 2021 teve 16 milhões de novos usuários (GitHub, Inc. 2021). Além disso, é considerado um portfólio, por vezes requisitado para vagas de emprego na área de tecnologia. Dentro da comunidade brasileira ele também se tornou uma ambiente de construção coletiva sobre materiais de estudo (PyLadies Paraíba 2021), (Pizza De Dados 2021), divulgação de vagas de trabalho (Onde Codar em Salvador 2021) e dicas sobre entrevistas em processos seletivos (Repositório de vagas Front-End 2015).

Pela popularidade e diversidade de projetos, o *GitHub* se mostra um local com potencial para realizar a busca por projetos de software voltados para utilização de DAGs educacionais. Além disso, a plataforma contém uma API - Application Programming Interface (Interface de Programação de Aplicativos), que é um conjunto de rotinas prontas para serem utilizadas via requisições HTTP permitindo assim a interação com o sistema e/ou seu banco de dados. Esse é um recurso importante pois possibilita a pesquisa e extração de dados sobre repositórios e usuários da plataforma de forma automatizada.

2.2.1 Dinâmica do *GitHub*

Como já é sabido, o *GitHub* é é uma local para fazer o versionamento e compartilhamento de arquivos através de repositórios *Git*. Para a criação de repositórios é preciso criar uma conta de usuário adicionando algumas informações, como e-mail, nome de usuário e senha. A customização do perfil de usuário também permite a adição de uma foto, redes sociais, localização e uma mini descrição. Sendo todas essas informações customizadas pelo usuário, não existe uma validação para garantir, por exemplo, a grafia correta da localidade informada, é possível preencher esse campo descrevendo qualquer cadeia de caracteres, conforme exibido na Figura 2.1.

Após a criação do perfil, o usuário pode fazer a criação de repositórios a partir de duas formas: 1) repositório vinculado ao seu próprio perfil; ou 2) através de criação de uma organização e em seguida a criação de um repositório dentro dela. Uma organização é uma local para centralizar o trabalho de uma ou mais pessoas, nesse local é possível adicionar um

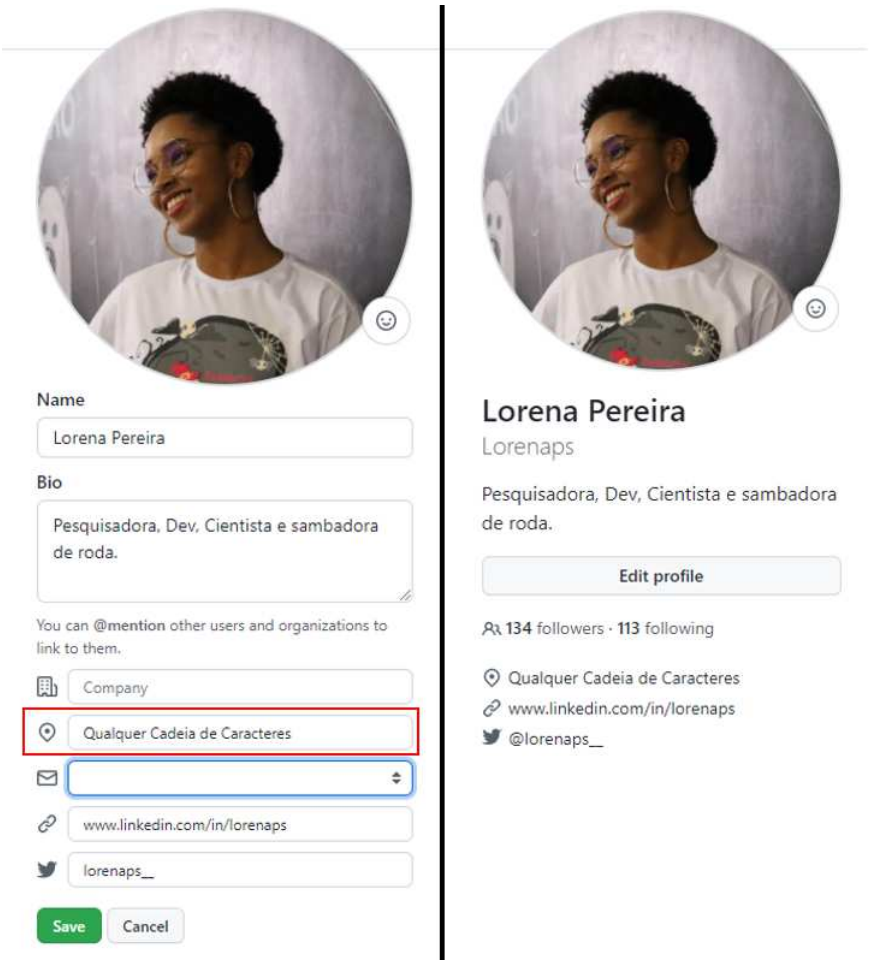


Figura 2.1: Edição de perfil do usuário no *GitHub*
Com destaque para a forma de preenchimento do campo referente a localização

e-mail da organização, imagem, localização, descrição e site, de forma a centralizar ali todas as informações pertinentes à organização, ficando assim independente do perfil do usuário que a criou. Seu objetivo é representar organizações reais (GitHub, Inc 2021), ou seja, um grupo de pessoas organizados com alguma finalidade. Após criar a organização também é possível permitir que outras pessoas tenham permissão de edição e colaboração sobre ela. Na Figura 2.2 é possível visualizar um perfil de usuário e na Figura 2.3 uma organização.

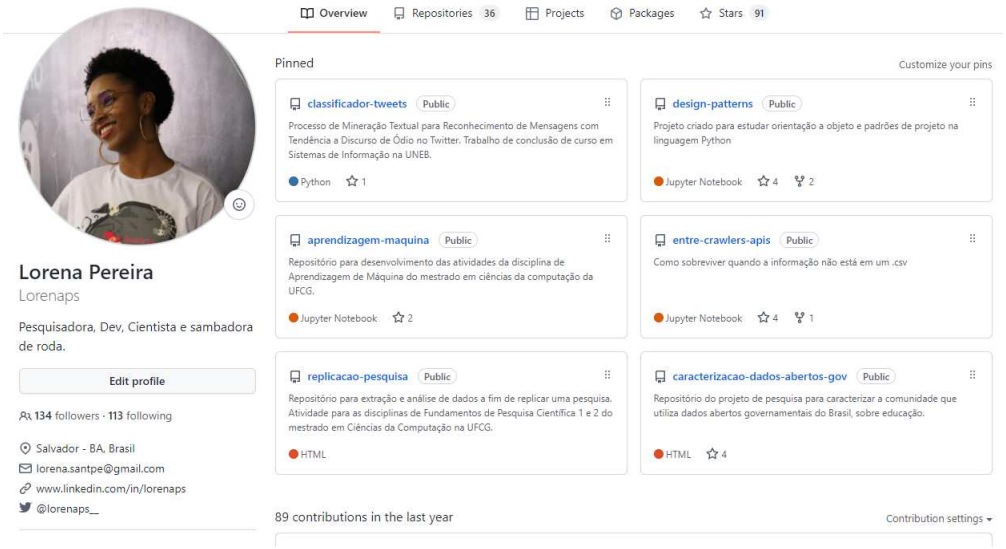


Figura 2.2: Visão geral do perfil de usuário no *GitHub*
Visão geral do perfil de usuário com as informações configuradas e repositórios destacados

Figura 2.3: Visão geral de uma organização do *GitHub*
Na organização da Prefeitura de São Paulo é possível visualizar alguns de seus repositórios e os membros que permitiram que outras pessoas vejam que ele colabora com àquela organização

A Figura 2.4 apresenta a disposição visual de um repositório. Eles possuem o local para a visualização dos arquivos, na lateral mostram a descrição, licenças e lista de contribuidores e na parte superior campos que permitem a interação entre usuários e repositórios. Os valores ao lado dos botões *Watch*, *Fork* e *Star* simbolizam, respectivamente, quantidade de pessoas que assinaram para receber notificação alterações naquele repositório, quantidade de pessoas que fizeram uma cópia do repositório e quantidade de pessoas que favoritaram o repositório. Já os *Commits* representam a quantidade de vezes que aquele repositório recebeu uma modificação. Ao fazer uma cópia do repositório a pessoa usuária denota a intenção de salvar o estado dos arquivos ali presentes e/ou de a partir daquela versão fazer uma alteração no conteúdo e possivelmente submeter essa alteração ao repositório original, cada alteração dessa contabiliza um *commit*. A relação de colaboração em repositórios no *GitHub* é estabelecida através da submissão de alterações e o respectivo aceite por parte dos demais colaboradores do repositório original.

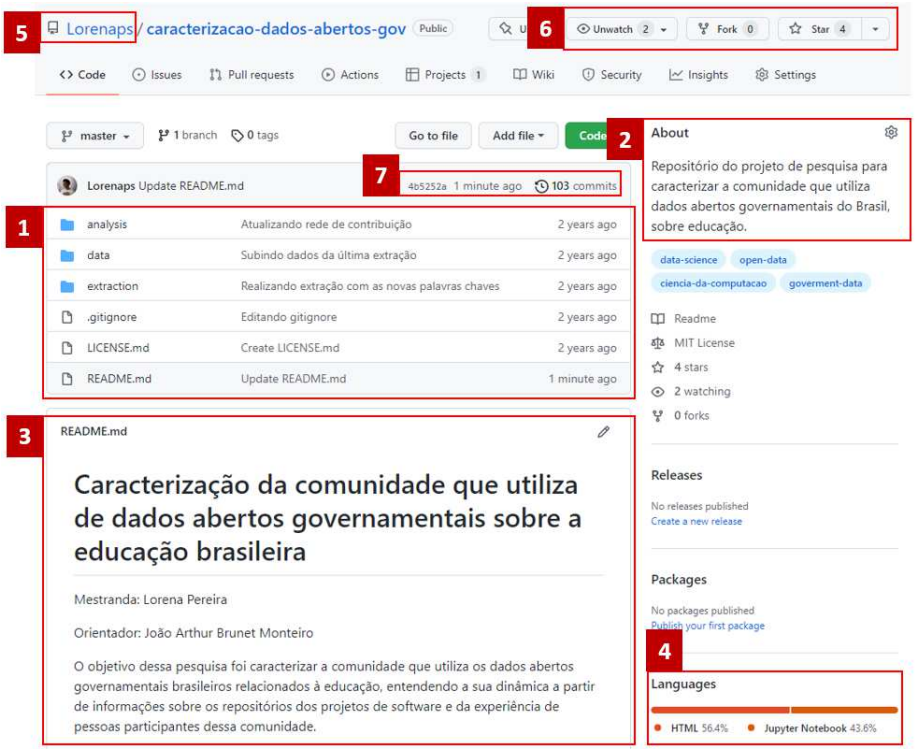


Figura 2.4: Visão geral de uma repositório do *GitHub*
1) Local de visualização da estrutura de arquivos; 2) Descrição sobre o repositório; 3) Renderização do arquivo *README.md*, esse arquivo é padrão dos repositórios e tem o objetivo de descrever as informações principais, como forma de execução do código ou instruções para colaborar com o projeto; 4) Identificação automática que o *GitHub* faz sobre a codificação dos arquivos (GitHub, Inc 2021); 5) Link do perfil dono daquele repositório, 6) botões de interação usuário/repositório e 7) quantidade de *commits* (alterações) nos arquivos daquele repositório

Essas informações podem ser retornadas via API conforme exibido na imagem Figura 2.5.


```
{
  "id": 206375532,
  "node_id": "MDEwOlJlcG9zaXRvcnkyMDYzNzU1MzI=",
  "name": "caracterizacao-dados-abertos-gov",
  "full_name": "Lorenaps/caracterizacao-dados-abertos-gov",
  "private": false,
  "owner": {
    "login": "Lorenaps",
    "id": 9660774,
    "node_id": "MDQ6VXNlcjk2NjA3NzQ=",
    "avatar_url": "https://avatars.githubusercontent.com/u/9660774?v=4",
    "gravatar_id": "",
    "url": "https://api.github.com/users/Lorenaps",
    "html_url": "https://github.com/Lorenaps",
    "followers_url": "https://api.github.com/users/Lorenaps/followers",
    "following_url": "https://api.github.com/users/Lorenaps/following{/other_user}",
    "gists_url": "https://api.github.com/users/Lorenaps/gists{/gist_id}",
    "starred_url": "https://api.github.com/users/Lorenaps/starred{/owner}{/repo}",
    "subscriptions_url": "https://api.github.com/users/Lorenaps/subscriptions",
    "organizations_url": "https://api.github.com/users/Lorenaps/orgs",
    "repos_url": "https://api.github.com/users/Lorenaps/repos",
    "events_url": "https://api.github.com/users/Lorenaps/events{/privacy}",
    "received_events_url": "https://api.github.com/users/Lorenaps/received_events",
    "type": "User",
    "site_admin": false
  },
  "html_url": "https://github.com/Lorenaps/caracterizacao-dados-abertos-gov",
  "description": "Repositório do projeto de pesquisa para caracterizar a comunidade qu
  "fork": false,
  "url": "https://api.github.com/repos/Lorenaps/caracterizacao-dados-abertos-gov",
```


Figura 2.5: Visão geral da reposta da API do *GitHub*
Resposta da consulta na API para o repositório Lorenaps/caracterizacao-dados-abertos-gov

2.3 Análise Textual Discursiva

Ao buscar aprofundar o entendimento sobre as experiências da comunidade que utiliza DAGs educacionais, se faz necessário debruçar-se sobre dados que possam conter ideias e motivações de forma mais extensa. E para isso, também foi preciso entender melhor sobre métodos de análise de dados relacionado a pesquisa qualitativa e a pesquisa exploratória.

A pesquisa qualitativa busca aprofundar a compreensão sobre um determinado contexto ou grupo social (Gerhardt e Silveira 2009). Tarefa para a qual o método descrito e empregado por Moraes e Galiazzi 2007 é capaz de ajudar. A Análise Textual Discursiva se baseia em uma experiência recursiva de decomposição, criação de relacionamentos e validação entre o

que se pretende analisar e as ideias emergentes da análise. Os autores costumam descrevê-la como "Uma tempestade de luz":

> "Esse processo em seu todo pode ser comparado com uma tempestade de luz. O processo analítico consiste em criar as condições de formação dessa tempestade em que, emergindo do meio caótico e desordenado, formam-se flashes fugazes de raios de luz iluminando os fenômenos investigados, que possibilitam, por meio de um esforço de comunicação intenso, expressar novas compreensões atingidas ao longo da análise."(Moraes 2003)

O processo a que (Moraes 2003) se refere são as etapas de:

- Unitarização: A partir do *corpus*, que é o conjunto de documentos a ser analisado, realiza-se um processo de decomposição do conteúdo, o separando em unidades de análise. Essas unidades são o núcleo da ideia presente no recorte em questão e precisa ter um significado completo em si mesma;

- Categorização: Após a identificação das unidades de análise busca-se agrupá-las em categorias. Moraes 2003 afirma que a "explicitação das categorias acontece por intermédio do retorno cíclico às unidades de análise, no intuito da construção gradativa do significado de cada categoria.".

- Comunicação: Processo de escrita que visa apresentar as ideias emergentes da análise, também chamado de *metatexto*. O produto final da ATD é construído a partir da descrição e interpretação das categorias definidas e suas justificativas. "Alguns textos serão mais descritivos, mantendo-se mais próximos do corpus original. Já outros serão mais interpretativos, pretendendo um afastamento maior do material original num sentido de abstração e teorização mais aprofundado"(Moraes 2003).

A ATD se baseia nesse exercício de decomposição e reconstrução e exige um aprofundamento, "envolvimento e impregnação"da pessoa pesquisadora, também imbuída dos seus conhecimentos prévios sobre o assunto, para encontrar os agrupamentos existentes entre as unidades, e partir disso criar categorias. As categorias e suas justificativas são confrontadas com as unidades que nela se decidiu por agrupar durante todo o processo.

Capítulo 3

Trabalhos Correlatos

Para aprofundar o entendimento no contexto do problema, foi realizada uma pesquisa procurando por trabalhos capazes de esclarecer os seguintes tópicos: *Como a comunidade em torno de dados abertos governamentais tem sido caracterizada?* e *O que sabemos sobre projetos de natureza educacional?*

3.1 Como a comunidade em torno de dados abertos governamentais tem sido caracterizada?

A systematic review of open government data initiatives de Attard et al. 2015 teve por objetivo explorar artigos estivessem relacionadas com a publicação e consumo de dados abertos governamentais. Como resultado trouxe uma visão geral sobre conceitos, terminologias, desafios e possíveis soluções, do ponto de vista cultural e técnico. O método utilizado para revisão sistemática da literatura foi composto dos seguintes passos:

1. Definição de termos de pesquisa;
2. Seleção de fontes de pesquisa;
3. Aplicação dos termos de pesquisa nas fontes selecionadas;
4. Seleção dos projetos primários através da aplicação de critérios de inclusão e exclusão.

A fim de alcançar publicações relevantes, os canais de pesquisa utilizados foram revistas científicas eletrônicas largamente utilizadas: ACM Digital Library, IEEE Xplore Digital

Library, Science Direct, Springer Link e ISI Web of Knowledge. Isso permitiu ao estudo analisar o estado da arte desse tipo de iniciativa e apresentar uma tabela com os principais desafios, natureza dos desafios e possíveis soluções, disponível na Tabela 3.1.

A principal contribuição de Attard et al. 2015 para o presente trabalho é a apresentação em detalhes de uma metodologia capaz de construir um conhecimentos de forma estruturada sobre o tema que se pretendia caracterizar. Contudo, o estudo atual se diferencia pois buscou estender o conhecimento sobre essa comunidade focando nos dados da área de educação e em projetos de software para além da publicação em revistas científicas.

Natureza do desafio	**Desafio**	**Possível solução**
Técnico	Formato	Usando um formato não proprietário e processável por máquina
	Ambiguidade	Usando um formato descritivo
		Adicionando documentação / metadados
	Descobrimento	Usando metadados de boa qualidade
		Mais ferramentas avançadas de pesquisa em portais
	Representação	Definindo e utilizando representação padronizada
		Usando gráficos nomeados para controle de versão
	Capacitação	Aplicando padrões
		Treinamento em grande escala
Político / Legal	Direitos de uso e licenciamento	Definindo políticas de dados padrão
	Regulamentos conflitantes	Definindo políticas de iniciativa de dados do governo aberto e estruturas legais
	Privacidade / Proteção de dados	Definindo regulamentações de privacidade
		Implementando mecanismos de controle de acesso (isso limita a abertura dos dados)
	Responsabilidade	Interação social
		Sensibilização
		Definindo estruturas legais
Econômico / Financeiro	Provisão de orçamento	Fornecer orçamento especificamente para iniciativas de dados abertos
Organizacional	Institucionalização	Reorganizar a atual estrutura organizacional
		Definir políticas de iniciativas governamentais abertas
	Escopo sobreposto	Usando metadados de proveniência
	Suporte técnico	Prestar apoio a entidades públicas com a execução de uma iniciativa de dados abertos
Cultural	Motivação	Aumentar a conscientização sobre a reutilização de dados abertos e seus benefícios
	Sensibilização	Destacando o valor e o potencial dos dados abertos
	Participação Pública	Sensibilização
		Fornecendo incentivos
	Concorrência	Fornecer dados específicos a uma taxa nominal (isso limita a abertura dos dados)

Tabela 3.1: Visão geral dos desafios em iniciativas de dados abertos governamentais. Fonte: Traduzido de Attard et al. 2015

Em *Open government data portals analysis: the Brazilian case*, Oliveira et al. 2016 fizeram uma análise mais específica considerando portais de dados abertos governamentais no Brasil, avaliando a qualidade no que diz respeito aos 8 princípios de dados abertos.

