CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Uma Avaliação de Dados Abertos Educacionais Brasileiros: Qualidade, Privacidade e *Learning Analytics*

Ciro X. Maretto[1, 2], Monalessa P. Barcellos[1]

[1]Núcleo de Estudos em Modelagem Conceitual e Ontologias (NEMO)
Universidade Federal do Espírito Santo (UFES) - Vitória, ES – Brasil

[2]Instituto Federal do Espírito Santo (Ifes) - Cariacica, ES – Brasil

ciro@ifes.edu.br, monalessa@inf.ufes.br

Abstract. *The growing demand for educational data, whether Open Data or Learning Analytics, highlights the challenge of balancing quantity, quality, and privacy. This article reports the experience of an evaluation on the quality of educational open data in Brazil. As one of the results, a supporting tool was developed and the main shortcomings and improvement opportunities were identified. The study concludes that the current state of data needs to advance in basic dimensions, whether due to lack of detail in legal guidelines or validations of the Brazilian Open Data Portal.*

Resumo. *A crescente demanda por dados educacionais, seja por meio de Dados Abertos ou Learning Analytics, ressalta o desafio de equilibrar quantidade, qualidade e privacidade. Este artigo relata a experiência de uma avaliação da qualidade de dados abertos educacionais no Brasil. Como um dos resultados tem-se a disponibilização para comunidade de uma ferramenta de apoio, assim como o apontamento das principais deficiências e oportunidades de melhoria observadas. O estudo conclui que o estado atual dos dados precisa evoluir em dimensões básicas, seja por falta de detalhes das orientações legais ou de validações do Portal Brasileiro de Dados Abertos.*

1. Introdução

Os dados abertos representam um dos pontos centrais para transparência governamental, permitindo uma visão mais clara e detalhada das ações do governo [Çaldağ e Gökalp 2022]. No Brasil, a relevância desses dados tem se destacado desde a Lei Complementar 131/2009, também conhecida como "Lei da Transparência", e foi ampliada a partir do momento em que o país se tornou membro da *Open Government Partnership* (OGP) em 2011, mesmo ano da promulgação da "Lei de Acesso à Informação" nº 12.527, de 18 de novembro de 2011 (LAI). A Lei da Transparência é focada na abertura da gestão fiscal e execução financeira. Já a associação à OGP representa o compromisso do Brasil com a transparência, a participação cidadã e a inovação por meio de reúso e combinação de dados. A LAI, por sua vez, é um marco legal que consagra o direito dos cidadãos a qualquer informação, exceto em casos em que o sigilo seja justificável.

No âmbito da área de Educação, mais especificamente no contexto das instituições de ensino federais, ainda se aplica o Decreto nº 8.777 de 11 de maio de 2016. Esse decreto exige a publicação de Planos de Dados Abertos a cada dois anos e

apoia a ideia de que os dados produzidos ou custodiados pelo governo são um ativo público e, como tal, devem estar disponíveis a todos, a menos que estejam expressamente protegidos por lei (LAI). Porém, em 2018 foi publicada a Lei nº 13.709/2018, também conhecida como Lei Geral de Proteção de Dados (LGPD), que entrou plenamente em vigor em 2021. Essa lei regula todas as atividades que envolvem o uso de dados pessoais no Brasil. A LGPD define as regras sobre a coleta, o armazenamento, o tratamento e o compartilhamento de dados pessoais, impondo um padrão mais elevado de proteção e penalidades para o não cumprimento.

Nesses mais de 10 anos de direcionamentos legais para a abertura de dados, é possível observar na Educação o crescimento da pesquisa em *Learning Analytics* (LA) [Sghir et al. 2022], que se beneficia com a crescente quantidade de dados coletados [Mougiakou et al. 2023]. Porém, isso pode evidenciar um cenário no qual há grande volume de dados, mas poucos são de fato usados [Drigas e Leliopoulos 2014]. Além disso, percebe-se que os dados vêm sendo coletados e tratados sem atenção às questões de proteção de dados e privacidade [Prinsloo et al. 2022].

