Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 26/Mar/2018; 1st round notif.: 03/Oct/2018
Camera ready: 20/Nov/2018; Edition review: 23/Nov/2018; Available online: 23/Nov/2018; Published: 01/Jan/2019

Análise exploratória sobre a abertura de dados educacionais no Brasil: como melhorar o ecossistema de dados na Web?

Exploratory analysis on the opening of education data in Brazil: how to improve the ecosystem of data on the Web?

Bruno Elias Penteado
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP) 13566-590 – São Carlos – SP – Brasil
brunopenteado@usp.br

Ig Ibert Bittencourt
Instituto de Computação – Universidade Federal de Alagoas (UFAL) – 57072-970 – Maceió – AL – Brasil
ig.ibert@ic.ufal.br

Seiji Isotani
Instituto de Ciências Matemáticas e da Computação – Universidade de São Paulo (USP) 13566-590 – São Carlos – SP – Brasil
sisotani@icmc.usp.br

Resumo

Os dados abertos educacionais trazem informações importantes sobre o cenário educacional de um país. Sua publicação traz impactos tanto em transparência quanto no aumento do potencial econômico para a sociedade como um todo por meio da gestão da aprendizagem e da tomada de decisão baseada em evidências. A Web traz muitas funcionalidades que podem estender esse potencial e, para isso, o W3C - organização responsável pelas padronizações na Web - criou a recomendação DWBP (Data on the Web Best Practices), para a produção de dados abertos que tirem proveito da arquitetura da Web. Contudo, até o momento não foram encontradas pesquisas que avaliem os dados de acordo com esse referencial. Assim, neste trabalho aplicamos um estudo de caso múltiplo, com a técnica de casos diversos considerando diferentes categorias de dados educacionais, analisando as práticas da DWBP. Como resultado, notamos poucas práticas plenamente atendidas atualmente e muitas práticas atendidas parcialmente. Isso ocorre, pois as informações disponibilizadas são de fácil acesso para o consumo humano, mas não para o processamento automático por agentes de software, inviabilizando a análise automática e contínua dos dados disponibilizados. Ao final, discutimos a adequação à recomendação, as possibilidades de intervenções tanto para a produção dos dados quanto para os portais que os hospedam e outras características percebidas durante a análise.
Palavras-Chave: *dados abertos governamentais; dados abertos educacionais; DWBP.*

Abstract

Open education data carry important information on the education landscape of a country. Its publishing impacts both transparency and economic potential to society through learning management and evidence-based decision making. The Web carries many features that can extend this potential and, to achieve that goal, W3C – organization responsible for the standards in the Web - developed the DWBP (Data on the Web Best Practices) recommendation, for the production and sharing of open data that realizes the potentials of Web architecture. Thus, in this work, we applied a multiple case study, with diverse cases technique considering diverse categories of education data analyzing the DWBP practices. As result, we noticed few current practices fully met and many practices just partially met. It occurs because the information are presented in an easy way for humans to read but not for automatic processing by software agents, rendering it infeasible to automatic and continuous analysis on the published data. In the end, we discuss the compliance to the recommendation, potential interventions both for data production and for data portals which host them and other characteristics perceived during the analysis.
Keywords: *open government data, open education data; DWBP.*

Cite as: Penteado, B.E., Bittencourt, I. I, Isotani, S. (2019). Exploratory analysis on the opening of education data in Brazil: how to improve the ecosystem of data on the Web? (Análise exploratória sobre a abertura de dados educacionais no Brasil: como melhorar o ecossistema de dados na Web?) Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação - RBIE), 27(1), 175-195. DOI: 10.5753/RBIE.2019.27.01.175