A fim de responder a questão de pesquisa *Como dados abertos governamentais tem sido publicados pela administração pública brasileira?* os pesquisadores realizaram uma investigação quantitativa dos portais existentes no Brasil. Para isso, seguiram as seguintes etapas:

- Definição de critérios de avaliação
- Seleção da população de amostra
- Coleta de dados
- Análise dos conjuntos de dados

As fontes de dados selecionadas para análise foram o portal do governo federal, os portais estaduais de: Alagoas, Distrito Federal, Minas Gerais, Pernambuco, Rio Grande do Sul, e São Paulo e outros 6 portais municipais de: Fortaleza, Curitiba, Porto Alegre, Recife, Rio de Janeiro e São Paulo. Sua análise cobriu os seguintes aspectos referente aos dados apresentados: Acesso, formato, atualização, metadados disponíveis, licenças e convenções na nomenclatura dos campos.

Por fim, observaram que embora a administração pública brasileira esteja divulgando seus dados a divulgação ainda não está sendo feita completamente de acordo com os 8 princípios. Problemas como formato proprietário ou não estruturado, não utilização de um vocabulário comum para dados do mesmo contexto e a ausência de metadados capazes de gerar o entendimento completo dos dados ainda são recorrentes nesses portais.

A contribuição de Oliveira et al. 2016 nos forneceu exemplos práticos de fontes de dados que não estão em conformidade com os princípios esperados, assim como exemplos de problemas que podem ser encontrados também pela comunidade de estudo do presente trabalho.

Em *Dados Abertos do Governo Brasileiro: Entendendo as Perspectivas de Fornecedores de Dados e Desenvolvedores de Aplicações ao Cidadão* (Araújo 2017) buscou compreender e comparar as perspectivas dos fornecedores de DAGs e desenvolvedores brasileiros que

usufruem desses dados, entendendo quais as suas motivações e as barreiras enfrentadas no fornecimento e utilização desse recurso.

Para isso, realizou um estudo qualitativo exploratório com entrevistas semi-estruturadas, primeiramente com 12 desenvolvedores que utilizaram DAGs e depois com 12 pessoas que trabalhavam no fornecimento de DAGs. Alguns dos principais entendimentos extraídos a partir da análise de (Araújo 2017) sobre sua questões de pesquisa foram:

1. Quais são as motivações e objetivos de desenvolvedores para criar ou colaborar com projetos que fazem uso de DAGs brasileiros?
    1.1. Sentem a necessidade de proporcionar à sociedade o acesso a uma melhor visualização dos dados disponibilizados pelos órgãos governamentais;
    1.2. Promover a tradução dos dados tal como eles são disponibilizados e melhorar o entendimento por parte do cidadão comum;
2. O que motiva fornecedores de DAGs brasileiros a trabalharem com a publicação de DAGs?
    2.1. Saber que as pessoas estão realmente fazendo uso das bases de dados;
    2.2. Promover melhorias das políticas públicas;
    2.3. Fornecer dados fáceis de serem entendidos, disponibilizando-os de forma acessível, utilizável e com qualidade.
3. Quais são as barreiras e desafios que desenvolvedores enfrentam para utilizar DAGs brasileiros?
    3.1. Falta de padronização dos dados;
    3.2. Dados inconsistentes, aleatórios ou com informações nulas;
    3.3. Problemas de colaboração e feedback entre fornecedores e desenvolvedores;
4. Quais são as barreiras e desafios que cidadãos enfrentam para utilizar DAGs brasileiros?
    4.1. Falta de incentivo ao cidadão para que este busque e alcance o entendendimento dos reais significados dos dados;

4.2. Necessidade do cidadão em possuir conhecimento técnico e instrutivo sobre os DAGS disponíveis;

5. Quais são as barreiras e desafios que fornecedores de DAGs brasileiros enfrentam ao tentar disponibilizar um dados aberto, hoje em dia no país?

   5.1. Não possuir a cultura em promover a abertura dos seus dados;

   5.2. Dificuldade de convencimento dos gestores;

   5.3. Pouca interação entre os níveis federal, estadual e municipal;

6. Que melhorias trariam benefícios para o trabalho dos desenvolvedores que utilizam DAGs?

   6.1. Aprimoramento das buscas dentro dos portais;

   6.2. Disponibilização de tutoriais, manuais e outros tipos de instruções sobre portais de DAGs;

   6.3. Padronização dos dados;

7. Que melhorias trariam benefícios para o trabalho dos fornecedores de DAGs?

   7.1. Criação de eventos;

   7.2. Criação de redes de colaboração entre diversos stakeholders;

   7.3. Promover a capacidade de TI e gestão da informação dentro dos órgãos;

Esse trabalho se faz fundamental para essa pesquisa pelo olhar aprofundado que permite sobre motivações, barreiras e sugestões de melhorias dos atores em questão. A partir dessa análise foi possível fazer alguns paralelos entre os resultados dos DAGs em geral e os encontrados quando focamos nos DAGs da área de Educação.

3.2 O que sabemos sobre projetos de natureza educacional?

Penteado e Isotani 2017 buscaram em seu trabalho responder a questão “Quais dados abertos educacionais estão disponíveis para a sociedade?”, considerando o contexto brasileiro.

Para isso, a partir de uma pesquisa no Portal Brasileiro de Dados Abertos, filtrando pelos órgãos MEC e INEP (considerados os principais relacionados com a educação brasileira), fizeram uma caracterização dos assuntos cobertos e periodicidade da publicação dos dados disponíveis.

Os autores organizaram os dados em conjunto de dados, informações relacionadas a um assunto específico, como ENEM, e em arquivo de dados, que são os arquivos em si. A partir disso, aplicaram os seguintes critérios de seleção: (i) conjuntos de dados que estivessem disponíveis; (ii) que tivessem publicação em mais de um ano e (iii) que não fossem duplicados. Por fim, ficaram com 35 conjuntos de dados e 325 arquivos.

Os conjuntos de dados vão de 1995 até 2016, com um crescimento acentuado ao longo dos anos. Os dados mais antigos foram do censo da educação básica, censo do ensino superior e do SAEB - Sistema de Avaliação da Educação Básica, com dados referentes a 1995 e seguindo a publicação anualmente.

E além dos citados anteriormente, Exame Nacional do Ensino Médio - ENEM, Programa Nacional de Alimentação Escolar - PNAE, Fundo de Financiamento Estudantil - FIES, dados do ensino superior, estrutura do ensino superior, estrutura do ensino básico, Programa Universidade para Todos - PROUNI, Brasil Alfabetizado, Programa Dinheiro Direto na Escola - PDDE, ensino técnico, Exame Nacional de Desempenho dos Estudantes - ENADE e matrículas do ensino superior, também são conjuntos com mais de 10 anos de publicação consecutiva (Penteado e Isotani 2017). Na Figura 3.1 podemos ver os conjuntos de dados avaliados pelos autores publicados a cada ano de referência.

Vale destacar que quando se fala em de dados do censo de 1995, não significa que eles estão disponíveis no formato em questão desde de esse ano, ou que esses dados não sejam atualizados ao longo do tempo. A data de publicação da documentação dos microdados do censo escolar de 1995 (que é disponibilizada junto ao conjunto de dados no portal no INEP [1]), por exemplo, data de novembro de 2006 (INEP e MEC 2006) e quando olhamos a última data de atualização do arquivo dos dados, data de junho de 2019.

Por fim, os autores também observaram a diminuição drástica da disponibilização dos dados a partir do ano de 2014. Essas foram importantes observações para o presente trabalho

[1]Microdados do censo escolar: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar

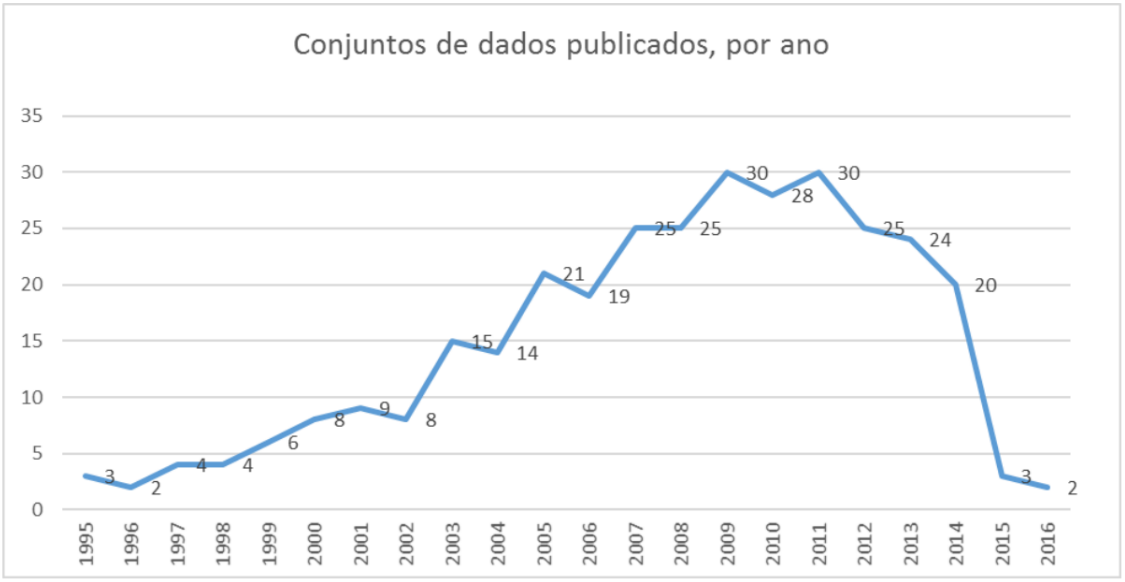


Figura 3.1: Total de conjuntos de dados publicados a cada ano, desde 1995
Fonte: Penteado e Isotani 2017

visto o panorama exposto das bases de dados existentes e o entendimento da dinâmica da publicação e periodicidade. Portanto, entendemos que existem diversos assuntos a serem explorados dentro da área de educação, com recorte temporal de 1995 a 2016, e que 15 deles, que descrevem desde de educação básica até a estrutura do ensino superior, possuem mais de 10 anos de dados.

Já Santos, Ferreira e Miranda 2017 fizeram uma revisão da literatura sobre trabalhos que utilizam dados abertos no contexto educacional, com destaque para os objetivos pedagógicos e origens das publicações. Eles partiram da pesquisa em periódicos, conferências, teses, dissertações e trabalhos de conclusão de curso. Buscavam por trabalhos que tivessem relação com educação ou sistemas computacionais, e usaram como palavras chaves:

- ”Dados abertos educacionais”;
- ”Dados abertos + educação”;
- ”INEP”;
- ”IDEB”;
- ”ENEM”.

Os autores consideraram os trabalhos entre 2008 e 2017 e aplicaram os seguintes critérios de exclusão: (i) Fora do período da pesquisa;(ii) Não utilizavam dados abertos educacionais; e (iii) Não utilizavam dados abertos para aplicações educacionais.

Com os 32 artigos restantes, eles ressaltaram o indício de o quanto a área estudada ainda estava em fase de crescimento no Brasil, devido à distribuição dos artigos ao longo dos anos, onde mesmo pesquisando por artigos desde de 2008 os primeiros artigos só apareceram para o ano de 2011, conforme mostrado na Figura 3.2

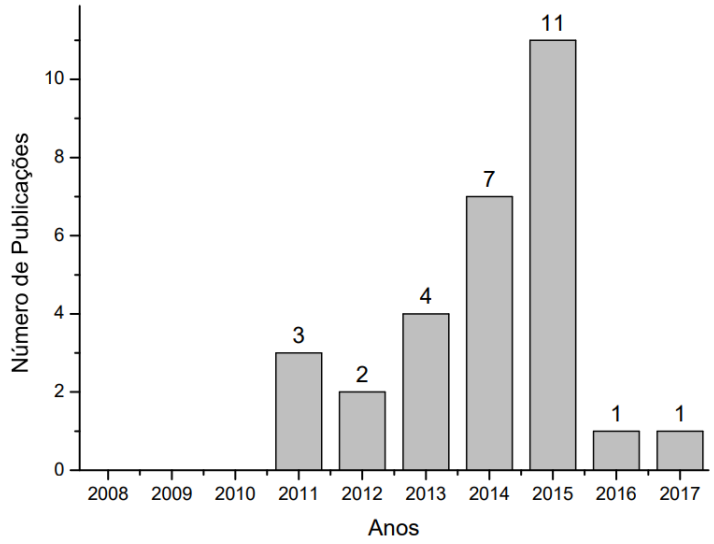


Figura 3.2: Número de publicações ao longo dos anos
Fonte: Santos, Ferreira e Miranda 2017. Essa visualização produzida pelos autores acabou não considerando 3 publicações, por isso que o total somado é 29 publicações ao invés de 32. Essa confirmação foi obtida através da troca de e-mails com os autores em janeiro de 2022. Na resposta os autores também enviaram a planilha completa dos artigos, e os 3 que faltavam na visualização acima foram publicados 1 em 2015 e 2 em 2016.

Com relação a análise dos artigos, os principais objetivos e suas subdivisões foram:

1. Mineração de dados aplicados a dados abertos educacionais: que poderia se subdividir em sistemas para apoiar a pessoa gestora na tomada de decisão, predição de desempenho de escolar e análise de algoritmos de aprendizagem de máquina focados para dados abertos educacionais;

2. Análise qualitativa: que diz respeito a qualidade dos dados, possibilidades de aplicações e estudos de contextos escolares;

3. Teórico: que descrevem métodos e desafios sobre a utilização dos dados;

4. Sistemas para apoio a aprendizagem colaborativa;

5. Visualização dos dados: aplicações que possibilitam a visualização dos dados de uma determinada região de forma mais acessível

Por fim, os autores fizeram uma discussão sobre os trabalhos que analisaram, sobre como ainda é um quantitativo pequeno de iniciativas e sobre como ainda existe um potencial para evolução da utilização desse recurso. Esse é um trabalho importante para a nossa pesquisa pela visão geral que traz sobre trabalhos desenvolvidos pela comunidade em publicações acadêmicas, sobre as palavras chaves de busca que utilizou para a pesquisa de artigos e os objetivos mapeados. Os objetivos apresentados podem permitir também uma análise de comparação entre os resultados das duas pesquisas.

Capítulo 4

Caracterização da Comunidade que Utiliza Dados Abertos Governamentais Sobre a Educação Brasileira

No contextos de DAGs educacionais brasileiros, existem esforços para o entendimento da adequação desse recurso com os princípios que eles devem seguir para facilitar a utilização (Penteado, Bittencourt e Isotani 2017), órgãos centrais que fazem a publicação dos mesmos abrangendo uma diversidade considerável de assuntos (Penteado e Isotani 2017) e o compromisso do governo brasileiro que desenvolve um plano nacional para a consolidação de práticas de governo aberto no país (Controladoria-Geral da União 2021). Mas não está disponível um mapeamento sobre forma de organização ao redor desses dados, quais projetos de software, para além daqueles publicados academicamente, são desenvolvidos, onde estão localizados, quais são as fontes dos dados e tecnologias utilizadas ou se existem problemas particulares do tema educacional.

Para avançar nesse conhecimento, foi realizada uma pesquisa exploratória, buscando em um primeiro momento entender características quantitativas e em seguida características qualitativas, com o objetivo fim de construir uma caracterização que pudesse descrever informações acerca da comunidade que utiliza dados abertos governamentais sobre a educação brasileira. Os objetivos específicos desse trabalho foram:

1. Entendimento dos projetos dessa comunidade: Descrição das características dos pro-

jetos, como autores, rede de contribuição, fonte dos dados, regionalidade, tecnologias utilizadas, popularidade, temporalidade, objetivo dos projetos e trabalhos realizados;

2. Entendimento das experiências dessa comunidade: Descrição das motivações na realização dos projetos, impactos percebidos, problemas encontrados, formas e canais de compartilhamento de soluções.

Esse capítulo visa descrever os materiais e métodos utilizados para alcançar os objetivos.

4.1 Entendimento dos projetos

Para mapeamento e coleta dos dados sobre os projetos foi replicado o método de pesquisa descrito em Attard et al. 2015. Em seu trabalho os autores tiveram o objetivo de explorar o cenário das iniciativas de dados abertos governamentais através de uma revisão da literatura que contou com 75 publicações, e como resultado apresentaram uma visão geral sobre conceitos, terminologias, desafios e possíveis soluções, do ponto de vista cultural e técnico. As etapas do método de pesquisa foram: Definição de termos de pesquisa; Seleção de fontes de pesquisa; Aplicação dos termos de pesquisa nas fontes selecionadas; Seleção dos projetos primários através da aplicação de critérios de inclusão e exclusão.

No entanto, diferentemente do trabalho de Attard et al. 2015, a presente pesquisa utilizou uma fonte que armazenasse repositórios públicos dos artefatos de software dos projetos, no caso o *Github*. Justifica-se a escolha da plataforma por sua popularidade e diversidade de projetos, conforme apresentado em Seção 2.2. Além disso, a plataforma contém uma API - Application Programming Interface (Interface de Programação de Aplicativos), que é um conjunto de rotinas prontas para serem utilizadas via requisições HTTP permitindo assim a interação com sua base de dados. Esse foi um recurso importante pois possibilitou a pesquisa e extração de dados sobre repositórios e usuários da plataforma de forma automatizada.

Para a definição dos termos de pesquisa foi feita uma análise das bases de dados existentes e dos temas citados em portais de referencia para a educação brasileira, a fim de construir uma lista com as principais palavras chaves, a exemplo de:

- Site do INEP: ENEM, Censo escolar, SAEB, MDE, Indicadores financeiros educacionais;

- Site do Ministério da Educação: SISU, ENEM, FIES, PROUNI, MEC, PME, PRONATEC, PNP, FIES;

- Site do FUNDEB (Fundo de Manutenção e Desenvolvimento da Educação Básica): FUNDEB, FNDE, SIOPE.