Considerando a necessidade de dados de qualidade, que permitam o entendimento dos dados abertos da Educação no Brasil e, assim, apoiem a definição e monitoramento de políticas públicas, este artigo visa realizar uma análise específica de dados abertos brasileiros, com foco particular nas instituições federais de ensino. O objetivo é avaliar a qualidade atual dos dados publicados por essas instituições após mais de uma década de orientações legais. Para tal, a avaliação é baseada em dimensões de qualidade já estabelecidas na literatura em geral, mas aqui com foco voltado para a Educação, diferentemente de outros trabalhos, tais como [Silva et al. 2020; Silva e Pinheiro 2018]).

No estudo aqui reportado, foi adotada uma abordagem de pesquisa quantitativa, com propósito exploratório de proporcionar uma visão abrangente acerca da qualidade dos dados abertos educacionais provenientes das instituições de ensino federais. A principal contribuição deste trabalho é fornecer subsídios iniciais que possam auxiliar na melhoria da qualidade desses dados. Além disso, foi desenvolvida a Ferramenta de Apoio à Avaliação para Dados Abertos Educacionais (FADAE), que está disponível em um repositório de código público [Maretto e Barcellos 2023]. A ferramenta viabiliza a reprodutibilidade do estudo, o acompanhamento de métricas para dimensões de qualidade, o aprimoramento por parte da comunidade e a obtenção de informações que podem ser usadas como base para propor melhorias visando à qualidade de dados.

O restante deste artigo está estruturado em 6 seções. Na Seção 2 é apresentado o referencial teórico, que aborda aspectos relacionados à qualidade em dados abertos, LA, qualidade de dados em LA e governança de dados. A Seção 3 introduz a ferramenta utilizada. Os resultados obtidos na execução do estudo são abordados na Seção 4. Na seção 5 é feita uma discussão sobre os resultados. A Seção 6 aborda trabalhos correlatos. Por fim, a Seção 7 apresenta as considerações finais do artigo.

2. Qualidade de Dados Abertos Educacionais

Os dados abertos educacionais representam uma considerável parcela dos dados governamentais abertos brasileiros e são fonte para algumas pesquisas no Brasil [Ferreira et al. 2021]. No contexto de dados educacionais, nem todo dado possível é produzido, já que a sua produção depende das práticas pedagógicas adotadas e do

contexto em que são aplicadas. Além disso, nem todo dado produzido é coletado, sobretudo de forma informatizada, pois a coleta depende de suporte tecnológico, que nem sempre está disponível. Dos dados que são coletados, uma parte é usada para as atividades internas das instituições (e.g., matrícula, emissão de histórico, etc.) e apenas uma porção dos dados coletados é de fato publicada de forma aberta. Esses dados que são publicados abertamente precisam atender requisitos de privacidade mais rigorosos e que exigem maior governança para permitir o correto atendimento ao seu propósito.

O Portal Brasileiro de Dados Abertos (*dados.gov.br*) é uma página web que centralizada dados governamentais abertos do Brasil. Esse Portal é uma implementação do CKAN (*ckan.org*), que é um sistema de gerenciamento de dados de código aberto usado por vários países [Silva et al. 2020]. Em junho de 2023, quando foi realizado o último levantamento para o estudo reportado neste artigo, constavam 232 organizações e 12.398 conjuntos de dados no Portal. Em outubro de 2020, eram 9.850 conjuntos de dados [Macedo e Lemos 2021].

Com cada vez mais dados sendo publicados, torna-se crucial avaliar a sua qualidade, pois a qualidade dos dados afeta diretamente seu potencial reúso e as decisões que são tomadas com base neles. Apesar de uma certa subjetividade ser intrínseca ao conceito de qualidade, em geral, ela pode ser considerada resultado da avaliação de um conjunto de dimensões [Silva e Pinheiro 2018]. Nesse sentido, há várias propostas para auxiliar na avaliação de dados abertos governamentais (e.g., [Çaldağ e Gökalp 2022; Laranjeiro et al. 2015; Silva et al. 2020; Silva e Pinheiro 2018]). Embora não sejam específicas para área de Educação, elas trazem princípios comuns no que diz respeito à avaliação de qualidade por meio de múltiplas dimensões, sendo recorrentes: **licenciamento** (permissão legal para uso), **completude** (se todos os dados permitidos estão presentes), **primariedade** (dados direto da fonte, sem transformações ou agregações), **acessibilidade** (disponibilidade do dado), **processabilidade** (facilidade de processamento por computador) e **atualidade** (atualizado o mais breve possível) e **privacidade e segurança** (atendimento a legislação e controle de acesso). Apesar das similaridades, não há uma definição comum de qualidade para dados educacionais até o momento [Mougiakou et al. 2023].