1 Introdução

Na última década, diversos países têm apresentado iniciativas de abertura de seus dados governamentais, no chamado movimento de *governo aberto*. O governo aberto pode ser compreendido como um método de governança que fornece espaço para a abertura, transparência e diálogo contínuo entre o governo e seus cidadãos (Parycek e Sachs, 2015). Para se atingir esses objetivos, os dados abertos são um componente-chave, ao permitir a publicação e o reuso de dados governamentais em formatos processáveis por computador. Ambos os conceitos são combinados nos chamados *dados abertos governamentais* que também trazem conjuntos de políticas que promovem a transparência, a prestação de contas e a criação de valor ao tornar dados governamentais disponíveis a todos os cidadãos (OECD, 2017). Ao liberar os dados em formatos não proprietários e sem licenças restritivas, o governo permite que atores da sociedade se apropriem das informações e gerem análises, produtos e serviços que retornem na forma de benefícios para a própria sociedade (Santos, 2014; Isotani & Bittencourt, 2015).

A liberação de informações de setores públicos na forma de dados abertos é considerada um dos pilares fundamentais do governo aberto. Com isso, este tópico tem adquirido importância crescente, com interesses desde dos criadores de políticas públicas até de profissionais da área, nas esferas governamentais e privada. Os dados abertos governamentais podem ser usados para auxiliar os cidadãos a compreender melhor o que o governo faz e seu desempenho nisso – o que é verdadeiro já que uma quantia considerável desses dados governamentais estão progressivamente se tornando mais acessíveis e podem ser usados em conjunto com informações de outras fontes, mesmo que proprietárias (Ubaldi, 2013).

De acordo com a definição da OpenDefinition.org, "*aberto significa que qualquer pessoa pode acessar, usar, modificar e compartilhar livremente para qualquer finalidade (sujeito, no máximo, a requisitos que preservam a procedência e a abertura)*". Além dessa definição, um grupo de defensores dos dados abertos se reuniu um 2007, em Sebastopol (EUA), e definiu 8 princípios básicos para a garantia de dados abertos governamentais, que têm servido de diretrizes para elaboração de políticas e avaliação do grau de abertura em iniciativas de dados abertos governamentais. Segundo eles, os dados devem ser: completos, primários, divulgados em tempo oportuno, acessíveis, processáveis por computadores, não discriminatórios, não proprietários e de licença de uso livre. Assim, trata-se de um tema multidisciplinar, envolvendo áreas como direito, ciências políticas, economia, comunicação social, ciência da informação, computação, etc., para garantir essas condições em ações como: garantir o respaldo legal de produção e uso dos dados, elaboração e monitoramento de políticas públicas, publicação e cruzamento de dados, divulgação e disseminação dos dados, a utilidade dos dados para fins econômicos, dentre outros aspectos.

No entanto, apesar da crescente adoção às políticas de dados abertos, a oferta de dados na Web ainda tem ocorrido em formatos que impõem limitações quanto a sua reutilização, pois em sua maioria são consumidos apenas por humanos, não permitindo que sejam processadas por agentes de software (Wood et al., 2014). A disponibilidade de dados abertos processáveis por agentes de software pode facilitar seu cruzamento com outros dados, apoiar análises e visualizações, aumentando a capacidade de seus usuários (criadores de políticas públicas, pesquisadores, cidadãos) para a análise e tomada de decisão em problemas complexos. Para isso, os consumidores desses dados esperam que eles sejam precisos, atualizados regularmente disponíveis o tempo todo (W3C, 2017).

A disseminação de dados abertos no Brasil ocorre em conjunto com os desejos da sociedade pela maior transparência dos órgãos públicos e com a instituição de legislações sobre o tema, como a Lei de Acesso à Informação, a qual exige que todos os órgãos governamentais publiquem seus dados para a sociedade como forma de fomentar o governo aberto, a transparência e a participação social, mediados pela tecnologia (Brasil, 2011). Além disso, o Brasil é um dos fundadores, também

em 2011, da Parceria para o Governo Aberto (em inglês, *Open Government Partnership* - OGP), uma iniciativa multilateral que estabeleceu um conjunto de princípios que tem por objetivo assegurar compromissos concretos dos governos para ações de promoção da transparência, combate à corrupção, capacitação dos cidadãos e do uso da tecnologia para fortalecer a governança (OGP, 2011).