O Portal Brasileiro de Dados Abertos também foi consultado, mas as palavras chaves encontradas já haviam sido cobertas pela análise dos outros sites. A partir da listagem feita após analisar cada portal, as palavras chaves eram testadas na API de busca do GitHub e discutidas entre orientanda e orientador até que foi possível descrever um escopo abrangente em variações dos termos. A lista final dos termos de pesquisa está descrita na Tabela 4.1.

Para extrair os dados acerca dos repositórios públicos do *GitHub* desenvolveu-se um script na linguagem de programação *Python*, a fim de comunicar-se com a API da plataforma. A API é pública mas para aumentar a quantidade de requisições por hora é preciso realizar autenticação, dessa forma passa-se a poder realizar 5000 requisições por hora. Uma outra limitação, na época que foi realizada a extração, era que para cada busca só era possível acessar os primeiros 1000 resultados. Após a pesquisa, os resultados foram salvos em um arquivo de formato aberto (.csv). O código utilizado para extração de dados está disponível no repositório público desse projeto de pesquisa no *GitHub*[1] assim como os dados extraídos.

[1] https://github.com/Lorenaps/caracterizacao-dados-abertos-gov

Palavras de busca para extração junto a API do *GitHub*	
dados educacao	analise inep
dados educacao basica	microdados inep
dados educacionais	dados sisu
analise educacao	analise sisu
analise educacao basica	dados fies
analise educacional	analise fies
censo educacao superior	dados prouni
dados educacao superior	analise prouni
analise educacao superior	dados mec
censo profissionais magistério	analise mec
dados profissionais magistério	dados pme
analise profissionais magistério	analise pme
censo escolar	dados pronatec
dados escola inep	analise pronatec
dados enade	dados pnp
analise enade	analise pnp
dados encceja	dados fundeb
analise encceja	analise fundeb
dados enem	dados fnde
analise enem	analise fnde
enem por escola	dados siope
dados prova brasil	analise siope
analise prova brasil	dados saeb
dados ideb	analise saeb
indicadores educacionais	dados mde
dados ies	analise mde
analise ies	indicadores financeiros educacionais
dados inep	

Tabela 4.1: Palavras de busca.

Após a extração dos dados do *GitHub*, a fim de garantir que análise seria feita apenas com repositórios que estivessem relacionados com DAGs educacionais brasileiros, cada repositório passou por uma avaliação manual considerando os critérios de inclusão:

1. Utilizar dados abertos fornecidos pelo governo através de fontes oficiais;
2. Utilizar dados abertos obtidos do governo através de fontes não oficiais (Canais derivados como Brasil IO);
3. Estar relacionado a iniciativas que utilizam dados abertos governamentais do Brasil na área de educação.

E os critérios de exclusão:

1. Não realização de processamento ou visualização dos dados, sendo apenas um catálogo com links de iniciativas. Isso é avaliado a partir da leitura do arquivo README.md[2] de cada projeto.
2. Repositório vazio. É preciso que haja algum arquivo de código além do arquivo README.md.
3. Não existência de um arquivo README.md ou a descrição existente não apresentar informações suficientes para caraterização do repositório, por exemplo no README.md haver apenas o título do projeto.

Os dados foram extraídos do *GitHub* no dia 19/05/2020 e a avaliação manual foi realizada entre o dia 22/05/2020 e 23/10/2020. Os repositórios aprovados, aqueles que passaram pelos critérios de inclusão e não satisfazerem nenhum dos critérios de exclusão, foram analisados manualmente a fim de extrair as seguintes informações:

- Organização responsável (autores do projeto);
- Objetivo do projeto (informativo, educacional, abertura ou ETL - Extract Transform Load, Extração, transformação e carregamento de dados);

[2]README.md é um arquivo escrito na linguagem de marcação *Markdown* com o propósito de armazenar informações descritivas do projeto em questão, suas dependências e até mesmo como executar e/ou contribuir.

- Trabalho realizado (Análise de dados, construção de uma aplicação, disposição de dados);
- Fonte dos dados;
- Local vinculado ao perfil dono do repositório (Estado);

As informações que se buscava extrair eram descritivas do projeto e não vinculadas a padrões de software, portanto a análise de arquivos de código não se fez necessária. Dito isso, a busca de informações teve como foco: campo de descrição do projeto (About), arquivo README.md e estrutura de arquivos, na Figura 4.1 é possível visualizar a aparência comum dos repositórios.

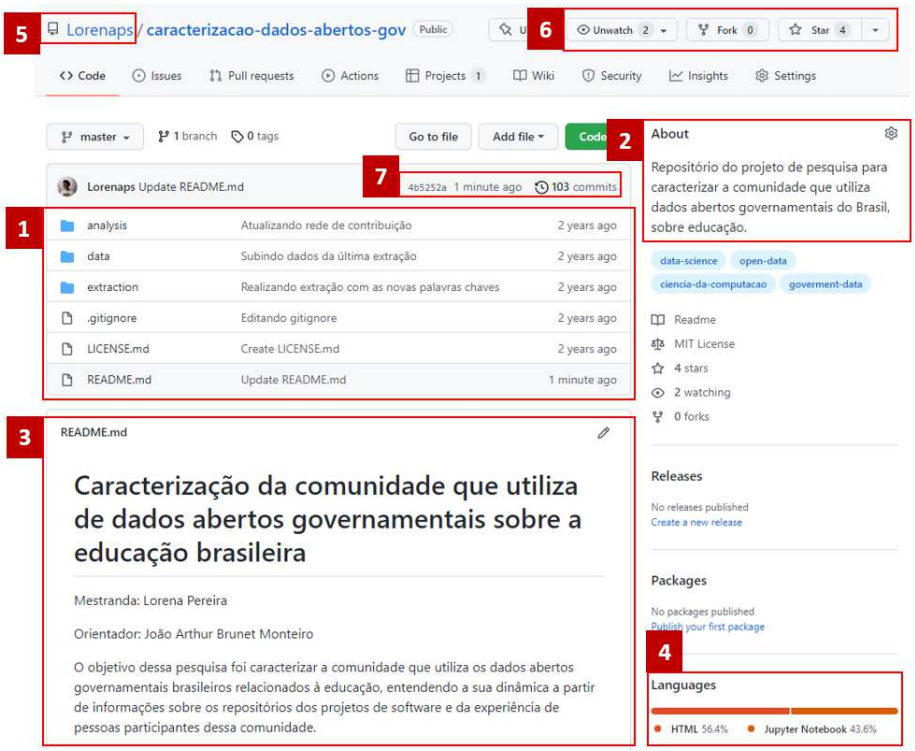


Figura 4.1: Visão geral de uma repositório do *GitHub*
1) Local de visualização da estrutura de arquivos; 2) Descrição sobre o repositório; 3) Renderização do arquivo *README.md*, esse arquivo é padrão dos repositórios e tem o objetivo de descrever as informações principais, como forma de execução do código ou instruções para colaborar com o projeto; 4) Identificação automática que o *GitHub* faz sobre a codificação dos arquivos (GitHub, Inc 2021); 5) Link do perfil dono daquele repositório, 6) botões de interação usuário/repositório e 7) quantidade de *commits* (alterações) nos arquivos daquele repositório

A estrutura de arquivos era analisada buscando identificar algum arquivo de código, de

forma a entender se o repositório não estava vazio, ou possuía mais do que a descrição do projeto. Por fim, o passo a passo realizado para cada repositório foi:

1. Abrir a URL do repositório;
2. Analisar se o repositório cumpria os critérios de inclusão e não satisfazia nenhum dos critérios de exclusão. Em caso positivo, os passos seguintes eram realizados:
    2.1. Buscar informações sobre autores e localização no perfil dono do repositório;
    2.2. Analisar *Sobre* e o arquivo README.md buscando identificar as informações descritas anteriormente;
    2.3. Caso houvesse um site indicado nas descrições (README.md ou campo *Sobre*) o recurso também era verificado. A observação feita nos sites buscava informações sobre quem desenvolve e localização, geralmente nas seções de rodapé, sobre e/ou contato.

4.2 Entendimento das experiências

A fim de entender informações qualitativas e vinculadas às experiências das pessoas/organizações pertencentes a essa comunidade, foi realizada uma pesquisa de opinião. O questionário da pesquisa contou com 1 seção descritiva do mesmo, onde também adicionamos um vídeo para complementar e humanizar o convite de participação, principalmente considerando que pesquisas impossibilitadas de ir a campo tiveram que se adaptar e aderir a formulários online e as dificuldades enfrentadas pela população durante a pandemia da Covid-19. O questionário completo pode ser encontrado no Apêndice A.

O questionário foi construído utilizando a plataforma do *Google Forms*, contou com 16 perguntas, sendo 6 abertas e 10 fechadas. Ele também foi dividido em 3 seções:

1. Primeira: para entender em detalhes a experiência na utilização de dados abertos governamentais;
2. Segunda: para entender sobre sua experiência e opinião sobre dados abertos no geral;
3. Terceira: para traçar um perfil das pessoas respondentes.

O questionário esteve aberto entre os dias 07/04/2021 e 21/04/2021 e teve 39 respostas, onde 38 foram válidas. O critério usado foi uso de fontes de dados relacionadas à educação.

Para a análise qualitativa das respostas das questões abertas foi aplicado o método ATD - Análise Textual Discursiva (Moraes e Galiazzi 2007). Esse método foi escolhido visto a sua aplicação a um contexto de uma pesquisa exploratória onde não existem categorias prévias ou já esperadas, e a sua capacidade de apoiar a produção e expressão de sentidos. As 6 questões abertas foram:

1. *Com qual base de dados você mais trabalhou?*
2. *Qual foi a motivação, sua ou da sua organização, na construção desse projeto?*
3. *Como você percebeu o impacto/resultado do seu projeto no respectivo público alvo?*
4. *Quais problemas tecnológicos e/ou na disponibilidade dos dados foram encontrados no desenvolvimento do projeto?*
5. *No caso de ter conseguido resolver o problema relatado, essa solução foi compartilhada com outras pessoas/organizações? De que forma?*
6. *Você conhece e/ou indica algum fórum ou canal para trocar conhecimentos sobre dados abertos governamentais de forma geral?*

Apenas as 5 primeiras foram consideradas para a aplicação do método, visto o seu objetivo de entender as experiências da comunidade. A última questão foi um campo aberto apenas para que fosse possível coletar canais diversos de compartilhamento.

Conforme descrito no Seção 2.3, a ATD se baseia em 3 etapas: unitarização, categorização e comunicação. Antes de detalhar como cada etapa foi realizada no trabalho, é importante retomar os conceitos definidos por Moraes e Galiazzi 2007:

- *Corpus*: Conjunto de documentos/materiais que se pretende analisar;
- Unidades de análise: Também chamadas de unidades de significado ou de sentido, são os fragmentos do corpus destacados pela pessoa pesquisadora e que denotam os principais significados ali contidos;
- Categorias: Resultado do agrupamento das unidades de análise afins;

- Metatexto: A partir do processo de fragmentação e desmontagem do conteúdo original, seguindo para a identificação de novas relações, o metatexto é a expressão do entendimento da pessoa pesquisadora sobre os significados e sentidos extraídos do processo.

Dito isso, na etapa de unitarização, que é onde ocorre a desconstrução dos textos através do destaque de seus elementos constituintes, foi realizada uma leitura detalhada de todas as respostas de cada uma das questões (*corpus*). Para cada questão o seu contexto era considerado, ou seja, ao analisar a questão *Com qual base de dados você mais trabalhou?*, a leitura era realizada considerando observar e destacar os locais que as pessoas indicaram como fonte dos dados em detrimento de outras informações que elas poderia ter expressado ao longo da resposta, exemplo: ***"A partir de uma palestra sobre evasão escolar procurei por sites governamentais que apresentavam dados sobre o tema, conseguimos a informação analisando as bases de dados do Censo Escolar e ENEM do INEP"*** a partir dessa resposta, as unidades de análise destacadas seriam ***"Censo Escolar"*** e ***ENEM***.

Da mesma forma, ao analisar as respostas da questão *Qual foi a motivação, sua ou da sua organização, na construção desse projeto?*, foi percebido o quanto os verbos eram importantes e se destacavam no entendimento da motivação das pessoas, exemplo: ***"Praticar análise e visualização de dados em uma disciplina"*** e ***"Análise de notas de redações do Enem para fundamentar aulas para o Ensino Médio; Informações do Enade para planejamento de programas de capacitação dos discentes de graduação para realização das provas"***. Na primeira resposta vemos que o verbo **"praticar"** mais a palavra **"disciplina"** são unidades de análise fundamentais para o entendimento da ideia expressada, que por sua vez liga a motivação ao cumprimento de trabalhos e atividades de disciplinas e/ou cursos. Assim como o trecho **"fundamentar aulas para o Ensino Médio"** na segunda resposta, com destaque para o verbo fundamentar, denotam de forma direta que o trabalho com essas bases de dados visava poder utilizar as informações extraídas para auxiliar gestores e professores.

Questão	Exemplos das unidades de análise
Com qual base de dados você mais trabalhou?	ideb, censo escolar, bolsas do CNPq, valor investido, cgee, enade, censo, tce, moodle, cpc, idd, merenda escolar
Qual foi a motivação, sua ou da sua organização, na construção desse projeto?	projeto de uma disciplina, entender como, criar visualizações, fiscalização de, praticar, para fundamentar aulas, monitorar, fortalecer controle social
Como você percebeu o impacto/resultado do seu projeto no respectivo público alvo?	publico interessado, forneceram sugestões, usaram como base, <tiveram o retorno>era disso que precisava, evidenciou <algo>, realizei uma análise, <complementou>às atividades do público, foi possível observar, não houve, aprendizado próprio, não publiquei
Quais problemas tecnológicos e/ou na disponibilidade dos dados foram encontrados no desenvolvimento do projeto?	parcialmente desorganizados, não estavam disponibilizadas nas bases, site de dificil acesso, metadados nem sempre claros, timeout para o bot acessar, falta de um padrão, dificuldade de entendero que os dados queriam dizer, Alguns sites não estavam disponíveis, foi preciso pedir a LAI
No caso de ter conseguido resolver o problema relatado, essa solução foi compartilhada com outras pessoas/organizações? De que forma?	o código foi disponibilizado de forma aberta no github, publicamos uma sessão de metodologia nos nossos estudos, foi citada em uma discussão da equipe, compartilhada no github e também no stackoverflow, por meio de reuniões de grupo trocando informações de como foi o processo

Tabela 4.2: Exemplos das unidades de análise

Em seguida, a categorização foi feita através de um ciclo de revisar as unidades de análise existentes entendendo o fator principal que estabelecia uma relação entre elas, por exemplo, na questão sobre as principais fonte de dados uma das categorias atribuídas foi a de *Ensino Superior*, que foi usada para agrupar os destaques de unidades de análise como "enade", "censo do ensino superior", "bolsas cnpq", "cpc"e "idd". CPC e IDD são indicadores do desempenho do ensino superior e sgnificam, respectivamente, Conceito Preliminar de Curso[3] e Indicador de Diferença entre os Desempenhos Observado e Esperado[4].

Já a etapa de comunicação se materializa com a apresentação das categorias e subcategorias resultantes de todo o processo, e as relações reconhecidas entre elas. Essa apresentação é feita no capítulo de resultados.

[3]Conceito Preliminar de Curso (CPC)

[4]Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD)

Capítulo 5

Resultados

Para caracterizar a comunidade que utiliza os dados abertos governamentais brasileiros relacionados à educação fizemos uso de duas bases de dados. A primeira reuniu informações sobre projetos de software, construídos e compartilhados entre a comunidade através do *GitHub*, e a segunda as respostas da pesquisa de opinião realizada com integrantes dessa comunidade. Este capítulo reúne os resultados de sua análise. Para facilitar o entendimento iremos nos referir às fontes de dados como *repositórios* e *respostas do questionário*, respectivamente. No decorrer das seções foi adicionado um destaque para resumir as considerações sobre as características estudadas.

Para permitir um compartilhamento mais amplo da pesquisa construímos uma visualização interativa dos resultados e está disponível de forma online através desse link. Os dados aqui apresentados também estão disponíveis no repositório público da pesquisa[1].

5.1 Caracterização dos Repositórios

De 389 repositórios extraídos, 217 passaram entre os critérios de aprovação e exclusão citados anteriormente. E para cada repositório, além das informações retornadas pela API da plataforma, analisamos e coletamos: organização responsável, objetivo do projeto, fonte dos dados, trabalho realizado e regionalidade do projeto.

[1]Repositório público da pesquisa: https://github.com/Lorenaps/caracterizacao-dados-abertos-gov

5.1.1 Tipo de Perfil Dono do Repositório

Considerando a dinâmica do *GitHub*, descrita na Seção 2.2, um primeiro ponto a ser destacado é o tipo de perfil dono do repositório. Tivemos 85% de perfis de usuários comuns e 15% de perfis de organização. Ambos podem ter vários repositórios, mas o que chama a atenção para entender essa característica é que a criação de uma organização pode retratar a reunião de colaboradores em prol de um objetivo comum e de forma contínua, como é o caso da Prefeitura de São Paulo[2] que, na presente data, tem 145 repositórios que versam sobre dados de educação, merenda escolar e sistemas de gestão.

Observando usuários e organizações encontramos semelhanças entre eles, o que nos levou a agrupá-los em categorias como, iniciativa individual, iniciativa em grupo, governo municipal, centros ou grupos de pesquisa, instituições de ensino superior e canais de mídia/-comunicação, conforme listamos na Tabela 5.1. Nessa mesma tabela apresentamos as organizações encontradas. Um caso interessante é o repositório do projeto "ENEM em Dados"[3][4] que, apesar de vinculado a um perfil de usuário comum, ao acessarmos o link disponível em sua descrição encontramos a explicação de que se trata de um projeto relacionado ao Laboratório de Visualização (LABVIS) em parceria com o Centro de Tecnologia da UFRJ e desenvolvido pelos estudantes Heitor Tomaz e Moisés Colares (Tomaz e Colares 2014).

As iniciativas em grupo retratam pessoas organizadas a fim de facilitar o acesso a dados abertos governamentais da área de educação, como é o caso de "Inep Dados Abertos"[5] e "Open Enade"[6], ou que buscam explorar esse tipo de recurso, como o "Grupo de Transparência e Dados Abertos de São José dos Campos"[7].