Os dados educacionais podem ser analisados sob a perspectiva de LA. Segundo [Lang et al. 2017], LA é uma área que abrange a coleta, mensuração, interpretação e divulgação de informações sobre os alunos e suas circunstâncias, com o objetivo de compreender e aperfeiçoar o processo de aprendizado, assim como o ambiente em que ele se realiza. As principais aplicações de LA, em ordem de quantidade de ocorrência são [Sghir et al. 2022]: mensurar a performance dos alunos, prever risco de falha ou evasão, medir o engajamento e satisfação, prever testes de admissão e fluxos de matrícula. Além das definições gerais de qualidade de dados, em [Scheffel et al. 2015] é apresentado um *framework* para avaliar ferramentas de LA, no qual são consideradas preocupações relacionadas a transparência, padrões de dados, dono do dado e privacidade. É possível observar que existe uma interseção entre dimensões consideradas para avaliação de dados abertos e ferramentas de LA.

Governança de dados desempenha um papel fundamental na garantia da qualidade dos dados ao longo do tempo, pois envolve a implementação de políticas, definição de responsabilidades e práticas que promovem o uso eficaz dos dados. Como observado em [Filgueiras e Lui 2023], governança de dados tem tido destacada

relevância recentemente, muito motivada pela LGPD. Entretanto, aplicar a governança envolvendo dados abertos Educacionais e LA é desafiador. É preciso equilibrar a necessidade de proteger a privacidade dos envolvidos, em especial os alunos, com a necessidade de coletar e analisar dados para melhorar a aprendizagem e seus ambientes.

3. Ferramenta de Apoio a Avaliação para Dados Abertos Educacionais (FADAE)

Para realizar a análise de dados reportada neste artigo, foi criada a ferramenta FADAE, desenvolvida em linguagem Python. Ela implementa um conjunto de funcionalidades para realização de Extração, Transformação e Carregamento (ETL - *Extract, Transform, Load*) de dados educacionais abertos. A Figura 1 ilustra as atividades realizadas com o apoio de FADAE, as quais são brevemente descritas a seguir.

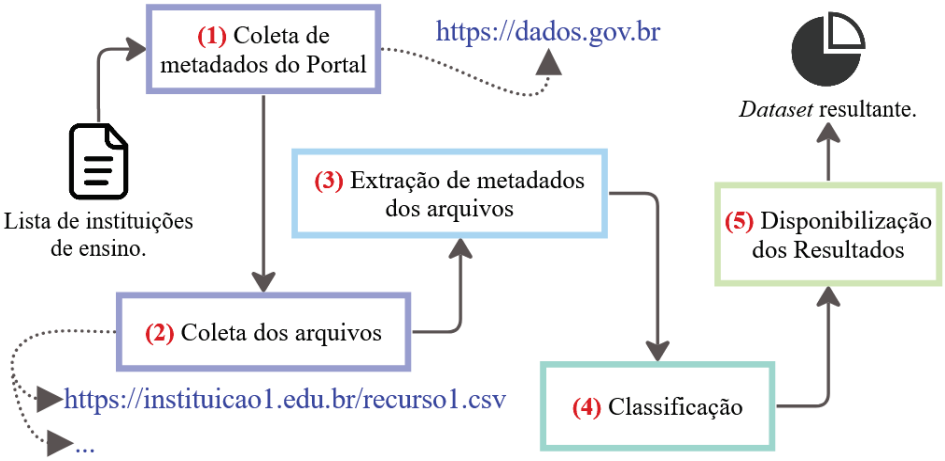

Figura 1. Atividades apoiadas por FADAE com dados ilustrativos.