Dentro deste contexto, a primeira iniciativa tecnológica de destaque foi o Portal da Transparência (*www.portaltransparencia.gov.br*), lançado em 2004 e que publica dados orçamentários e financeiros do governo federal. Na educação, a prática ainda é incipiente, tendo como os primeiros *datasets* publicados os censos de instituições escolares (tanto da educação básica quanto da superior), e os dados de desempenho escolar no SAEB (Sistema Nacional de Avaliação da Educação Básica)[1]. Em ambos os casos, os primeiros conjuntos de dados foram liberados institucionalmente em 1995 e o movimento de publicação ganhou força com a liberação de diversos conjuntos de dados a partir de meados dos anos 2000 (Penteado, Isotani, 2017).

Dentro deste contexto, este trabalho busca entender ***como os dados abertos educacionais estão prontos para serem usados no ecossistema de dados na Web?*** Para isso fizemos um estudo empírico do estado atual de abertura dos dados, analisando os dados disponíveis no portal de dados do governo federal brasileiro (*dados.gov.br*). Para verificar o grau de disponibilidade e adequação dos dados educacionais publicados utilizou-se aspectos cobertos por uma das recentes recomendações da W3C, a *Data on the Web Best Practices* (W3C, 2017), como licenciamento, qualidade, versionamento, formatos, enriquecimento dos dados, além de outras informações relevantes - componentes básicos para o ecossistema de dados na Web. Este artigo estende Penteado, Bittencourt e Isotani (2017) ao trazer informações complementares e atualizadas sobre a disponibilização dos dados educacionais, como novas funcionalidades no portal *dados.gov.br*, além de uma discussão mais aprofundada sobre as implicações dos resultados.

Na seção 2 é contextualizado o cenário de dados abertos educacionais no Brasil e outros países do mundo. Na seção 3 detalhamos a recomendação DWBP para publicação e consumo de dados abertos. Trabalhos relacionados ao levantamento de dados abertos educacionais são discutidos na seção 4. A seção 5 traz a metodologia utilizada nesta pesquisa. A seção 6 traz os resultados, que são discutidos na seção 7. A partir da discussão, são feitas sugestões para ambientes de catálogo de dados e para produção de dados abertos educacionais, que podem auxiliar na construção de ambientes mais adaptados a este cenário de dados abertos na Web.

2 Dados abertos educacionais

Os dados abertos educacionais são relativamente uma nova área de interesse, com literatura ainda incipiente. Guy (2016) argumenta que o termo "dados abertos educacionais" ainda é vagamente definido, podendo significar tanto: i) dados abertos disponíveis que podem ser usados para propósitos educacionais e ii) dados abertos divulgados por instituições educacionais. No primeiro caso, os dados abertos educacionais podem ser considerados como um subconjunto dos *recursos educacionais abertos* (REA), em que os conjuntos de dados (*datasets*) são disponibilizados para uso no ensino – esses dados podem não ter sido projetados para uso educacional, mas reutilizado livremente com novo propósito. No segundo caso, o interesse está principalmente na divulgação de dados de instituições acadêmicas sobre seu desempenho e de seus estudantes, tais como: dados referenciais, como a localização das instituições; dados internos como registros de empregados, orçamentos, identidade de alunos; dados de currículo, cursos e objetivos de aprendizagem; dados

[1] Ambos os *datasets* originais se encontram em: http://portal.inep.gov.br/web/guest/microdados

gerados pela interação dos usuários; dados padronizados para comparações e acompanhamento de políticas públicas e de transparência.

Santos (2014) classifica os dados educacionais em 3 níveis: macro, meso e micro. No *macro*, são analisados os dados de políticas educacionais nacionais ou regionais, como censos escolares, indicadores de desempenho, gastos com programas educacionais. No nível *meso*, são usados dados de gestão em nível escolar, como indicadores de desempenho das escolas, dados administrativos (acadêmicos ou de gestão). No *micro*, são dados coletados por aluno, como dados pessoais, histórico acadêmico, por vezes conflitando com dados privados.