Seguindo para a categoria de Instituições de Ensino Superior (IES), temos repositórios relacionados com as instituições em si e repositórios relacionados com as disciplinas de cursos das respectivas instituições. No caso da relação direta com a instituição temos o repositório do "Insper"[8]. E no caso da relação com disciplinas temos os repositórios: "Deep Learning UnB"[9], relacionado a disciplina homônima ofertada na escola de Engenharia de

[2]GitHub da Prefeitura de São Paulo: https://github.com/prefeiturasp
[3]Site do projeto ENEM em Dados - Projeto de Heitor Tomaz e Moisés Colares
[4]Repositório do projeto ENEM em dados
[5]https://github.com/inepdadosabertos/api
[6]https://github.com/OpenEnade/API
[7]https://github.com/sjcdigital/sjc-edu
[8]https://github.com/Insper
[9]https://github.com/deeplearningunb

Quantidade de repositórios	**Tipo de Perfil**	**Tipo de Organização**	**Organização**
183	Usuário	Iniciativa individual	Usuários comuns
1	Usuário	Iniciativa individual	Laboratório de Visualização (LABVIS) e Centro de Tecnologia da UFRJ
20	Organização	IES	Datascienceimih
2	Organização	Governo municipal	Prefeitura de São Paulo
1	Organização	Centro ou grupo de pesquisa	Colaboratório de Desenvolvimento e Participação
1	Organização	Centro ou grupo de pesquisa	Laboratório de Inovação em Políticas Públicas
1	Organização	IES	Insper
1	Organização	IES	Curso de Engenharia de Software - UnB
1	Organização	IES	Deep Learning UnB
1	Organização	IES	ProjetosDW-2018-2
1	Organização	Iniciativa em grupo	Grupo de Transparência e Dados Abertos de São José dos Campos
1	Organização	Iniciativa em grupo	Inep Dados Abertos
1	Organização	Iniciativa em grupo	Open Enade
1	Organização	Mídia	Gênero e Número
1	Organização	Mídia	Núcleo Digital Grupo RBS (Agência de publicidade)

Tabela 5.1: Tipo de usuários e organizações

Software da UnB; e "EPS/MDS"[10] referente às disciplinas Métodos de Desenvolvimento de Software e Engenharia de Produto, relacionado ao Curso de Engenharia de Software da UnB. Nessa mesma linha também destacamos a organização "Datascienceimih"[11] que apresenta repositórios com atividades dos cursos de Ciências de Dados e Bioinformática do Centro Universitário Metodista Izabela Hendrix. Assim como "ProjetosDW-2018-2"[12] que fez a exploração da base de dados do censo escolar brasileiro do INEP para um projeto de Data Warehousing e Business Intelligence do curso de Sistemas de Informação da UFRPE.

Como governo municipal temos o perfil da Prefeitura de São Paulo, citado anteriormente, cujo 2 de seus repositórios foram retornados na coleta de dados via API do *GitHub*. Já os centros ou grupos de pesquisa são representados pelo "Colaboratório de Desenvolvimento e Participação"[13], centro de pesquisa vinculado a USP que trabalha tanto no desenvolvimento de software quanto na formação de pessoas (Colaboratório de Desenvolvimento e Participação), e o "Laboratório de Inovação em Políticas Públicas"[14], que é um laboratório formado por alunos da Fundação Getúlio Vargas - FGV atuando na experimentação e facilitação de inovação no setor público (LAB.ipp - Laboratório de Inovação em Políticas Públicas).

Por fim, em tipo de organização temos os canais de mídia representados pela "Gênero e Número"[15], que fez uma extração e análise dos microdados do censo da educação, e "Núcleo Digital Grupo RBS"[16] que fez um ranking das escolas do Rio Grande do Sul sobre o seu desempenho do ENEM 2018.

> A comunidade se desenvolve, em sua maior parte, a partir de projetos pessoais, porém existem também instituições da sociedade civil como faculdades, centros de pesquisa e canais de mídia/comunicação, como Gênero e Número, e instituições governamentais, como a prefeitura de São Paulo.

[10] https://github.com/fga-eps-mds
[11] https://github.com/datascienceimih
[12] https://github.com/ProjetosDW-2018-2/censo-escolar-inep
[13] https://github.com/COLAB-USP
[14] https://github.com/LABipp
[15] https://github.com/generonumero/educacao
[16] https://github.com/rbsdev/rank-escolas-2018

5.1.2 Tecnologias Utilizadas

Em seguida observamos as tecnologias utilizadas para construção dos projetos de software. O *GitHub* analisa cada repositório e indica a codificação predominante nos arquivos ali contidos (GitHub, Inc 2021). Para esse estudo resolvemos usar o termo tecnologias utilizadas a fim de não denotar única e exclusivamente linguagens de programação. Alguns dos repositórios analisados tiveram como codificação indicada *Rich Text Format*[17] e *TeX*[18], ao verificar a estrutura de arquivos desses repositórios percebemos que além de uma análise de dados da área de educação o repositório armazena também o relato de projetos acadêmicos realizados a partir desse recurso. No total identificamos 18 tecnologias, listadas na Tabela 5.2.

Tecnologia	Repositórios
Jupyter Notebook	116
R	23
Python	18
HTML	15
Não Informado	15
JavaScript	7
Java	6
Ruby	3
CSS	2
Shell	2
TSQL	2
TypeScript	2
C#	1
Go	1
PHP	1
Rich Text Format	1
TeX	1
Vue	1

Tabela 5.2: Tecnologias utilizadas

Jupyter Notebook é a tecnologia mais utilizada, presente em 53% dos repositórios, e apesar de poder ser usada com a linguagem Python também permite a integração com outras linguagens como Ruby, Julia e Scala (Project Jupyter 2021). Em seguida vemos a utilização de R e Python que são linguagens de programação conhecidas por serem eficientes para

[17] https://github.com/krieggerms/Analise-da-Insercao-do-Aluno-Negro-com-BI
[18] https://github.com/fnw/dados-enade-visualizacao-2018-1

análise de dados. Depois notamos a presença de HTML, JavaScript e CSS o que denota a construção de portais e sites como um outro resultado prático das atividades dessa comunidade. Java, Ruby, Shell, TSQL, TypeScript, C#, Go, PHP, Vue também são tecnologias presente nos repositórios, mas em menores proporções. 15 repositórios não tiveram a indicação automática de linguagem, para eles adicionamos o rótulo de Não Informado. Optamos por não tentar identificar manualmente a tecnologia para não aumentar a complexidade e assim comprometer a análise dos demais repositórios, considerando que um repositório pode ter dezenas ou centenas de arquivos.

> Jupyter Notebook se mostra uma tecnologia importante dentro dessa comunidade. Trata-se de uma aplicação WEB de código aberto que permite a criação de relatórios mais descritivos através da junção de células de código com células de texto, onde é possível também a visualização de gráficos e fórmulas matemáticas, sendo assim um recurso muito útil para o compartilhamento de informação com outras pessoas, visto que ao mesmo tempo que pode descrever uma análise através de tabelas, textos e gráficos, permite a revisão e reexecução do código fonte.

5.1.3 Objetivo e Trabalho Realizado

Seguindo para o entendimento do objetivo e trabalho realizado de cada projeto foi preciso estabelecer e revisar, a medida em que se conhecia os repositórios, uma série de critérios a serem seguidos para classificação dos mesmos. Com relação ao objetivo, os repositórios foram agrupados em: informativo, educacional e abertura/ETL. Já para o trabalho realizado as categorias foram: análise, aplicação e disponibilização. A seguir descrevemos os critérios utilizados para cada um deles:

Critérios quanto ao objetivo:

- Informativo ou exploratório: projetos cujo objetivo é informar sobre uma determinada nuance dos dados, mesmo que para apresentar gráficos informativos o projeto tenha que fazer algum processamento ou limpeza dos dados a sua função principal se mantém sendo informar sobre o assunto. Assim como outros projetos descrevem apenas o objetivo de construir WEB crawlers para a extração dos dados;

Exemplo: https://github.com/ggpereira/visio-edu

- Educacional: quando o objetivo está relacionado com uma atividade de uma disciplina ou projeto de um curso, graduação ou pós graduação. Geralmente faz menção a um trabalho em equipe para uma disciplina ou curso. Faz menção a desafios como "Quarentena de Dados Alura"ou "Kaggle". Projetos de pesquisa acadêmica também são considerados nessa categoria;
  Exemplo: https://github.com/yuriats/quarentena-dados

- Abertura/ETL: quando o objetivo é fazer extração, tratamento ou limpeza dos dados, dispor dicionários de dados ou esquemas de banco de dados, ajudando com que a fonte atenda aos princípios de dados abertos. Também engloba projetos que visam fazer a importação desses dados para disponibilizá-los em uma aplicação, sendo assim o projeto responsável por fazer a carga desses dados.
  Exemplo: https://github.com/turicas/cursos-prouni

Critérios quanto ao trabalho realizado:

- Análise: analisa os dados visando informar alguém sobre alguma coisa, fazendo uso de tabelas e gráficos;
  Exemplo: https://github.com/repitta/CienciaDeDadosEducacionais

- Aplicação: constrói uma aplicação ou sua infraestrutura para a apresentação da análise.
  Exemplo: https://github.com/wellingtonf-souza/dash_enem

- Disponibilização: realiza a libertação de dados ou dispõe de meio para realizar seu carregamento em uma aplicação;
  Exemplo: https://github.com/sombriks/microdados-brutos

A partir da análise dos 217 repositórios entendemos que o objetivo mais comum é o informativo com 66%, e que a partir desse objetivo o trabalho mais realizado é o de análise com 90% seguido do trabalho de aplicação com 10%. Na Figura 5.1 exibimos as relações entre o objetivo informativo e os trabalhos realizados.

É importante separar os trabalhos dentro do objetivo informativo porque os projetos que fazem uma análise sem necessariamente formatar esse conteúdo de uma aplicação tem o alcance de um público que precisa ter alguma familiaridade com ferramentas para análise, como o Jupyter Notebook, já quando projeto faz a construção de uma aplicação, um site, ele consegue estender o alcance através de uma interface mais comum e que não precisa de uma ferramenta específica para poder visualizar a informação. Geralmente o segundo tipo de projeto afirma querer "facilitar a visualização".

"Este e um trabalho de analise de dados públicos em relação ao nivel de instrução educacional entre brasileiros de ate 25 anos." - Repositório NivelDeInstrucaoEducacionalBrasil que realizou um trabalho de análise.

"O EducaBrasil.org é um mashup para facilitar a visualização de gastos em educação nos municípios cearenses utilizando dados do TCM-CE." - Repositório EducaBrasil que realizou um trabalho de aplicação.

Figura 5.1: Relação entre o objetivo informativo e os trabalhos realizados

O objetivo **informativo** está presente em 143 repositórios, o que representa 66% do total. Dentro desse objetivo os trabalhos realizados são de **análise** em 129 repositórios (90%) e **aplicação** em 14 repositórios (10%)

O segundo objetivo com maior presença é o educacional com 26% que tem também a

análise como maior trabalho realizado com 86%, seguido de aplicação com 10,5%. Dentro desse objetivo houveram 2 repositórios para os quais não foi possível identificar o trabalho realizado a partir da descrição disponível. Na Figura 5.2 exibimos as relações entre o objetivo educacional e os trabalhos realizados.

O terceiro objetivo mais frequente é abertura/ETL com 8%, que tem disponibilização como maior trabalho realizado com 71%, seguido de aplicação com 29%. Na Figura 5.3 exibimos as relações entre o objetivo abertura/ETL e os trabalhos realizados.

Figura 5.2: Relação entre o objetivo educacional e os trabalhos realizados

O objetivo **educacional** está presente em 57 repositórios, o que representa 26% do total. Dentro desse objetivo os trabalhos realizados são de **análise** em 49 repositórios (86%), **aplicação** em 6 repositórios (10,5%) e para 2 repositórios não foi possível identificar (3,5%).

Figura 5.3: Relação entre o objetivo abertura/etl e os trabalhos realizados

O objetivo **abertura/ETL** está presente em 17 repositórios, o que representa 8% do total. Dentro desse objetivo os trabalhos realizados são **disponibilização** em 12 repositórios (71%) e **aplicação** em 5 repositórios (29%).

Figura 5.4: Trabalho realizado de forma isolada

Observando o trabalho realizado de forma isolada vemos que **análise** esta presente em 178 repositórios (82%), **aplicação** 25 repositórios (12%) e **disponibilização** em 12 repositórios (6%).

Se olharmos o trabalho realizado de forma isolada vemos que análise esta presente em 82% dos repositórios, aplicação em 12% e disponibilização em 6%. Exibimos essa distribuição na Figura 5.4.

> Os projetos realizados majoritariamente objetivam informar determinado público sobre uma nuance dos dados em questão, e considerando as tecnologias utilizadas isso é feito a partir da construção e compartilhamento de relatórios com Jupyter Notebook, Projetos R e Sites. Uma outra característica que podemos extrair é a utilização desse recurso para incentivar e testar o conhecimento de pessoas que estudam a área de análise de dados. E podemos notar também como essa comunidade por vezes trabalha com o objetivo de dar acesso aos dados, expandindo assim o público que pode fazer uso deles.

5.1.4 Popularidade através de *Stars*, *Forks* e *Commits*

Por popularidade e colaboração temos alguns conceitos interessantes trazidos pela plataforma *GitHub*. *Stars*, *Forks* e *Commits* simbolizam, respectivamente, quantidade de pessoas que favoritaram o repositório, quantidade de pessoas que fizeram uma cópia do repositório e quantidade de vezes que aquele repositório recebeu uma modificação.

Analisando a popularidade dos objetivos aqueles que são informativos possuem mais *Forks* e *Stars*. Ao fazer uma cópia do repositório a pessoa usuária denota a intenção de salvar o estado dos arquivos ali presentes e/ou de a partir daquela versão fazer uma alteração no conteúdo e possivelmente submeter essa alteração ao repositório original, estabelecendo assim uma ação de colaboração com o respectivo projeto. Na Figura 5.5 podemos visualizar a popularidade e volume de alterações por objetivo do projeto.

Seguindo a lista de maior popularidade podemos observar que as posições entre os objetivos abertura e educacional se invertem, em comparação a maior quantidade de repositórios por objetivo. Projetos que objetivam a abertura de dados possuem 3,6 vezes mais *Stars* do que os projetos educacionais.

Podemos interpretar essa inversão entendendo que os projetos de abertura geralmente trazem um benefício para mais pessoas além das desenvolvedoras, pois a partir dos seus projetos outras pessoas poderão ter o acesso a dados que antes só estavam acessíveis através de muitas barreiras ou em formato não aberto. Já no caso dos projetos educacionais muitas

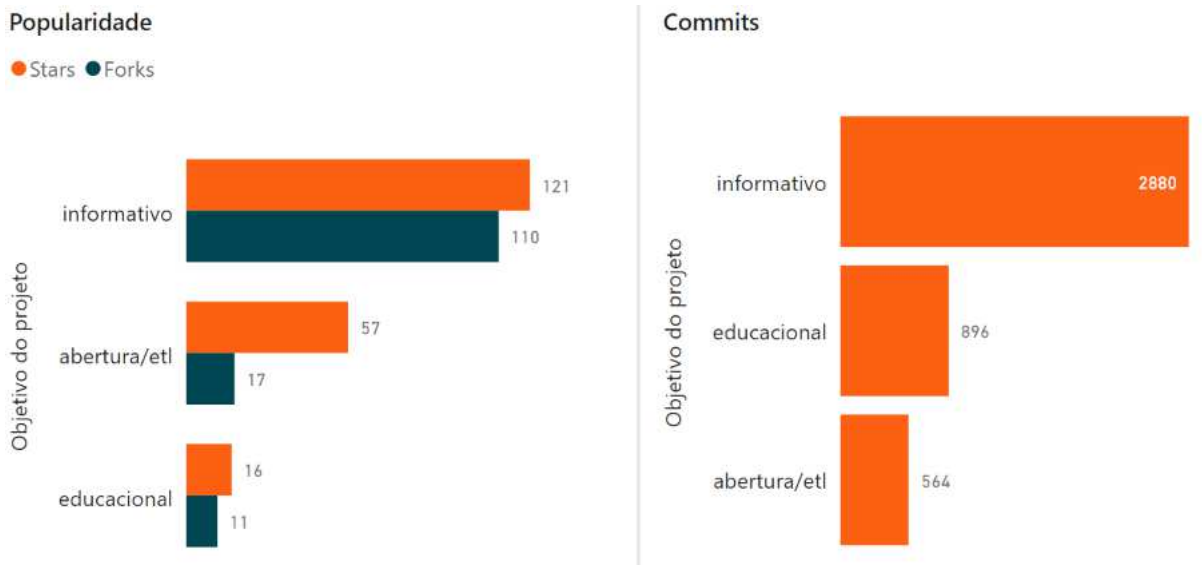


Figura 5.5: Popularidade, através das *Stars* e *Forks*, e volume de alterações por objetivo do projeto.

vezes são repositórios que realizam uma atividade de uma disciplina ou curso e isso é um requisito direto para aprovação das pessoas desenvolvedoras e não necessariamente uma atividade que tem por prioridade levar benefícios direto à comunidade.

Já observando os *Commits* por objetivo, a mesma ordem da maior quantidade de repositórios por objetivo se mantêm: informativo, educacional, abertura/ETL. Onde o objetivo informativo chega a ter 3,2 vezes mais *Commits* do que o educacional. A quantidade de *Commits* representa a quantidade de vezes que um conjunto de alterações foi incorporado ao repositório, e entendendo o caráter descritivo dos projetos que objetivam informar um público sobre determinado contexto, é de se esperar a maior quantidade de edições devido aos ajustes na aparência do relatório, seja nos textos ou visualizações de dados geradas.

Descendo na hierarquia de objetivo para trabalho realizado visualizamos na Figura 5.6 que os projetos que realizam análise são os mais populares, porém notamos a inversão das posições entre disponibilização e aplicação quando comparamos com Figura 5.4, indicando que os projetos de disponibilização são mais populares do que aplicação mesmo estando o segundo em maior número na quantidade de repositórios analisados. Já observando o volume de alterações análise e aplicação ocupam as 2 primeiras posições e possuem, respectivamente, 4,9 e 3,5 vezes mais alterações do que quando o trabalho realizado é disponibilização.