(A) Coleta de metadados do Portal: inicialmente, FADAE deve ser alimentada com uma lista das instituições de ensino a serem consideradas. Isso é necessário, pois o Portal aborda dados abertos de diferentes tipos de organização governamental. Usando uma API, a ferramenta extrai metadados para cada conjunto de dados das instituições de ensino informadas. Cada conjunto de dados pode ter diferentes recursos. Por sua vez, cada recurso possui um conjunto de metadados, por exemplo, URL para o download do arquivo de dados, a quantidade de downloads, o formato e se está disponível ou não. Alguns dos metadados disponíveis na API são expostos no Portal.

(B) Coleta dos arquivos: nesta atividade é realizado o download dos arquivos das URLs coletadas do passo anterior. Além disso, atualiza o metadado de disponibilidade caso o link apresente erro.

(C) Extração de metadados dos arquivos: nesta atividade FADAE valida se o formato reportado no Portal corresponde ao formato do arquivo baixado e tenta extrair outros metadados a partir do arquivo. Esses novos metadados são os cabeçalhos das tabelas de dados armazenados em formato tabular (i.e., presentes em arquivos .csv, .ods, xls ou .xlsx), o nome do arquivo e sua extensão. Então, é feita a agregação dos metadados coletados do Portal (A) com os coletados a partir dos arquivos (B). Alguns problemas podem ocorrer na extração dos dados. Por exemplo, arquivos .csv podem ser criados com diferentes *charsets* e dialetos (delimitadores), o que pode implicar em problemas de *parse* e codificação. Esses problemas são contornados em FADAE com as

bibliotecas e funções csv.Sniffer (*docs.python.org*) e chardet.detect (*github.com/chardet*), que tentam identificar os padrões apropriados e geralmente produzem excelentes resultados. Porém, ainda pode apresentar falhas, principalmente em recursos com poucos dados.

(D) Classificação: esta atividade busca identificar os dicionários de dados (i.e., recursos que descrevem o que significa cada atributo, por exemplo, a descrição de cada cabeçalho em um recurso disponibilizado em arquivo .csv), os potenciais riscos à proteção de dados e a presença de atributos usados em LA. Para isso, são utilizadas algumas técnicas de *Natural Language Processing* (NLP) por meio da biblioteca SpaCy (spacy.io): tokenização, lematização, remoção de *stop words* e *Levenshtein distance*. A identificação dos potenciais riscos à proteção de dados é realizada por meio de atributos como CPF, RG, documento, matrícula, identidade, nome, aluno, sexo, data nascimento, e-mail, entre outros. Além da questão de privacidade, nesta atividade são buscados atributos comumente usados nas pesquisas sobre LA. Para isso, FADAE baseia-se na classe de atributos de [Issah et al. 2023]. Esse trabalho aponta os principais atributos usados nos trabalhos de LA da literatura (entre 2016 e 2022) voltados para a previsão de desempenho dos alunos. Partindo dos atributos mais utilizados para os menos utilizados, tem-se: Desempenho acadêmico (média de notas, nível de escolaridade, número de matérias feitas por semestre); Demográfico (sexo, nacionalidade, local de nascimento, idade); Comportamental (participações em sala, visita aos recursos, satisfação escolar, frequência escolar); Psicológico (personalidade, motivação, estratégias de aprendizagem); Características familiares (escolaridade de mãe e pai, renda familiar, endereço dos pais); Ambiente escolar (tamanho da escola). Outra classificação dos conjuntos de dados é feita por similaridade semântica com base nos metadados do conjunto de dados ("Nome" e "Descrição") e uma adaptação das categorias definidas em [Sghir et al. 2022]: Aluno, Professor e Institucional (cursos, financeiro, administrativo).

(E) Disponibilização dos Resultados: é a atividade na qual FADAE exporta para um arquivo .csv a compilação de todos os dados, incluindo metadados acrescentados e classificações. Isso habilita os dados serem processados por outras ferramentas especializadas em visualização de dados, permitindo a criação de gráficos e relatórios personalizados. O *dataset* resultante da execução dos passos anteriores representa o cenário de dados abertos educacionais brasileiros em um dado instante e contempla os seguintes atributos: instituição, sigla, nome do conjunto de dados (*dataset*), descrição do *dataset*, nome do recurso, descrição do recurso, formato, data criação, data de modificação do *dataset* e do recurso, quantidade de downloads do *dataset* e do recurso, flag de disponibilidade, caminho do arquivo, URL, formato, flag erro no parse (.csv), charset (.csv), delimitador (.csv), cabeçalhos (dados tabulares), atributos de risco à proteção de dados, atributos para LA, categoria (classificação semântica), flag para dado atualizado, periodicidade, grupo (metadado do Portal) e licença. Além do *dataset*, algumas estatísticas podem ser obtidas diretamente por meio da FADAE. Parte delas são apresentadas a seguir.