Os dados educacionais podem ser usados para diferentes finalidades, como: planejamento de metas e objetivos a serem alcançados pelos gestores de uma região; avaliar a efetividade de medidas adotadas no contexto educacional; desenvolvimento de pesquisas; produtos de empresas que atuam no mercado educacional, buscando trazer tanto benefício econômico quanto de transparência, e servindo de base para a avaliação e proposição de melhorias no sistema educacional (Bandeira et al. 2015). Siqueira et al. (2017) argumentam sobre a importância dos dados abertos no desenvolvimento de sistemas de informação que promovam o reuso e a transparência, apontando para as vantagens dos dados abertos conectados, como a expressividade semântica e a possibilidade de processamento automático.

Guy (2016) ainda argumenta que os *datasets* educacionais são de interesse para uma variedade de atores, como educadores, estudantes, instituições, governos, pais e o público em geral – seja para a melhoria do sistema educacional ou para a análise e mineração de dados para influenciar as políticas públicas ou explorar a monetização de *datasets*. Os dados educacionais apresentam grande potencial para muitos e sua exploração é inevitável e necessária. Pelo lado econômico, estima-se que a liberação de dados abertos educacionais pode gerar cerca de 1 trilhão de dólares na economia global, a partir do mapeamento de necessidades específicas de aprendizado em função do histórico de desempenho escolar de cada aluno (McKinsey, 2013).

No Brasil, Penteado & Isotani (2017) realizaram um levantamento da evolução das bases de dados governamentais sobre educação no Brasil, coletadas a partir do portal oficial de publicação de dados abertos federais (*dados.gov.br*). Para tanto, foram filtrados os *datasets* produzidos pelo Ministério da Educação (MEC) e Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) – órgãos responsáveis pela regulação e acompanhamento da educação brasileira em todos os seus níveis. A Figura 1 apresenta a contagem de *datasets* únicos produzidos a cada ano. As primeiras publicações oficiais encontradas datam de 1995, ano em que os conjuntos de dados do censo escolar da educação básica, o censo escolar do ensino superior e os microdados[2] de respostas dos alunos nos testes e questionários contextuais do SAEB, e que desde então têm sido regularmente publicados. Outros *datasets* regulares têm sido publicados desde então: microdados do Exame Nacional do Ensino Médio (ENEM), Programa Nacional de Alimentação Escolar (PNAE), Programa 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, microdados do Exame Nacional de Desempenho de Estudantes (ENADE) e matrículas do ensino superior.

[2] Microdados se referem ao menor nível de desagregação de dados recolhidos por pesquisas, avaliações e exames realizados.