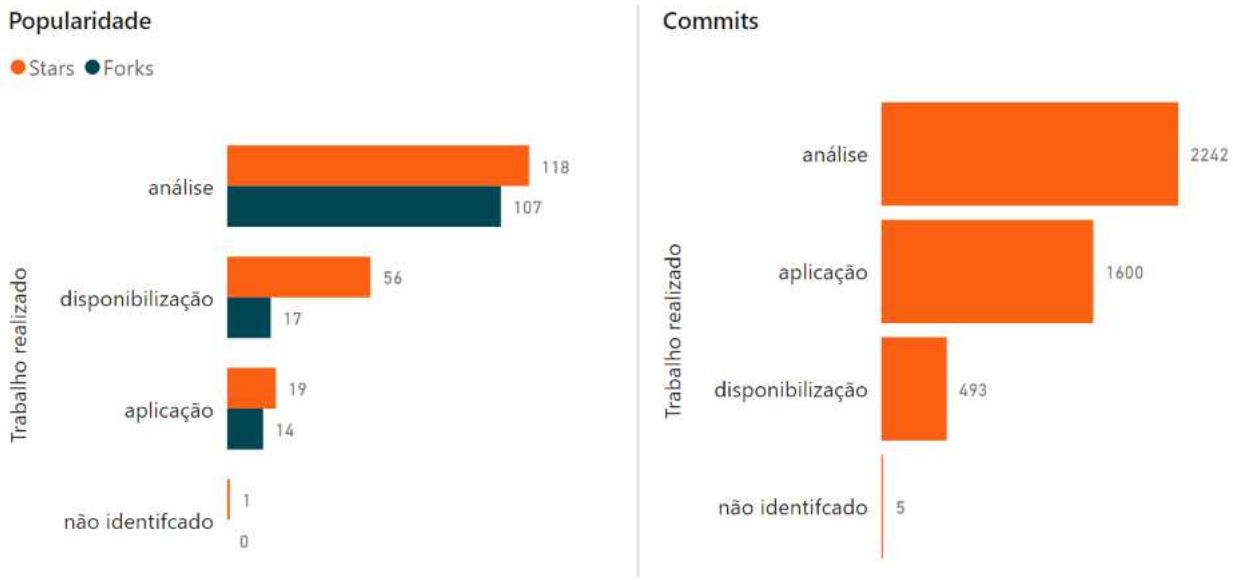


Figura 5.6: Popularidade, através das *Stars* e *Forks*, e volume de alterações por trabalho realizado no projeto.

Ainda sobre popularidade, olhando sob uma perspectiva de colaboração, construímos uma rede de ligações entre os repositórios e seus contribuidores. Dos 217 Repositórios 4 deles ao terem a lista de contribuidores acessada através da API retornam uma lista vazia. Dos 213 repositórios restante destacamos a quantidade de repositórios e a quantidade de pessoas contribuidoras separando por tipo de perfil de usuário, podemos visualizar nas Tabela 5.3 e Tabela 5.4.

Quantidade de repositórios	**Quantidade de pessoas contribuidoras**
149	1
24	2
4	3
3	4

Tabela 5.3: Rede de contribuições x tipo de perfil Usuário

Quantidade de repositórios	Quantidade de pessoas contribuidoras
23	2
8	1
1	7
1	16

Tabela 5.4: Rede de contribuições x tipo de perfil Organização

A rede de contribuição completa pode ser encontrada em Apêndice B. Nas figuras 5.7 e 5.8 é possível visualizar os 2 casos de maior contribuição, com 7 e 16 colaboradores.

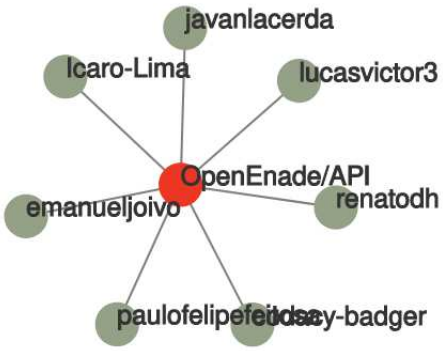


Figura 5.7: Repositório "OpenEnade/API"[19]com 7 contribuintes.

O repositório, que foi classificado com o objetivo **abertura/etl**, se descreve como: "Provê uma maneira programática de acesso aos dados do ENADE.

[19]https://github.com/OpenEnade/API
[20]https://github.com/fga-eps-mds/2017.2-MerendaMais

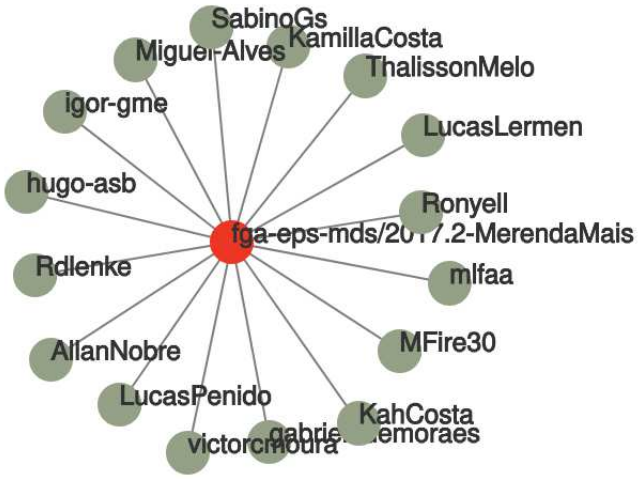


Figura 5.8: Repositório "fga-eps-mds/2017.2-MerendaMais"[20]com 16 contribuintes.
O repositório, que foi classificado com o objetivo **informativo**, se descreve como "Aplicação mobile para auxiliar conselheiros na fiscalização da merenda escolar das escolas de sua região, desde planejar uma visita até a consolidação dos dados.

> Podemos observar que a maioria dos repositórios possuem apenas a pessoa dona do repositório como colaboradora, são 157 repositórios o que representa 74%.

5.1.5 Criação e Atualização dos Repositórios

Seguindo para uma observação temporal, os repositórios analisados também mostram uma evolução do número de projetos criados e atualizados ao longo dos últimos 9 anos.

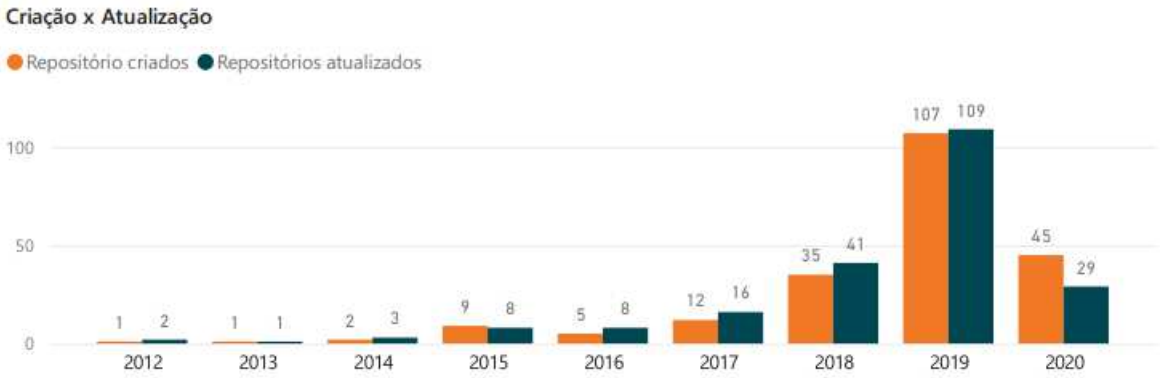


Figura 5.9: Evolução da criação e atualização dos repositórios

Com isso podemos relacionar tanto a popularização da ferramenta *GitHub*, inclusive com

seu uso crescente fora da América do Norte - só o Brasil teve um aumento de 1.691.776 para 2.369.096 de usuários em 2021(GitHub, Inc. 2021). Quanto os esforços para a construção de uma ambiente de participação cidadã em processos decisórios governamentais através das práticas de governo aberto, principalmente as práticas referentes oficinas e cursos de formação(Controladoria-Geral da União 2021), inclusive quando observamos a existência de repositórios com objetivos educacionais, conforme falamos anteriormente.

Uma outra observação temporal que podemos fazer é que nesse conjunto de repositórios, projetos educacionais começam a ser criados em 2016, com uma participação mais expressiva em 2018. Antes disso, encontramos apenas a participação, ainda pequena, de projetos informativos e abertura de dados.

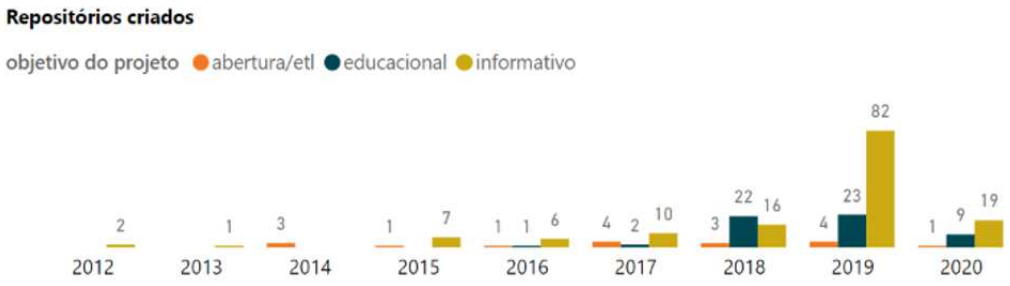


Figura 5.10: Evolução da criação dos repositórios por objetivo do projeto

5.1.6 Fontes de Dados

A medida em que catalogamos cada repositório buscamos também entender as fontes de dados utilizadas e seus possíveis cruzamentos. Para isso consideramos a fonte direta citada nas descrições (Brasil.io, Educação Inteligente, Secretaria de Educação Municipal) e/ou o tema descrito da base de dados, ou seja, se a descrição apresentada ENEM ou Censo escolar consideramos INEP, se dizia PROUNI, consideramos MEC. Na Tabela 5.5 podemos visualizar a lista das fontes citadas. É possível visualizar a frequência maior para o uso do INEP e MEC, mas também a utilização de dados das Secretarias de Educação, o que denota a importância e impacto da publicação de dados por esses órgãos governamentais.

Base de dados	**Repositórios**
INEP	165
MEC	17
Não Identificado	14
Educação Inteligente	2
IBGE	2
IES	2
Portal Brasileiro de Dados Abertos	2
Secretaria de Educação	2
Secretaria Municipal de Educação	2
Censo escolar, PNAE, SIMEC, PRONAF, QSA, CEIS, RAIS e CAQi	1
Pronatec, Ancine e Prouni	1
Brasil.io	1
E-MEC	1
FNDE	1
GEO Open Data e INEP	1
INEP e Sucupira	1
INEP e UNESCO	1
TCM-CE	1

Tabela 5.5: Fonte dos dados

5.1.7 Regionalidade

E por fim, para entendermos a regionalidade do projeto usamos o local indicado no perfil dono do repositório ou site da aplicação, caso existisse um link externo. De 217 repositórios, 107 foram localizados em estados brasileiros, 2 no exterior e para 108 não foi possível identificar uma localidade exata, porém os repositórios em questão estão descritos em português e tratam de dados abertos governamentais brasileiros. Mesmo com uma ausência de precisão para a localidade de 108 repositórios, consideramos um resultado expressivo para discussão pois mostra que dessa coleta existem projetos de 16 estados e presente em todas as regiões brasileiras.

Figura 5.11: Regionalidade dos projetos

5.2 Caracterização das Experiências

As dificuldades trazidas pela pandemia da COVID-19 para projetos de pesquisa que necessitam do contato com outras pessoas ocasionaram uma alta demanda para formas de coleta de dados online. Nesse cenário, a nossa pesquisa de opinião teve 39 respostas, das quais 38 foram válidas, o que consideramos um resultado satisfatório, entendendo a natureza exploratória deste trabalho. Consideramos satisfatório também devido ao alto nível de detalhe fornecido pelos respondentes, o qual iremos explorar através do resultado da Análise Textual Discursiva. Através das experiências descritas nessa pesquisa de opinião, foi possível entender as motivações e problemas de pessoas que integram essa comunidade, assim como suas características e dinâmicas de colaboração.

5.2.1 Perfil dos Respondentes

Para entender o perfil das pessoas alcançadas pelo questionário, estruturamos as perguntas de forma a aferir a experiência dos participantes com o recurso estudado nessa pesquisa e entender suas características sociodemográficas. Os respondentes são 71% do gênero masculino e 29% do gênero feminino, de idade entre 19 e 64 anos como podemos visualizar nas Figura 5.12 e Figura 5.13. Com relação a localização as maiores participações foram dos estados Paraíba, São Paulo, Minas Gerais e Santa Catarina com 34%, 26%, 7% e 7%, respectivamente. No total foram pessoas localizadas em 13 estados brasileiros, podemos ver a lista completa na Tabela 5.6.

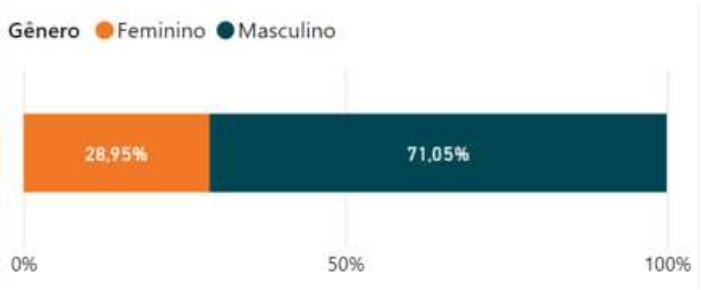


Figura 5.12: Gênero das pessoas respondentes

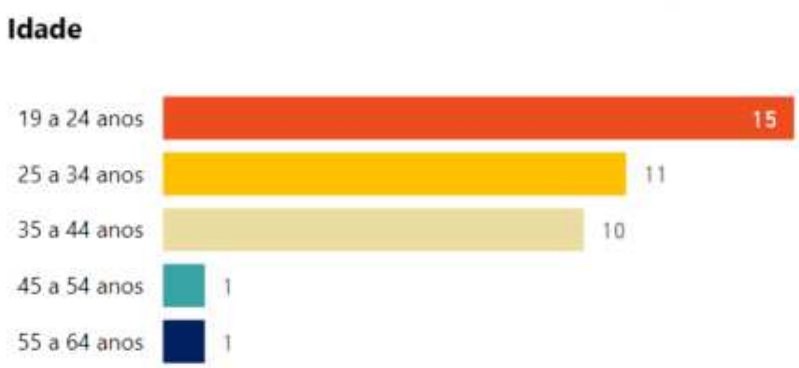


Figura 5.13: Idade das pessoas respondentes

Estado	Quantidade de respondentes
Paraíba - PB	13
São Paulo - SP	10
Minas Gerais - MG	3
Santa Catarina - SC	3
Alagoas - AL	1
Bahia - BA	1
Ceará - CE	1
Distrito Federal - DF	1
Goiás - GO	1
Mato Grosso - MT	1
Pernambuco - PE	1
Piauí - PI	1
Rio de Janeiro - RJ	1

Tabela 5.6: Localidade das pessoas respondentes no momento da participação

Com relação a tempo de experiência com projetos que utilizam dados abertos de forma geral 34% tem entre 3 e 5 anos, 29% entre 1 e 2 anos, 24% menos de 1 ano e 13% mais de 5 anos. Também perguntamos com quantos projetos relacionados com dados abertos governamentais ligados a educação as pessoas já trabalharam, 47% participaram de apenas 1 projeto, 26% participaram de 2, 16% de 5 ou mais, 8% de 3 e 3% de 4 projetos. Esses dados podem ser visualizados nas figuras Figura 5.14 e Figura 5.15.

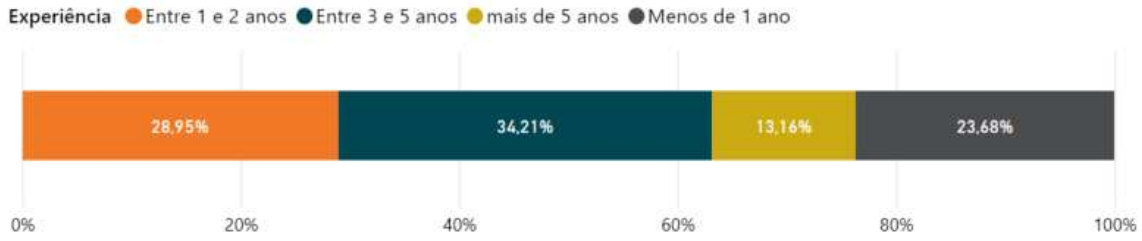


Figura 5.14: Tempo de experiência

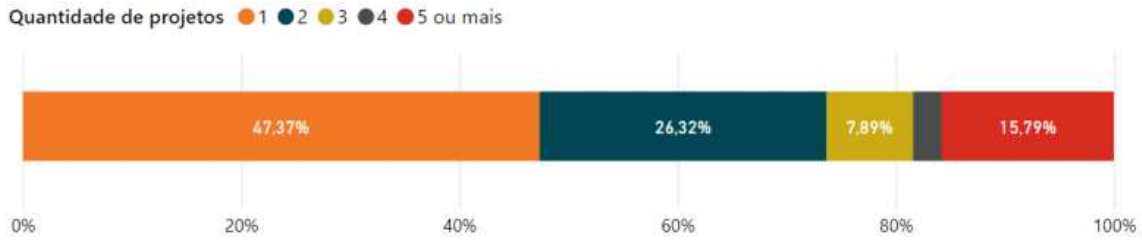


Figura 5.15: Quantidade de projetos que já participaram

Sobre o perfil entendemos que as pessoas respondentes estão localizadas em 13 estados brasileiros e que a presença do gênero masculino é 2,4 vezes maior do que o gênero feminino, estando essas pessoas, em sua maioria, com idade entre 19 a 24 anos e se estendendo até a faixa entre 55 e 64 anos.

87% dessas pessoas tem 5 anos ou menos de experiência com dados abertos governamentais de forma geral e 47% participou de apenas 1 projeto com dados abertos governamentais sobre educação, esse cenário está alinhado com o perfil mais jovem dos respondentes além do, ainda recente, crescimento dos esforços e iniciativas para democratização do acesso a esse recurso.

Muitos são os papéis que uma pessoa pode desempenhar em projetos dessa natureza, esse foi um campo de resposta que deixamos de forma multivalorada. Temos maior presença de estudantes, cientistas e analistas de dados e desenvolvedores de software. Mas para entender a diversidade e alcance da utilização desse tipo de dado vale destacar a presença de professores, consultores da área de educação, designers e jornalistas. Podemos ver a lista completa de papéis desempenhados na Figura 5.16.