4. Execução do Estudo e Resultados Obtidos

O estudo consistiu em utilizar FADAE para analisar a qualidade de dados de instituições de ensino federais em relação às seguintes dimensões de qualidade (descritas na Seção 2): licenciamento, completude, acessibilidade, processabilidade,

atualidade, utilidade e legalidade. Inicialmente, foram selecionadas no site do MEC (*mec.gov.br*) 113 instituições de ensino do país (em sua maior parte institutos e universidades federais). Esse conjunto de instituições foi, então, fornecido como entrada para a atividade (A) "Coleta de metadados do Portal", o que resultou em um total de 1.957 conjuntos de dados e 10.187 recursos. Das 113 instituições selecionadas, 32 não estão presentes no Portal e três estão presentes, mas não possuem conjuntos de dados publicados. Na atividade (B) "Coleta dos arquivos" foram baixados 21,9 Gb de dados e 13.214 arquivos. Na atividade (C) "Extração de metadados dos arquivos" foram processados 9.878 arquivos tabulares para obtenção de novos metadados. Na atividade (D) "Classificação" esses arquivos foram classificados. Finalmente, na atividade (E) "Disponibilização dos Resultados" os resultados das atividades anteriores foram compilados em um único *dataset*. Os dados foram, então, analisados considerando-se as dimensões mencionadas. Um resumo dos resultados é apresentado a seguir.

Licenciamento: foram identificadas anomalias em relação à licença dos conjuntos de dados. No estudo, todas as licenças encontradas são abertas e em geral muito similares. A maior parte, 69,66%, é CC-BY[1]. 18,03% dos recursos não possuem especificação da licença ou ela é genérica Other (Open). Dois recursos (0,01%) possuem a licença UK-OGL, que é específica do Reino Unido.

Completude: foram analisados os metadados usados. No Portal existem recursos sem atributos essenciais, como título, periodicidade ou licença. Em 90,64% dos recursos foi constatada falta de algum dos metadados coletados. 41,49% dos dados possuíam algum recurso com a função de dicionários de dados. Havia grande variação na quantidade de conjuntos de dados publicados, com média de 25 e máximo de 151.

Acessibilidade: a verificação da disponibilidade do recurso pode ser feita por meio dos metadados do Portal e da tentativa de download. 92,49% dos arquivos estão disponíveis. Alguns não estavam acessíveis, ocorrência do código de resposta HTML 404 ou armazenado no Google Drive com exigência de solicitação de permissão para acesso, ou, ainda, com redirecionamento que não levava ao recurso, mas a uma página web qualquer (e.g., a página inicial da instituição) como resultado.

Processabilidade: esta dimensão foi avaliada na atividade (C) "Extração de metadados dos arquivos", ao se tentar capturar os cabeçalhos dos dados. A maior parte dos arquivos está no formato .csv (75,2%), sendo os principais delimitadores usados "," (54,6%) e ";" (39,7%). Os arquivos .csv apresentaram problemas quanto à forma, tais como, linhas em branco no início do arquivo, cabeçalhos de algumas colunas em branco e principalmente linhas de título antes de cabeçalhos dos dados tabulares. Considerando os problemas de acesso (atividade B) e a processabilidade dos dados (atividade C), 27,03% dos dados estão indisponíveis ou não são processáveis.

Atualidade: essa dimensão pode ser avaliada observando-se a última atualização do conjunto de dados. 36,31% dos conjuntos de dados foram atualizados há mais de um ano, enquanto 64,41% foram atualizados nos últimos seis meses.

Utilidade: o número de downloads e a presença de dados utilizados por LA são indicadores para essa dimensão. Apenas 15,13% dos recursos possuem mais de 10

[1] Licença que permite a redistribuição e modificação do trabalho original, inclusive para fins comerciais, desde que o autor original seja creditado.

downloads e 36,24% possuem 0 download. Porém, é preciso considerar que os recursos não são armazenados no Portal, o que torna a métrica quantidade de downloads pouco precisa, já que downloads direto na fonte não são computados no Portal. Em relação aos dados utilizados em LA, aproximadamente 23,17% dos dados foram classificados como referente a alunos. Atributos usados em LA foram observados em apenas 2,5% dos recursos e, na grande maioria, com poucos atributos ou sem referência ao desempenho do aluno.

Legalidade: a presença de dados pessoais típicos pode ser um indicador de atendimento à legislação e ocorrência de riscos à LGPD. 3,76% dos recursos apresentaram dados pessoais, estando distribuídos em 46,75% das instituições. Como exemplo de ocorrência, tem-se um dado recurso contendo matrículas e deficiências (relativas à saúde). Nesse caso, a matrícula foi usada como uma forma de ocultar o nome do aluno. Porém, em uma breve pesquisa por qualquer uma das matrículas usando o buscador do Google, foi possível encontrar o nome completo do aluno no próprio site da instituição. Identificando, assim, a pessoa e qual tipo de deficiência ela possui. Essa é uma informação relacionada à saúde e deveria estar protegida.

5. Discussão

Os resultados obtidos no estudo a partir da execução das atividades de FADAE oferecem uma visão geral da qualidade dos dados abertos educacionais e destacam as principais deficiências nas dimensões observadas. A seguir são discutidas algumas questões considerando-se os resultados obtidos. A discussão não é exaustiva e visa destacar alguns pontos para contribuir para uma reflexão geral sobre a qualidade dos dados e oportunidades de melhoria.

A predominância da publicação dos dados em arquivos .csv não garante que serão processáveis por máquina de forma satisfatória. A falta de padrão de forma e conteúdo dos arquivos pode impossibilitar o uso dos arquivos pelos diferentes sistemas das instituições. A solução recomendada para essa questão é a utilização de formatos mais específicos. O principal deles é uma especificação da *World Wide Web Consortium* (W3C), o RDF (*Resource Description Framework*). Esse formato é largamente usado para dados abertos ligados. Considerando que as instituições de ensino possuem dificuldades para publicação em .csv, que é um formato relativamente fácil e conhecido, acredita-se que exista ainda um caminho a ser percorrido para implantação do uso de arquivos RDF.

Como observado nos dados coletados, o problema relacionado à proteção de dados pessoais pode não estar diretamente no recurso publicado, mas se manifestar quando há a combinação dele com outros dados públicos. Além disso, conforme destacado em [De Queiroz e Motta 2015], fica evidente que técnicas simples de anonimização podem não ser adequadas para garantir um alto nível de proteção contra a reidentificação. O cuidado com a privacidade é uma das principais preocupações em LA, sendo uma das questões básicas abordadas em *frameworks* da área, como o SHEILA [Tsai et al. 2018]. A publicação de dados protegidos é um indicativo de baixa governança de dados na instituição.

Em relação a LA, nota-se que os dados publicados diretamente pelas instituições de ensino avaliadas não trazem relevância expressiva. Como apontado na Seção 2, quanto mais próximo do dado aberto, mais restrito é o conjunto de dados, o que impacta

nas suas possibilidades de uso. Porém, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP (*www.gov.br/inep*) possui dados abertos relevantes para trabalhos de LA [Ferreira et al. 2021]. O INEP e o MEC utilizam informações fornecidas pelas próprias instituições de ensino e avaliações de seus alunos para consolidar estatísticas importantes sobre a educação no Brasil. Normalmente, são publicados relatórios ou *datasets* anuais, muitos desses no Portal Brasileiro de Dados Abertos. Diferentemente do INEP, não faz parte da atividade-fim das instituições de ensino o tratamento e publicações de dados. Porém, cada vez mais dados são requisitados e a sua correta manipulação é essencial para garantir a qualidade. A área de LA é a maior beneficiada com o amadurecimento da governança de dados dentro das instituições de ensino e, como consequência, a qualidade dos dados abertos educacionais também seria melhorada.