Pessoas estudantes ou professoras, geralmente, estão explorando esse tipo de dado e/ou buscando informações sobre um contexto de interesse. Já no papel de consultoria, geral-

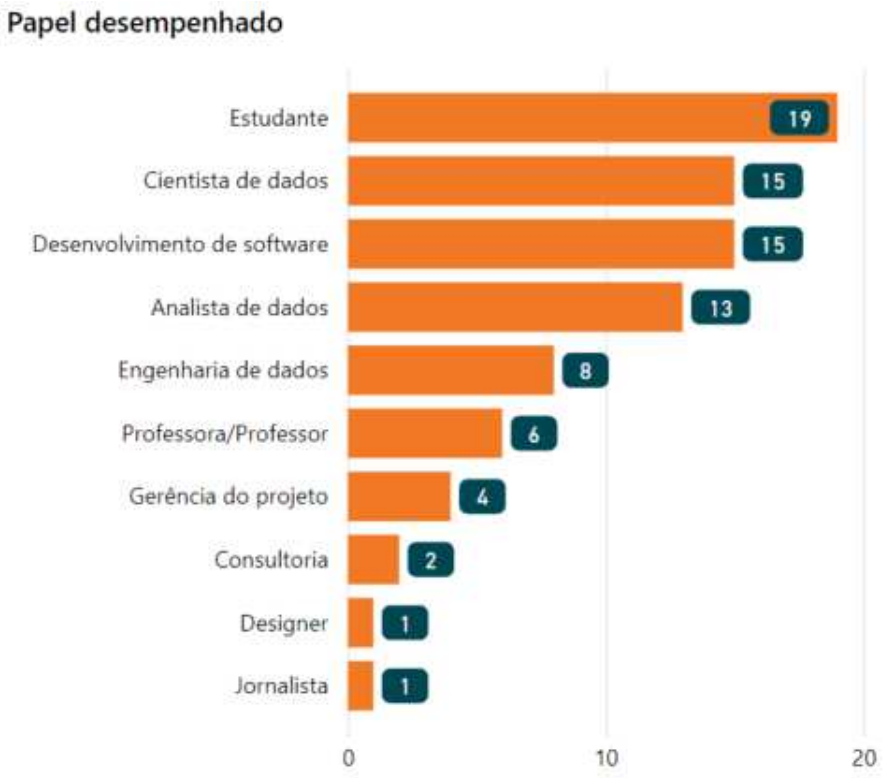


Figura 5.16: Papel desempenhado

mente, são representantes da área de educação e/ou órgão diretamente relacionado ao projeto em questão, e têm a capacidade de sanar duvidas semânticas e metodológicas para a análise de dados.

Com relação a tecnologia utilizada também deixamos como campo multivalorado e, em consonância com os papéis mais desempenhados, temos 33 usos de linguagem de programação e frameworks (Python, R, GoLang, React, SOLR), 13 de planilha eletrônica (Excel ou Google Sheets) e 9 de softwares de visualização de dados (Power BI ou Tableau), conforme Figura 5.17.

Sobre o contexto e utilização dos dados abertos governamentais de forma geral, perguntamos o que os respondentes pensavam sobre fóruns ou canais para tirar dúvidas e compartilhar soluções relacionadas a esse recurso. 58% responderam muito importante, 39% importante e 3% moderado.

Perguntamos também sobre a opinião dos integrantes da comunidade sobre: 1) Facilitar a utilização desse recurso; 2) Divulgar como esse recurso é utilizado; 3) Monetizar o acesso a esse recurso. Sobre a facilitação 97% concordam totalmente e 3% concordam. Sobre a

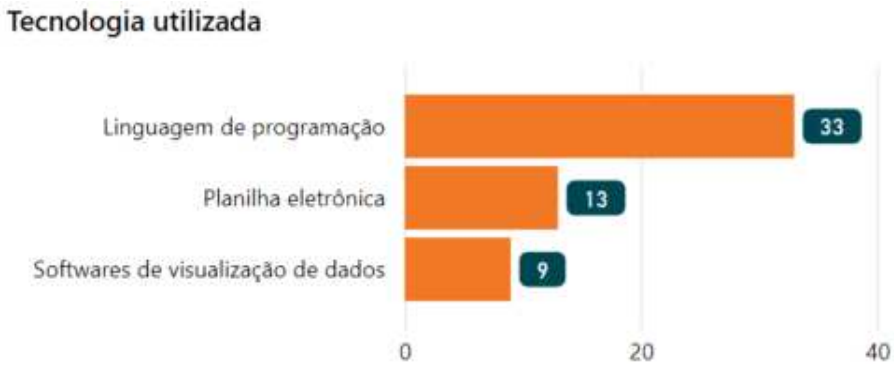


Figura 5.17: Tecnologia utilizada

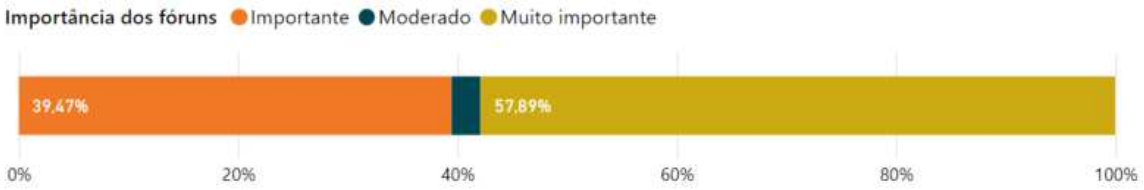


Figura 5.18: Importância dos fóruns

divulgação 90% concordam totalmente, 5% concordam e 5% não estão decididos. Já sobre a monetização 68% discordam totalmente, 13% discordam, 10% não estão decididos, 5% concordam totalmente e 2% concordam. Podemos ver essas distribuições em Figura 5.19, Figura 5.20 e Figura 5.21.

Figura 5.19: Opinião sobre facilitar acesso ao recurso

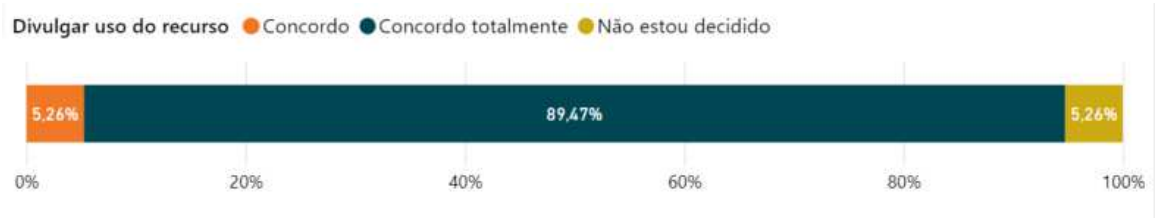


Figura 5.20: Opinião sobre divulgar a utilização

Figura 5.21: Opinião sobre monetização do recurso

5.2.2 Análise das Experiências

Aqui chamamos de experiência as práticas, considerações e motivações relatadas pelos respondentes. As experiências foram examinadas a partir das seguintes questões:

1. Com qual base de dados você mais trabalhou?
2. Qual foi a motivação, sua ou da sua organização, na construção desse projeto?
3. Como você percebeu o impacto/resultado do seu projeto no respectivo público alvo?
4. Quais problemas tecnológicos e/ou na disponibilidade dos dados foram encontrados no desenvolvimento do projeto?
5. No caso de ter conseguido resolver o problema relatado, essa solução foi compartilhada com outras pessoas/organizações? De que forma?
6. Você conhece e/ou indica algum fórum ou canal para trocar conhecimentos sobre dados abertos governamentais de forma geral?

Para análise das respostas foi utilizado o método de Análise Textual Discursiva (ATD) (Moraes e Galiazzi 2007), a seguir iremos apresentar as ideias emergentes da análise, tendo em vista a motivação do estudo em questão. As respostas para cada questão são apresentadas a partir de categorias, e elas foram consideradas a partir das unidades de análise, pertinência com a ideia indicada, relação com as outras ideias presentes e frequência.

No decorrer da apresentação das categorias foram adicionadas algumas citações dos próprios respondentes, para embasar a discussão. Elas serão exibidas da seguinte forma:

"Exemplo de citação de resposta do participante"[Código atribuído ao participante]

Categorias
Censo escolar
ENEM
Ensino superior
Índices de desenvolvimento, investimentos e licitações
Outros dados INEP, MEC e SAEB
Secretaria de educação
Datasus e Censo
Dados de degurança pública e INSS
Sistemas educacionais (moodle)

Tabela 5.7: Categorias da ***Questão 1: Com qual base de dados você mais trabalhou?***

Apresentamos na Tabela 5.7 as categorias que descrevem as bases de dados utilizadas pelos respondentes. Dessas 9 categorias, 5 estão relacionadas com o Inep visto que é o órgão diretamente responsável pela coleta das informações, são elas: Censo escolar; ENEM; Ensino superior; Índice de desenvolvimento, investimentos e licitações e Outros dados INEP, MEC e SAEB.

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) foi criado em 1937 e passou a fazer parte do Ministério da Educação (MEC) em 1997, trata-se do órgão federal responsável por avaliações e exames educacionais, pesquisas estatísticas e indicadores, além da gestão do conhecimento sobre estudos educacionais (Inep, Ministério da Educação 2021). Portanto, uma fonte de dados crucial para iniciativas relacionadas à educação brasileira.

As principais fonte de dados são o Censo escolar, ENEM e Ensino superior. Dentro da categoria ensino superior encontramos referências ao Exame Nacional de Desempenho dos Estudantes (Enade), censo da educação superior, bolsas do CNPq e indicadores de qualidade do ensino superior como Conceito Preliminar de Curso (CPC) e Índice Geral de Cursos (IGC).

"Censo escolar, dados gerais do INEP" [P1]

"microdados do ENEM, censo do ensino superior e censo escolar" [P2]

"microdados do ENEM" [P3]

"Quantidade de Bolsas do CNPq e valor investido pelo órgão"[P4]

Em Índices de desenvolvimento, investimentos e licitações encontramos referências ao Índice de Desenvolvimento da Educação Básica (Ideb), merenda escolar e Sistema Integrado de Monitoramento Execução e Controle (Simec Obras).

"IDEB, censo escolar, licitações de merenda" [P5]

"Dados de compras públicas de merenda escolar" [P6]

"SIMEC obras"[P7]

Já para Outros dados INEP, MEC e SAEB temos referências a Prova Brasil, Programa Mais Educação (PME), Programa Universidade para Todos (Prouni), Sistema Nacional de Informações da Educação Profissional e Tecnológica (Sistec) e Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE). Em Secretaria de educação encontramos referencia a secretarias municipais e estaduais. Para Datasus e censo, Dados de segurança pública e INSS e Sistemas educacionais (moodle) foram menções individuais que falam dessas exatas referências.

"Dados do censo escolar. Dados do SIOPE. Dados abertos no INEP" [P8]

"DataSUS, microdados ENEM" [P9]

"censo escolar, secretaria da educação, plataforma de pesquisa do INSS, secretaria de segurança pú1blica do estado de são paulo" [P10]

"Moodle UFPB uab"[P11]

Com relação a associação entre as fontes destacamos o utilização de Censo escolar com Ideb, Censo escolar com Censo do ensino superior e Censo escolar com o ENEM.

> *"IDEB, Censo Escolar"* [12]

> *"censo escolar do ensino médio e censo do ensino superior, ambos disponibilizados pelo INEP"* [P13]

> *"microdados do ENEM, censo do ensino superior e censo escolar"* [P14]

Ao descrever as motivações em si entendemos que elas podem ter um caráter investigativo, técnico/exploratório, semântico ou de oportunidade, conforme relacionamos na Tabela 5.8.

Categorias	**Subcategorias**
Investigativa	Investigar uma questão/acontecimento
	Fiscalizar algo em específico/Controle social
Técnica ou Exploração	Aprendizado/Curiosidade
	Trabalhos e atividades de disciplinas e/ou cursos
	Facilitar acesso
Semântica	Utilização da semântica dos dados para auxiliar gestores/professores
Oportunidade	Circunstância oportuna

Tabela 5.8: Categorias da ***Questão 2: Qual foi a motivação, sua ou da sua organização, na construção desse projeto?***

Quando falamos em caráter investigativo queremos dizer que essas pessoas fazem a análise de dados a partir de um cenário delimitado do que querem examinar, elas visam **investigar uma questão ou acontecimento** ou **fiscalizar algo em específico**, exercendo assim o controle social almejado pelas iniciativas de transparência e governo aberto. "entender como <questão delimitada>", "fiscalização de", "monitorar", "entender perfil"e "avaliar impactos, investimento, custeio"são algumas formas de como costumam descrever sua motivação.

> *"Fiscalização de compras e qualidade da merenda nas escolas de ensino básico e apoio a grupos civis"* [P15]

> *"Fortalecer controle social, traduzindo a situação das políticas públicas de educação"* [P16]

> *"Acessar dados para análise e construção de narrativas (...) narrativas sobre mulheres negras centrada em dados."* [P17]

> *"Entender o perfil de alunos e instituições de ensino e cruzar esses dados com financiamento universitário"* [P18]

> *"A partir da diminuição de bolsas para projetos no IFPB-CG"* [P19]

A motivação técnica ou exploratória trata das práticas que são atreladas ao **aprendizado e curiosidade** ou a **trabalhos e atividades de disciplinas ou cursos**. Essas pessoas estão buscando por exercitar determinado conhecimento ("treinar", "estudo de ferramentas", "disciplina da universidade"), fazer uma exploração dos dados ou facilitar o seu acesso e entendimento ("analise exploratoria", "facilitar acesso e compreensão").

> *"Entender os dados e criar algumas visualizações"* [P20]

> *"Praticar análise e visualização de dados em uma disciplina."* [P21]

> *"Treinar habilidades de analytics e data science"* [P22]

> *"Matéria da faculdade sobre ciência dos dados."* [P23]

Quando a motivação tem um caráter semântico as pessoas estão interessadas em fazer uma análise daqueles dados e a partir dos resultados semânticos da análise aplicar no cenário educacional novamente, usando essas informações **para auxiliar gestores/professores**.

> *"Utilização profissional para formação de professores"* [P24]

> *"Análise de notas de redações do Enem para fundamentar aulas para o*

> *Ensino Médio (...) Informações do Enade para planejamento de programas de capacitação dos discentes de graduação para realização das provas."* [P25]

> *"A motivação em ambos os projetos foi prover uma ferramental que sintetizasse e mostrasse de uma forma mais amigável para usuário e gestores públicos dados vinculados a indicadores educacionais na educação básica, de modo a facilitar processos decisórios e também levar informação a população."* [P26]

> *"realizar uma análise dos conteúdos que eram efetivamente passados em sala de aula, e o desempenho dos alunos na prova brasil, realizando uma análise comparativa com os outros dados indicadores na secretaria de estado."* [P27]

Por fim, foi encontrada também a motivação de caráter de oportunidade, trata de um momento que foi propício para a participação da pessoa no projeto.

> *"A oportunidade surgiu"* [P28]

Categorias
É utilizado pelo público
Público interessado e participativo/Fizeram sugestão de melhorias
Entendimento sobre uma questão
Não houve ou ainda está em desenvolvimento ou não foi publicado
Dificuldade de engajamento

Tabela 5.9: Categorias da ***Questão 3: Como você percebeu o impacto/resultado do seu projeto no respectivo público alvo?***

Com relação ao impacto/resultado do projeto, apresentado na Tabela 5.9, os respondentes se dividem entrem a utilização pelo público alvo, participação do público em sugestão de melhorias, atingir o entendimento sobre uma questão - isso por si só já se mostra um resultado considerado positivo, principalmente considerando os projetos cuja a motivação é técnica e exploratória. Há ainda aqueles projetos que ainda estão em desenvolvimento, portanto não

foram liberados para utilização, e os caso onde não foi possível notar resultado/impacto ou a percepção foi de dificuldade de engajamento do público.

> *"Demais. Secretários da educação de todo o país aprovaram o projeto, prefeituras usaram como base e muita gente disse que era disso que precisava."* [P29]

> *"O projeto funcionou como complemento às atividades do público alvo, servindo também como apoio a insights."* [P30]

> *"Evidenciou uma diminuição no investimento de bolsas em relação a anos anteriores de crescimento"* [P31]

> *"Realizei uma análise de dados utilizando os dados do CGEE, para entender melhor o panorama de mestres no Brasil, mostrando diferença de remuneração e quantidade de mestres empregados por estado. O impacto acontece ao entender o privilégio de alguns estados em detrimento de outros, na quantidade de vagas de mestrado e na diferença de remuneração quando comparado a outros, facilitando o entendimento das desigualdade no mesmo contexto."* [P32]

> *"O impacto ainda não foi mensurado ou percebido no momento."* [P33]

> *"O projeto ainda está em fase de desenvolvimento (...)"* [P34]

> *"O resultado foi positivo, porém tenho dificuldade na divulgação desse projeto"* [P35]

> *"Pouco engajamento por parte do público alvo, após o levantamentos dos dados e informações"* [P36]

Para a Questão 4 a maior parte dos participantes relatam problemas na estrutura dos da-

dos, falta de padronização da mesma base em anos diferentes e que às vezes mesmo com a presença de um dicionário de dados era difícil entender o que os dados descreviam. Outra queixa é sobre a dificuldade de navegação nos sites que armazenam os dados, alguns respondentes reclamaram sobre burocracia no acesso e *timeout* ao tentar extrair os dados. Algumas bases também apresentam dados incompletos com relação a colunas faltantes ou quando há ausência de dados para anos específicos, para alguns caso foi necessário usar a LAI.

Categorias
Problemas na estrutura e organização dos dados - Falta de padronização - Baixa qualidade
Não houve problema
Dificuldade em usar o site que armazena o dado - Burocracia
Ausência do dado
Dados inconsistentes se comparados com diferentes fontes que deveria falar sobre o mesmo assunto
Foi preciso entrar em contato com o órgão - LAI

Tabela 5.10: Categorias da ***Questão 4: Quais problemas tecnológicos e/ou na disponibilidade dos dados foram encontrados no desenvolvimento do projeto?***

> *"Dificuldade enorme em conseguir fazer um script de leitura dos dados que servisse para os diferentes níveis de ensino e diferentes anos do censo."* [P37]

> *"Conectar diferentes fontes de dados (e.g. integrar tribunais de contas diferentes)"* [P38]

> *"Nossa base havia sido previamente estudada para minimizar o atrito, mas ainda encontramos inconsistências no relacionamento entre as entidades da base e tivemos dificuldade no processamento devido a modelagem confusa. Quando tivemos que expandir a base, em outra base encontramos informações importantes faltando (colunas de identificação) que impossibilitava o cruzamento e consequentemente a geração da informação."* [P39]

> *"O site do INEP não possui link para todos os indicadores em todos os anos no formato desejado (CSV). Algumas bases de dados em alguns anos só*

> *estavam disponíveis em XLS. Os nomes das colunas não são padronizados entre um ano e outro, nem entre outras bases do INEP, o que dificulta a agregação e cruzamento. Os valores categóricos também não são padronizados entre anos e bases de dados, dificultando a análise."* [P40]

> *"Alguns sites não estavam disponíveis, outros estavam mas com abas muito específicas, era preciso um "grande mapa"para saber navegar por ele e encontrar o dado necessário. Em alguns casos foi preciso pedir a LAI, algumas planilhas do Excel vieram em formato pouco intuitivo, foi preciso tratar todos os dados, descobrir as categorias e simplificar a linguagem."* [P41]

Um outro problema que chamou a atenção foi a inconsistência encontrada quando os respondentes comparavam informações de um mesmo assunto porém com uma versão usando microdados e outra versão com os dados agregados, ou mesmo quando comparavam os dados de uma planilha com dados disponíveis no portal.