É importante destacar que apesar de abordar questões relevantes quanto à qualidade de dados educacionais, a avaliação realizada como um observador externo das instituições de ensino limita a percepção real sobre a condição da fonte dos dados. Nesse sentido, as dimensões completude e primariedade possuem restrições de alcance nesse tipo de avaliação. Essa limitação pode ser vencida caso a própria instituição faça sua avaliação com o propósito de melhoria de qualidade e conformidade legal. Para isso, a implantação da governança de dados precisa ser considerada pelas instituições de ensino como uma questão estratégica.

6. Trabalhos Correlatos

Apesar da vasta quantidade de trabalhos relacionados a dados abertos na literatura, poucos são os estudos na área educacional [Penteado et al. 2017]. Em [Silva et al. 2020], é feita uma análise de qualidade do Portal Brasileiro de Dados Abertos considerando os oito princípios de *Open Government Data* (*opengovdata.org*) e 17 dimensões específicas de qualidade de informação. Já em [Silva e Pinheiro 2018] é proposta DGABr, uma métrica para avaliar o potencial de reúso dos dados governamentais abertos disponibilizados no Brasil. Essa avaliação possui 28 dimensões considerando as perspectivas: dados abertos, legal, técnica, gerencial e reúso. A iniciativa *Open Data Barometer* (*opendatabarometer.org*) traz índices para implementação de boas práticas para dados abertos e separa os resultados por categoria de dados, sendo uma delas educação. Porém, essa iniciativa apenas publicou dados de 2013 a 2017. Nos trabalhos supracitados a avaliação dos dados é feita independente de área, o que limita uma análise mais profunda no cenário específico da educação. Outro ponto comum a esses trabalhos é a concentração no que é apresentado no Portal, sem que haja uma avaliação no nível dos dados.

Em [Penteado et al. 2017], os dados abertos educacionais brasileiros são avaliados quanto à recomendação da *Data on the Web Best Practices* (DWBP), criada pela W3C. A avaliação é feita em apenas quatro conjuntos de dados e considera como uma das dimensões a "qualidade". Embora não deixe claro o que exatamente seria qualidade, sugere como ponto de melhoria para essa dimensão a utilização da extensão *qa* para o CKAN. Até a data da elaboração deste artigo essa extensão não foi aplicada e avalia apenas o nível em relação ao formato do arquivo publicado com base na classificação *5 Star Linked Data* da W3C. Além disso, nesse estudo, o tema de proteção de dados e privacidade não é abordado.

O estudo abordado neste artigo contribui para a área e se diferencia dos citados por avaliar: uma grande quantidade de conjuntos de dados ao invés de uma pequena amostra pré-selecionada; desenvolver, utilizar e disponibilizar ferramenta de apoio à análise de dados abertos educacionais brasileiros; considerar o conteúdo dos arquivos além dos metadados disponíveis no Portal; apresentar uma avaliação atualizada; e abordar proteção de dados pessoais e *learning analytics*.

7. Conclusão

Este artigo apresentou uma análise de dados educacionais conduzida com auxílio de FADAE, que requer menor intervenção manual e não se limita aos metadados disponíveis no Portal Brasileiro de Dados Abertos. Sob essa perspectiva abrangente, as evidências de qualidade nos dados abertos educacionais representam, também, um indicativo da governança de dados e do ambiente de aplicação de LA das instituições. Os resultados indicam que ainda há necessidades de melhorias, mesmo em dimensões básicas para qualidade de dados.

Com cada vez mais dados sendo publicados, é essencial o apoio de ferramentas e orientações mais robustas. Da mesma forma que as legislações foram as impulsionadoras do atual cenário, outras orientações podem ser usadas para nortear as políticas de melhoria de qualidade das informações disponibilizadas no Portal. Seja por meio de notas técnicas complementares, configuração de validadores no CKAN, aplicação de ontologias de referência e até mesmo direcionamentos para quais conjuntos de dados podem ou devem ser publicados.