> *"(...); inconsistências entre microdados e dados agregados; (...)"* [P42]

> *"(...). Fizemos scrapping na plataforma e dados da planilha nem sempre batem com o do portal."* [P43]

Das 29 pessoas que relataram a existência de problemas 23 (79,3%) afirmaram que eles são recorrentes, 3 (10,3%) negaram e 3 (10,3%) marcaram que "Não se aplica". 9 pessoas afirmaram que não tiveram problemas, que os tutorias de dados ajudaram na utilização:

> *"Sem problemas, os tutorias de entendimento da formatação dos dados sempre me atendeu"* [P44]

> *"Não tive nenhum problema, consegui coletar os dados facilmente na pagina do governo"* [P45]

Quando questionados sobre se e como compartilharam as soluções, os respondentes se dividem entre plataformas como *GitHub*, *Stackoverflow* ou relatando que o código fonte é aberto mas também dispõem de documentações com a metodologia utilizada no projeto. Para

alguns projetos realizados em sala de aula a solução foi compartilhada localmente com as pessoas ali envolvidas (grupos de trabalho, professores), e há quem destaque ter recebido o auxílio de outras pessoas para conseguir resolver o problema, o que traz a ideia de comunidade. Outros alegaram que o projeto ainda estava em desenvolvimento, que não conseguiram resolver o problema ou que a solução não foi compartilhada.

Categorias
GitHub/Código aberto/Stackoverflow
Compartilhamento local
Comunidade
Documentação/Publicação
Em desenvolvimento
Não foi compartilhada
Problemas não resolvido

Tabela 5.11: Categorias da ***Questão 5: No caso de ter conseguido resolver o problema relatado, essa solução foi compartilhada com outras pessoas/organizações? De que forma?***

> *"O código fonte é aberto e sempre endossamos no nosso discurso."* [P46]

> *"Foi compartilhada no github e também no stackoverflow"* [P47]

> *"Sugerimos aos colegas, durante as aulas, como poderiam realizar estes ajustes nos dados de maneira mais trivial"* [P48]

> *"Em geral, o professor compartilhava o entendimento que ele já tinha obtido previamente."* [P49]

> *"Na verdade eu fui ajudada por outras pessoas que já haviam passado por essa situação."* [P50]

Por fim, perguntamos de forma aberta também se os participantes conheciam ou indicavam algum fórum ou canal para trocar conhecimentos sobre dados abertos governamentais de forma geral, 24 pessoas não informaram ou afirmaram não conhecer/indicar, dos demais as indicações foram:

- Kaggle
- Canal no Telegram: Dados Abertos
- Canal no Telegram: Colaboradados
- Canal Youtube: Python para Zumbis, Fernando Masanori
- Slack - Data Hackers
- Projeto Serenata de Amor
- Stackoverflow, Facebook
- Fórum Brasileiro de Segurança Pública
- Chat do Brasil IO e Discord da OKBR
- Canal no Telegram: Dados Abertos .BR
- Fórum Jornalismo de Dados

5.3 Considerações Sobre Projetos e Experiências

Figura 5.22: Resumo sobre características dos repositórios

Figura 5.23: Resumo sobre características das experiências

O estudo de descrição e caracterização realizado aqui contou, por um lado, com 217 repositórios de projetos de software, e por outro, com o relato de 38 pessoas sobre suas práticas e experiências. Os repositórios permitiram o olhar sobre questões quantitativas e norteadoras: forma de organização em torno desses projetos, principais objetivos e trabalhos realizados - com destaque para o objetivo educacional, tecnologias relacionadas e tipos de entregáveis dos projetos - incluindo a presença de trabalhos acadêmicos, popularidade dos projetos voltados para a abertura de dados, evolução temporal na quantidade de repositórios e presença em 16 estados brasileiros.

Já os relatos de experiências nos permite entrar no detalhe das vivências relacionadas a integrantes dessa comunidade. Sabemos que os respondentes têm um perfil jovem-adulto, estão localizados em 13 estados brasileiro e sua maioria possui 5 anos ou menos de experiência com projetos dessa natureza. No entendimento da base de dados mais utilizada os respondentes reforçam a importância e utilização dos dados do Inep - importância observada também com os repositórios. E a associação de assuntos, como censo escolar com Ideb, censo escolar com censo do ensino superior e censo escolar com o ENEM.

Com relação à motivação, os participantes se dividem entre investigar uma questão ou acontecimento ou fiscalizar algo em específico, exercendo assim o controle social. Se motivam também por aprendizado e curiosidade ou por trabalhos e atividades de disciplinas ou cursos, estando dessa forma em consonância com os 26% de repositórios que objetivam o fim educacional. E finalmente, também são incentivados por utilizar os resultados da análise para incidir no próprio contexto educacional auxiliando no trabalho de professores e gestores.

Ainda sobre a utilização desses dados em atividades educacionais, relatam que um impacto positivo de seu trabalho vai desde o próprio entendimento de uma questão, até a utilização final pelo público alvo.

Com relação aos problemas encontrados, pudemos ver os relatos já conhecidos na comunidade de dados abertos governamentais de forma geral, sobre falta de padronização e dados incompletos. O que chama atenção nessas experiências é a inconsistência de informações que são de fontes diferentes mas deveriam descrever o mesmo assunto e a dificuldade na navegação pelo site que armazena o dado. Mas também devemos destacar o não enfrentamento de problemas com a base por 9 dos respondentes (23%), e o destaque feito sobre a

importância de tutoriais de utilização dos dados.

Já para a forma de compartilhamento de soluções entendemos que comunidade se vale da utilização de código aberto, plataformas como *GitHub* e *Stackoverflow* e compartilhamento na comunidade que se está inserido no momento, como salas de aula. A partir das experiências, não notamos a existência de um canal central e categorizado sobre problemas recorrentes na utilização dos dados.

Destacamos também a relação de semelhança entre os objetivos e motivações encontradas por essa pesquisa, com os objetivos encontrados na revisão da literatura feita por Santos, Ferreira e Miranda 2017. Em "Dados Abertos Educacionais Uma Revisão da Literatura Brasileira", os autores agruparam os artigos avaliados em 5 categorias que se dividem da seguinte forma:

1. Mineração de dados aplicados a dados abertos educacionais: que poderia se subdividir em sistemas para apoiar a pessoa gestora na tomada de decisão, predição de desempenho de escolar e análise de algoritmos de aprendizagem de máquina focados para dados abertos educacionais;

2. Análise qualitativa: que diz respeito a qualidade dos dados, possibilidades de aplicações e estudos de contextos escolares;

3. Teórico: que descrevem métodos e desafios sobre a utilização dos dados;

4. Sistemas para apoio a aprendizagem colaborativa;

5. Visualização dos dados: aplicações que possibilitam a visualização dos dados de uma determinada região de forma mais acessível.

É possível reconhecer como reencontramos com motivações que visam usar os DAGs educacionais de forma incidir novamente nas estruturas educacionais brasileiras, seguindo, assim, um ciclo: geração dos dados por meio da rotina das instituições e pessoas, consolidação e divulgação desses dados, captura e nova análise a partir da pluralidade de olhares que diferentes atores podem ter, direcionamento e aplicação dessas análises. O que por sua vez poderá impactar sobre as instituições e pessoas, que por sua vez farão a geração de outros dados. O que antes vimos como sistemas para a apoiar a tomada de decisão, revemos

de forma direta como a motivação semântica dos respondentes. Reencontramos também o cunho educacional e tecnológico dos dados, que antes foi identificado como análise de algoritmos de aprendizagem de máquina, vemos como a motivação de exploração de dados atrelada ao aprendizado, dos respondentes, e o objetivo educacional dos repositórios. Até mesmo o que vimos antes como analisar a qualidade dos dados e descrição de métodos e desafios sobre a utilização desses dados, podemos relacionar com o objetivo de abertura/etl dos repositórios, a partir do esforço comum que é feito no entendimento dos problemas em consumir algumas bases dispostas.

O que nos parece novo aqui é a motivação investigativa, que através da fiscalização de um contexto em específico aplica o controle social almejado pelo governo aberto. Mas que também se mostra como uma evolução relacionada à crescente construção da cultura de transparência e participação cidadã.

Quando olhamos para o trabalho de Araújo 2017, também conseguimos estabelecer o paralelo entre a idéia expressada pelos desenvolvedores, ao afirmarem que uma das dificuldades do cidadão seria a necessidade de possuir conhecimento técnico e instrutivo sobre os dados disponíveis, com a visão geral que temos dos problemas listados para utilização de DAGs educacionais - no caso, são pais de alunos que podem não conseguir acompanhar como foi o desempenho da escola dos filhos no ENEM ou como é feito o investimento nas escolas da sua região.

Capítulo 6

Conclusão

Esse estudo possibilitou entender e refletir sobre características e dinâmicas da comunidade que utiliza dados abertos governamentais brasileiros da área de educação. Foram 217 repositórios de projetos de software e 38 relatos de participantes dessa comunidade.

Descrevemos um perfil de quem trabalha com esse recurso e sua forma de organização. Em sua maioria, são iniciativas individuais porém com organizações de diversidade expressiva. São setores como governo municipal, instituições de ensino e iniciativas cidadãs, mídia de comunicação e centros de pesquisa. O perfil dos respondentes é jovem-adulto com 5 anos ou menos de experiência com DAGs no geral. Projetos e respondentes, juntos, estão em 17 estados brasileiros.

Um aspecto interessante é a segunda mais expressiva participação ser de instituições de ensino superior, e através de atividades de disciplinas, o que denota o uso proveitoso dessas bases de dados na formação de estudantes. Para além do controle social, elas servem como recurso a fortalecer o ensino, seja no processamento técnico dessas grandes bases (predição de dados do ENEM) ou na reflexão necessária para ir além dos quantitativos ali contidos e relacionar aquele contexto com outras dinâmicas sociais (demanda de vagas de alunos com necessidades especiais, avaliar a infraestrutura em escolar para pessoas com necessidades).

Os projetos realizados, majoritariamente, objetivam informar determinado público sobre uma nuance dos dados em questão, se preocupando com uma forma agradável de comunicar aquele contexto. E por vezes, também possuem o objetivo de dar acesso aos dados, expandindo assim o público que pode fazer uso deles, o que é imprescindível para a construção da cultura de participação cidadã e governo aberto.

A investigação/fiscalização de uma questão ou acontecimento e a vontade de incidir diretamente na área de educação auxiliando pessoas gestoras e professoras, são motivações com potencial de ter resultados que retroalimentam esse ecossistema.

A partir dos resultados alcançados, esperamos que o entendimento desse contexto possa proporcionar benefícios diretos para organizações governamentais e a sociedade civil. As instituições públicas responsáveis pelos dados disponíveis podem usar os problemas relatados para a correção e prevenção de inconsistências, inclusive em outros contextos. Assim como, investir em infraestrutura e arquitetura da informação dos sites, para evitar os problemas como dificuldade na navegação e falha por demora no tempo de resposta do site.

Entender o perfil de pessoas e suas experiências pode ajudar no planejamento de ações para o fortalecimento dessa comunidade. Sabendo que para lidar com os problemas relatados é preciso um conhecimento em extração e tratamento de dados e que essa não é uma formação básica nem acessível da maioria da população, investir na capacitação e acesso de, pelo menos, trabalhadores da área de educação se torna primordial para conseguirmos uma participação cidadã realmente efetiva, principalmente pensando nas estruturas educacionais com menores recursos e maior quantidade de alunos.

Ao entender as formas de compartilhamento utilizadas, a comunidade pode se valer também da criação de um canal central e categorizado para o compartilhamento de soluções e troca de experiências, inclusive entre fornecedores e consumidores de DAGs educacionais. Outras pesquisas também podem utilizar e evoluir os materiais utilizado nessa, a exemplo das palavras chaves de busca, script de extração de dados do *GitHub* e dados extraídos, que estão disponíveis no repositório público da pesquisa, também no *GitHub*.

Com relação às limitações e vulnerabilidades do estudo, acreditamos que a análise manual de um conjunto de dados não estruturado é um ponto com potencial para a influência de vieses e que por isso, exige atenção, rigor e frequente revisão dos critérios e categorizações realizadas. Pensando nisso, deixamos descrito na dissertação os critérios utilizados para a análise dos repositórios, e com relação às categorias advindas da Análise Textual Discursiva.

Sobre a representatividade dos repositórios, salientamos que são diretamente relacionados com as palavras chave de busca, e que buscamos no processo de coleta refinar as palavras chaves a partir dos resultados que obtínhamos nos testes com a API, e também através de pesquisa nos sites oficiais relacionados com a educação brasileira. Porém, ainda pode existir

uma possibilidade de aumento do alcance com palavras chaves menos formais, o que talvez seria possível de avaliar a partir do conhecimento de grupos mais específicos dentro desse ecossistema, como conselhos escolares ou secretarias de educação.

Para a pesquisa de opinião, destacamos que se trata de um trabalho exploratório, com uma amostra com 38 respondentes coletada em um período pandêmico, e que isso pode ter interferido na diversidade e alcance dos respondentes. Além disso, destacamos também, inclusive como trabalho futuro, a necessidade da extensão desse estudo focado em dinâmicas dos órgãos educacionais municipais, para que seja possível estudar a comunidade ouvindo um de seus segmentos basilares. Trazemos também como trabalho futuro uma forma de ampliar as pessoas participantes desse estudo, de forma a alcançarmos uma visão expandida sobre fornecedores, consumidores e as pessoas que atuam diretamente na geração desses dados - professores e alunos.

Por fim, entendemos as realizações que a comunidade vem alcançando, suas dificuldades e um campo amplo para evolução e melhorias, para o qual acreditamos que a construção coletiva é fundamental. Por isso, e entendendo as relações com os trabalhos de Araújo 2017 e Santos, Ferreira e Miranda 2017, fazemos a recomendação de fortalecimento dessa comunidade e seus benefícios através do investimento em eventos que possam cultivar a ideia de colaboração entre governo, secretarias de educação e diversos outros atores da sociedade, como os *Hackathons* e *Hackfests*, porém voltando o olhar para as dinâmicas da rede pública municipal de ensino, que são responsáveis por 48,4% dos estudantes brasileiros (Cristaldo 2021).

Bibliografia

Araújo 2017 ARAÚJO, N. M. d. ***Dados abertos do governo brasileiro: entendendo as perspectivas de fornecedores de dados e desenvolvedores de aplicações ao cidadão***. Dissertação (Mestrado) — UFRN, 2017.

Attard et al. 2015 ATTARD, J. et al. **A systematic review of open government data initiatives**. *Government Information Quarterly*, Elsevier, v. 32, n. 4, p. 399–418, 2015.

Beghin, Zigoni et al. 2014 BEGHIN, N.; ZIGONI, C. et al. **Avaliando os websites de transparência orçamentária nacionais e subnacionais e medindo impactos de dados abertos sobre direitos humanos no Brasil**. Instituto de Estudos Socioeconômicos, 2014.

Bitbucket 2019 BITBUCKET. ***Celebrating 10 million Bitbucket Cloud registered users***. 2019. Disponível em: <https://bitbucket.org/blog/celebrating-10-million-bitbucket-cloud-registered-users>. Acesso em: 01.02.2022.

Brasil 2011 BRASIL. ***Lei nº 12.527, de 18 de novembro de 2011. Art. 1º***. [S.l.], 2011. Disponível em: <http://www.presidencia.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm>. Acesso em: 10.01.2020.

Brasil 2011 BRASIL. ***Lei nº 12.527, de 18 de novembro de 2011. Seção I - Do Pedido de Acesso***. [S.l.], 2011. Disponível em: <http://www.presidencia.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm>. Acesso em: 10.01.2020.

Chen e Ganapati 2018 CHEN, C.; GANAPATI, S. Is transparency the best disinfectant? 2018. Disponível em: <https://www.opengovpartnership.org/documents/is-transparency-the-best-disinfectant/>. Acesso em: 21.01.2022.

CKAN 2018 CKAN. ***What is CKAN?*** 2018. Disponível em: <http://docs.ckan.org/en/2.9/user-guide.html#what-is-ckan>. Acesso em: 01.02.2022.

Colaboratório de Desenvolvimento e Participação COLABORATóRIO DE DESENVOLVIMENTO E PARTICIPAçãO. ***COLAB: Histórico e Missão***. Disponível em: <https://colab.each.usp.br/historico-e-missao/>. Acesso em: 14.09.2021.

Controladoria-Geral da União 2021 CONTROLADORIA-GERAL DA UNIãO. ***5º Plano de Ação Nacional em Governo Aberto***. [S.l.], 2021. Disponível em: <https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/5o-plano-de-acao-brasileiro/5-plano-acao-nacional-final>. Acesso em: 17.01.2022.

Controladoria-Geral da União 2021 CONTROLADORIA-GERAL DA UNIãO. ***Relatório Final de Autoavaliação - 4º Plano de Ação Nacional em Governo Aberto***. [S.l.], 2021. 24-30 p. Disponível em: <https://www.gov.br/cgu/pt-br/governo-aberto/noticias/2021/11/relatorio-final-de-autoavaliacao-do-4o-plano-esta-aberto-para-contribuicoes/relatorio-de-autoavaliacao-final-4plano.pdf>. Acesso em: 29.11.2021.

Cristaldo 2021 CRISTALDO, H. ***Censo Escolar 2020 aponta redução de matrículas no ensino básico***. 2021. Disponível em: <https://agenciabrasil.ebc.com.br/educacao/noticia/2021-01/censo-escolar-2020-aponta-reducao-de-matriculas-no-ensino-basico>. Acesso em: 23.01.2022.

Datapedia 2015 DATAPEDIA. ***Para Cidadãos***. 2015. Disponível em: <https://datapedia.info/#cidadaos>. Acesso em: 11.11.2019.

Dietrich et al. 2009 DIETRICH, D. et al. Open data handbook. *Open Knowledge International*, 2009.

Eaves 2009 EAVES, D. ***The Three Laws of Open Government Data***. 2009. Disponível em: <https://eaves.ca/2009/09/30/three-law-of-open-government-data/>. Acesso em: 16.01.2022.

Gerhardt e Silveira 2009 GERHARDT, T. E.; SILVEIRA, D. T. *Métodos de pesquisa*. [S.l.]: Plageder, 2009.

Git 2014 GIT. ***GitHub - Account Setup and Configuration***. 2014. Disponível em: <https://git-scm.com/book/en/v2/GitHub-Account-Setup-and-Configuration>. Acesso em: 08.01.2022.

Git s.d. GIT. ***Git***. s.d. Disponível em: <https://git-scm.com/>. Acesso em: 08.01.2022.

GitHub, Inc 2021 GITHUB, INC. ***GitHub Docs: Glossário do GitHub***. 2021. Disponível em: <https://docs.github.com/pt/get-started/quickstart/github-glossary>. Acesso em: 17.10.2021.

GitHub, Inc 2021 GITHUB, INC. ***GitHub Docs: Sobre linguagens do repositório***. 2021. Disponível em: <https://docs.github.com/pt/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/about-repository-languages>. Acesso em: 17.10.2021.

GitHub, Inc. 2021 GITHUB, INC. ***Let's look back at the code and communities built on GitHub this year***. 2021. Disponível em: <https://octoverse.github.com/#lets-look-back-at-the-code-and-communities-built-on-git-hub-this-year>. Acesso em: 01.02.2021.

GitHub, Inc. 2021 GITHUB, INC. ***The 2021 State of the Octaverse - Future of developer communities***. 2021. Disponível em: <https://octoverse.github.com/#future>. Acesso em: 29.11.2021.

GitLab 2022 GITLAB. ***About GitLab***. 2022. Disponível em: <https://about.gitlab.com/company/>. Acesso em: 01.02.2022.

Gonçalves 2014 GONÇALVES, S. The effects of participatory budgeting on municipal expenditures and infant mortality in brazil. *World development*, Elsevier, v. 53, p. 94–110, 2014.

INEP e MEC 2006 INEP; MEC. *Micodados do Censo Escolar 1995*. [S.l.: s.n.], 2006.

Inep, Ministério da Educação 2021 INEP, MINISTéRIO DA EDUCAçãO. ***Inep: Sobre***. 2021. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/institucional/sobre>. Acesso em: 08.01.2022.

LAB.ipp - Laboratório de Inovação em Políticas Públicas LAB.IPP - LABORATóRIO DE INOVAçãO EM POLíTICAS PúBLICAS. ***About***. Disponível em: <https://medium.com/lab-ipp/about>. Acesso em: 14.09.2021.

Ministério da Transparência e Controladoria-Geral da União 2018 MINISTéRIO DA TRANSPARêNCIA E CONTROLADORIA-GERAL DA UNIãO. ***4º Plano de Ação Nacional em Governo Aberto***. [S.l.], 2018. Disponível em: <http://governoaberto.cgu.gov.br/a-ogp/planos-de-acao/4o-plano-de-acao-brasileiro/4o-plano-de-acao-nacional_portugues.pdf>. Acesso em: 11.11.2019.

Moraes 2003 MORAES, R. Uma tempestade de luz: a compreensão possibilitada pela análise textual discursiva. *Ciência & Educação (Bauru)*, SciELO Brasil, v. 9, n. 2, p. 191–211, 2003.

Moraes e Galiazzi 2007 MORAES, R.; GALIAZZI, M. do C. *Análise textual: discursiva*. [S.l.]: Editora Unijuí, 2007.

Oliveira et al. 2016 OLIVEIRA, M. I. S. et al. **Open government data portals analysis: the Brazilian case**. In: ACM. *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*. [S.l.], 2016. p. 415–424.

Onde Codar em Salvador 2021 ONDE CODAR EM SALVADOR. ***Onde Codar em Salvador***. 2021. Disponível em: <https://github.com/devssa/onde-codar-em-salvador>. Acesso em: 01.02.2022.

Open Government Partnership 2018 OPEN GOVERNMENT PARTNERSHIP. ***Guia do Governo Aberto para Céticos***. [S.l.], 2018. Disponível em: <https://www.opengovpartnership.org/wp-content/uploads/2018/07/Guia-de-Governo-Abierto-para-Ceticos.pdf>. Acesso em: 17.01.2022.

Open Government Partnership 2022 OPEN GOVERNMENT PARTNERSHIP. ***Members***. 2022. Disponível em: <https://www.opengovpartnership.org/our-members/>. Acesso em: 16.01.2022.

Open Knowledge Foundation s.d. OPEN KNOWLEDGE FOUNDATION. ***Open Government Data and Content***. s.d. Disponível em: <https://opendefinition.org/government/>. Acesso em: 10.01.2020.

Open Knowledge Foundation s.d. OPEN KNOWLEDGE FOUNDATION. ***The Open Definition***. s.d. Disponível em: <https://opendefinition.org/>. Acesso em: 16.01.2022.

Parlametria 2019 PARLAMETRIA. ***Dados (mais) abertos no congresso - Barreiras encontradas e propostas para avançar***. 2019. Disponível em: <https://parlametria.org/assets/reports/GargalosdeTransparênciadeDadosnoCongresso.pdf>. Acesso em: 10.01.2020.

Penteado, Bittencourt e Isotani 2017 PENTEADO, B.; BITTENCOURT, I. I.; ISOTANI, S. **Dados abertos educacionais no Brasil e sua preparação para os dados abertos na web**. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 526.

Penteado e Isotani 2017 PENTEADO, B.; ISOTANI, S. Dados abertos educacionais: que informações temos disponíveis. In: *VI Congresso Brasileiro de Educação*. [S.l.: s.n.], 2017. v. 4, p. 1933–1938.

Pinho 2018 PINHO, M. D. C. Dados abertos governamentais e democracia digital: o estado da arte e uma aplicação aos portais de dados abertos de seis prefeituras brasileiras. Faculdade de Comunicação da UFBA, 2018.

Pizza De Dados 2021 PIZZA DE DADOS. ***Guia do Cientista de Dados das Galáxias***. 2021. Disponível em: <https://github.com/PizzaDeDados/datascience-pizza>. Acesso em: 01.02.2022.

Portal Brasileiro de Dados Abertos 2020 PORTAL BRASILEIRO DE DADOS ABERTOS. ***O que são dados abertos?*** 2020. Disponível em: <http://dados.gov.br/pagina/dados-abertos>. Acesso em: 10.01.2020.

Project Jupyter 2021 PROJECT JUPYTER. ***The Jupyter Notebook***. 2021. Disponível em: <https://jupyter.org/index.html>. Acesso em: 17.10.2021.

PyLadies Paraíba 2021 PYLADIES PARAíBA. ***O que posso fazer com Python?*** 2021. Disponível em: <https://github.com/pyladiespb-org/python-world>. Acesso em: 01.02.2022.

Repositório de vagas Front-End 2015 REPOSITóRIO DE VAGAS FRONT-END. ***Questões para entrevista de profissionais Front-end***. 2015. Disponível em: <https://github.com/vagasfrontend/perguntas-para-entrevista>. Acesso em: 01.02.2022.

Santos, Ferreira e Miranda 2017 SANTOS, P.; FERREIRA, R.; MIRANDA, P. Dados abertos educacionais: Uma revisao da literatura brasileira. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 11.

Schalkwyk, Willmers e Czerniewicz 2014 SCHALKWYK, F. van; WILLMERS, M.; CZERNIEWICZ, L. Open data in the governance of south african higher education. *Available at SSRN 2557342*, 2014.

Serenata de Amor 2016 SERENATA DE AMOR. ***Operação Serenata de Amor***. 2016. Disponível em: <https://serenata.ai/>. Acesso em: 11.11.2019.

Tauberer 2014 TAUBERER, J. Open government data definition: The 8 principles of open government data. *Open Government Data: The Book*, 2014.

Tomaz e Colares 2014 TOMAZ, H.; COLARES, M. ***ENEM em dados***. 2014. Disponível em: <http://heitortomaz.github.io/EnemGit/index.html>. Acesso em: 14.09.2021.

Apêndice A

Pesquisa de Opinião

Utilização de dados abertos governamentais brasileiros sobre a área de educação.

Esse questionário tem o objetivo de entender as motivações e dificuldades na utilização dos dados abertos governamentais brasileiros sobre educação (Exemplos: dados sobre o censo escolar, ENADE, PROUNI, licitações para escolas). Sua aplicação faz parte do projeto de mestrado intitulado: Caracterização da Comunidade que Utiliza Dados Abertos Governamentais Brasileiros Sobre a Área de Educação. O projeto é realizado pela mestranda Lorena Pereira com orientação do Prof. Dr. João Arthur Brunet no Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande.

A pesquisa se destina a pessoas que já utilizaram esse tipo de dado independentemente da finalidade.

Ter essas informações é importante porque entender melhor as características dessa comunidade poderá beneficiar instituições do governo que buscam otimizar seus serviços de divulgação de dados abertos, e a própria sociedade civil através do compartilhamento de experiências.

Esse questionário é anônimo, o login realizado foi apenas para evitar que você responda mais de uma vez. Porém em nenhum momento seu e-mail é registrado junto às respostas fornecidas aqui.
A sua ação de respondê-lo é um ato voluntário, pelo qual eu agradeço muito!
Você pode desistir a qualquer momento sem penalização ou prejuízo a você.

A duração estimada é de 10 minutos.

Obrigada pela participação!
Qualquer dúvida pode entrar em contato através do email:
lorenapereira@copin.ufcg.edu.br

*Obrigatório

http://youtube.com/watch?v=Xd-DotiTvvk

1. De quantos projetos ou atividades com dados abertos governamentais educacionais brasileiros você já participou? *

Dados abertos governamentais relacionados a educação brasileira (Exemplos: microdados do ENEM, dados da secretaria de educação do seu estado, censo escolar, entre outros).

Marcar apenas uma oval.

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5 ou mais

2. Com qual base de dados você mais trabalhou? *

Dados abertos governamentais relacionados a educação brasileira (Exemplos: microdados do ENEM, dados da secretaria de educação do seu estado, censo escolar, entre outros).

_____

3. Qual foi a motivação, sua ou da sua organização, na construção desse projeto? *

Em caso de mais de um projeto, por favor descrever o projeto que na sua visão foi o mais marcante. Mas pode ficar a vontade e descrever mais de um, se assim desejar.

_____

_____

_____

_____

_____

4. Como você percebeu o impacto/resultado do seu projeto no respectivo público alvo? *

Em caso de mais de um projeto, por favor descrever o projeto que na sua visão foi o mais marcante. Mas pode ficar a vontade e descrever mais de um, se assim desejar.

5. A partir de que ferramenta você fez uso desses dados? *

Marque todas que se aplicam.

- [ ] Softwares de visualização de dados (Exemplo: Power BI ou Tableau)
- [ ] Planilha eletrônica (Exemplo: Excel ou Google Sheets)
- [ ] Scripts em alguma linguagem de programação (Exemplo: Python ou R)

Outro: [ ] _____

6. Qual papel que você desempenhou nesse projeto ou experiência em questão? *

Marque todas que se aplicam.

- [ ] Consultoria da área de educação e/ou órgão diretamente relacionado
- [ ] Desenvolvimento de software (Exemplo: sites; sistemas; plataformas)
- [ ] Engenharia de dados
- [ ] Cientista de dados
- [ ] Analista de dados
- [ ] Gerência do projeto
- [ ] Jornalista
- [ ] Designer
- [ ] Estudante (Explorando esse tipo de dado e/ou buscando informações sobre um contexto de interesse)
- [ ] Professora/Professor (Explorando esse tipo de dado e/ou buscando informações sobre um contexto de interesse)

Outro: [ ] _____

7. Quais problemas tecnológicos e/ou na disponibilidade dos dados foram encontrados no desenvolvimento do projeto? *

Exemplos de problemas: site indisponível, formato do dado não era aberto, dificuldade em entender o que os dados descreviam, entre outros. Em caso de não ter encontrado problema pode responder com "Não se aplica."

8. Você diria que esses problemas são recorrentes? *

Marcar apenas uma oval.

- Sim
- Não
- Não se aplica

9. No caso de ter conseguido resolver o problema relatado, essa solução foi compartilhada com outras pessoas/organizações? De que forma? *

Em caso de não ter encontrado problema ou resolvido pode responder com "Não se aplica."

Está quase acabando. Faltam apenas 7 perguntas.

São mais 6 perguntas fechadas e 1 aberta de resposta curta.

Dados abertos governamentais de forma geral

10. O que você pensa sobre fóruns ou canais para tirar dúvidas e/ou compartilhar soluções relacionadas à utilização de dados abertos governamentais de forma geral? *

Marcar apenas uma oval.

- ◯ Não é importante
- ◯ Pouco importante
- ◯ Moderado
- ◯ Importante
- ◯ Muito importante

11. Você conhece e/ou indica algum fórum ou canal para trocar conhecimentos sobre dados abertos governamentais de forma geral?

_____

12. Com relação à utilização de dados abertos governamentais, qual a sua opinião sobre: *

Marcar apenas uma oval por linha.

	Discordo totalmente	Discordo	Não estou decidido	Concordo	Concordo totalmente
Facilitar a utilização desse recurso	◯	◯	◯	◯	◯
Divulgar como esse recurso é utilizado	◯	◯	◯	◯	◯
Monetizar o acesso a esse recurso	◯	◯	◯	◯	◯

Sobre Você

Perguntas para que possamos entender um perfil das pessoas participantes desse questionário.

13. Há quanto tempo você atua em projetos dessa natureza? *

Marcar apenas uma oval.

- Menos de 1 ano
- Entre 1 e 2 anos
- Entre 3 e 5 anos
- mais de 5 anos

14. Qual a sua idade? *

Marcar apenas uma oval.

- Até 18 anos
- 19 a 24 anos
- 25 a 34 anos
- 35 a 44 anos
- 45 a 54 anos
- 55 a 64 anos
- 65 ou mais

15. Como você se identifica com relação a gênero? *

Caso as opções existentes não lhe represente, por favor, adicione a identidade de gênero que melhor lhe representa usando o campo aberto na opção outros.

Marcar apenas uma oval.

- ◯ Agênero
- ◯ Não binário
- ◯ Feminino
- ◯ Masculino
- ◯ Prefiro não responder
- ◯ Outro: _____

16. De que estado você está respondendo esse questionário? *

Marcar apenas uma oval.

- Estou fora do Brasil
- Acre (AC)
- Alagoas (AL)
- Amapá (AP)
- Amazonas (AM)
- Bahia (BA)
- Ceará (CE)
- Distrito Federal (DF)
- Espírito Santo (ES)
- Goiás (GO)
- Maranhão (MA)
- Mato Grosso (MT)
- Mato Grosso do Sul (MS)
- Minas Gerais (MG)
- Pará (PA)
- Paraíba (PB)
- Paraná (PR)
- Pernambuco (PE)
- Piauí (PI)
- Rio de Janeiro (RJ)
- Rio Grande do Norte (RN)
- Rio Grande do Sul (RS)
- Rondônia (RO)
- Roraima (RR)
- Santa Catarina (SC)
- São Paulo (SP)
- Sergipe (SE)
- Tocantins (TO)

Apêndice B

Rede de Contribuição

Rede de contribuição construída dentro da ferramenta Power BI a partir dos dados coletados sobre os repositórios via API do GitHub.

Repositórios coletados em: 19 de maio de 2020 Repositórios analisados entre: 19 de maio e 27 de outubro de 2020 Página inicial

Rede de Contribuição

213

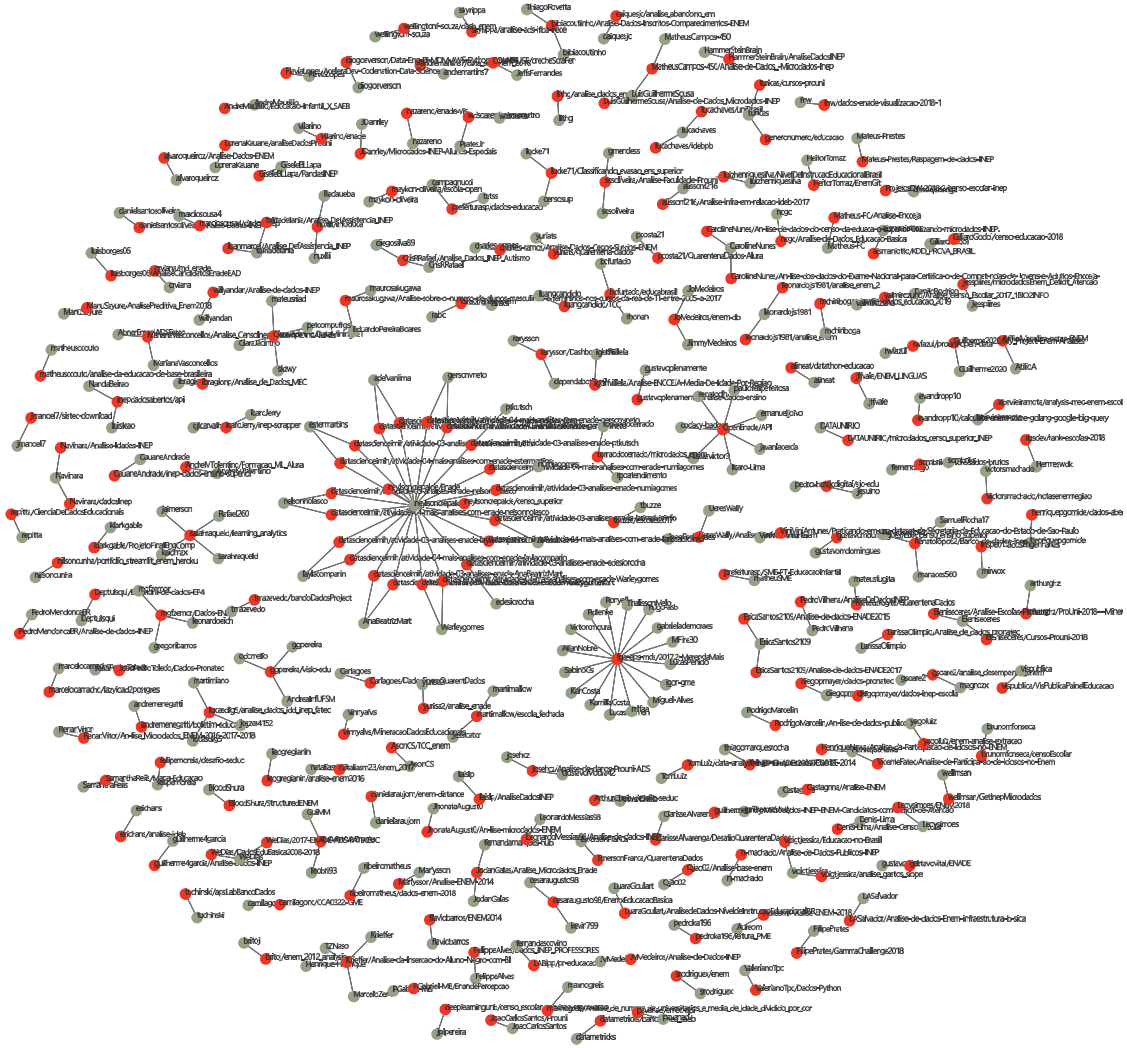
Repositórios

Dos 217 Repositórios 4 deles ao terem a lista de contribuidores acessada através da API retornam vazio.

Repositórios

Contribuidores

Experiências



Quantidade de colaborações

1 2 3 4 7 16

Tipo de perfil dono do repositório

Organização Usuário

Tipo de usuário

Bot User

Tecnologia

Todos

Objetivo do projeto

Todos

Trabalho realizado

Todos

Forks no repositório

Todos

Stars no repositório

Todos