Este trabalho, por se tratar de uma avaliação com base em uma ferramenta aberta, permite reprodutibilidade e adaptação a fim de melhorar e acompanhar as características dos dados abertos educacionais. Dessa forma, pode-se direcionar os esforços nos pontos mais deficitários e identificar pontos divergentes para aprimoramento das políticas públicas. Nas limitações deste trabalho se destacam o processamento apenas de dados tabulares, a intervenção manual do arquivo de entrada e o método simplificado para classificação semântica. Como trabalho futuro, além de aprofundamento das investigações conduzidas e do uso de novas formas de classificação dos dados, uma variação de FADAE pode ser criada para ser aplicada internamente nas organizações, antes mesmo que os dados sejam publicados de forma pública.

Agradecimento

Esta pesquisa foi apoiada pela FAPES - Fundação de Amparo à Pesquisa do Estado do Espírito Santo (Processo 2023-5L1FC e T.O. 1022/2022).

Referências

Çaldağ, M. T., Gökalp, E. (2022). The maturity of open government data maturity: a multivocal literature review. *Aslib Journal of Information Management*, v. 74, n. 6, p. 1007-1030.

De Queiroz, M. J., Motta, G. H. (2015). Privacidade e Transparência no Setor Público: Um Estudo de Caso da Publicação de Microdados do INEP. In *XV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*. SBC, p. 362-365

Drigas, A. S., Leliopoulos, P. (2014). The Use of Big Data in Education. *International Journal of Computer Science Issues*, v. 11, n. 5, p. 58-63.

Ferreira, L. A., Rodrigues, R. L., De Souza, R. N. (2021). Dados abertos educacionais brasileiros: Um mapeamento sistemático da literatura. In *XXXII Simpósio Brasileiro de Informática na Educação - SBIE*. SBC, p. 1186-1195.

Filgueiras, F., Lui, L. (2023). Designing data governance in Brazil: an institutional analysis. *Policy Design and Practice*, v. 6, n. 1, p. 41-56.

Issah, I., Appiah, O., Appiahene, P., Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision Analytics Journal*, v. 7, p. 100204.

Lang, C., Wise, A., Siemens, G., Gasevic, D. (2017). *Handbook of Learning Analytics*. New York: SOLAR, Society for Learning Analytics and Research.

Laranjeiro, N., Soydemir, S. N., Bernardino, J. (2015). A Survey on Data Quality: Classifying Poor Data. In *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing - PRDC,* p. 179-188.

Macedo, D. F., Lemos, D. L. D. S. (2021). Dados abertos governamentais: iniciativas e desafios na abertura de dados no Brasil e outras esferas internacionais. *AtoZ: novas práticas em informação e conhecimento*, v. 10, n. 2, p. 14.

Maretto C. X., Barcellos M. P. (2023) Ferramenta de Apoio à Avaliação para Dados Abertos Educacionais - FADAE. http://dx.doi.org/10.6084/m9.figshare.23650068.

Mougiakou, S., Vinatsella, D., Sampson, D., et al. (2023). *Educational Data Analytics for Teachers and School Leaders*. Cham: Springer International Publishing.

Penteado, B., Bittencourt, I. I., Isotani, S. (2017). Dados abertos educacionais no Brasil e sua preparação para os dados abertos na web. In *XXVIII Simpósio Brasileiro de Informática na Educação - SBIE*. SBC, p. 526-535.

Prinsloo, P., Slade, S., Khalil, M. (2022). The answer is (not only) technological: Considering student data privacy in learning analytics. *British Journal of Educational Technology*, v. 53, n. 4, p. 876-893.

Scheffel, M., Drachsler, H., Specht, M. (2015). Developing an evaluation framework of quality indicators for learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, p. 16-20.

Sghir, N., Adadi, A., Lahmer, M. (2022). Recent advances in Predictive Learning Analytics: A decade systematic review (2012-2022). *Education and Information Technologies*, p. 1-35.

Silva, A. de A. P., Monteiro, D. A. A., Reis, A. de O. (2020). Qualidade da informação dos dados governamentais abertos: análise do portal de dados abertos brasileiro. *Revista Gestão em Análise*, v. 9, n. 1, p. 31-47.

Silva, P. N., Pinheiro, M. M. K. (2018). DGABr: Metric for evaluating Brazilian open government data. *Informação & Sociedade: Estudos*, v. 28, n. 3.