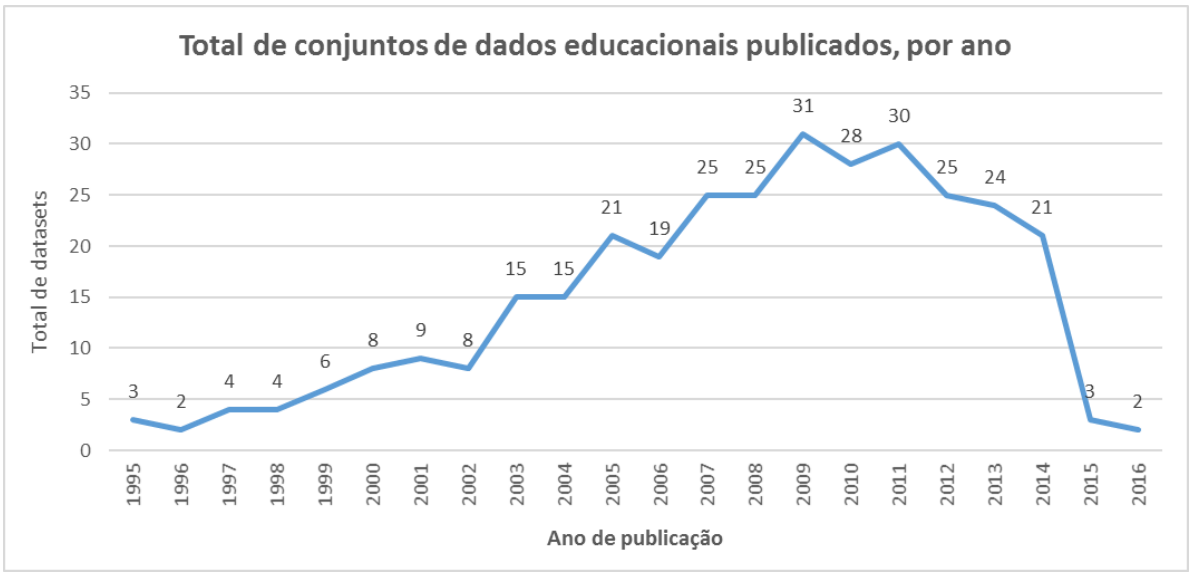

Figura 1. Total de conjuntos de dados publicados a cada ano, desde 1995, pelos órgãos selecionados neste trabalho (adaptado de Penteado & Isotani, 2017).

Os conjuntos de dados educacionais encontrados ficam em torno de 4 eixos principais, conforme apontado em Santos (2014): os **censos** (educação básica, ensino superior, dentre outros), a **prestação de contas** das políticas públicas (quantidade de itens e valores repassados em programas como Prouni, PNAE, Pronatec, PNLD, dentre outros), os microdados de **avaliações diagnósticas** (Prova Brasil, ENEM, ANA, dentre outros) e os dados agregados de resultados em **índices de desempenho**. Tal tendência é alinhada às missões das organizações que os publicam: o MEC (Ministério da Educação) e o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira).

3 Melhores Práticas para Dados na Web

Para conduzir a análise dos dados educacionais, foi escolhido como referencial teórico a recomendação da W3C[3], *Melhores Práticas para Dados na Web,* do Inglês *Data on the Web Best Practices* (DWBP). A DWBP é uma recomendação da W3C que traz um conjunto de práticas relacionadas à publicação e uso de dados na Web, projetada para apoiar um ecossistema autossustentável (W3C, 2017). Seu objetivo principal é fazer com que os dados sejam encontrados e compreendidos tanto por humanos quanto por computadores, facilitando a interação entre produtores e consumidores dos dados, oportunizando dados precisos, regularmente atualizados e disponíveis o tempo todo. Seu objetivo é fomentar a expansão continuada da Web como meio para a troca de dados e, com isso, tirar proveito máximo das suas capacidades, como a habilidade de ligar um fato a outro, descobrir recursos relacionados e criar visualizações interativas.

Enquanto que classificações como o esquema de 5 estrelas (Berners-Lee, 2006) buscam apresentar um grau de abertura quanto ao formato dos arquivos, a recomendação DWBP traz práticas mais amplas, ancoradas nas capacidades técnicas da arquitetura da Web, e envolve como os dados são entregues, codificados, mantidos e do relacionamento entre os produtores e consumidores desses dados, estendendo os princípios dos dados abertos governamentais. Esta recomendação tem como prática geral o reaproveitamento de padrões já estabelecidos na Web e

[3] W3C: *World Wide Web Consortium* é a principal organização de padronização da Web, composta por empresas, órgãos governamentais e organizações independentes.

para a recuperação das informações. A Figura 2 demonstra um esquema geral dos principais conceitos relacionados à DWBP.

Neste trabalho adotamos a nomenclatura usada na recomendação: *datasets* (conjunto de dados) e distribuições. O *dataset* é uma coleção de dados, publicada ou mantida por um único agente, não sendo necessariamente na forma de um arquivo em torno de um assunto específico. A *distribuição* representa um recorte específico do *dataset*, podendo ser arquivos, APIs ou feeds RSS. É também chamada de *recurso* em algumas plataformas.

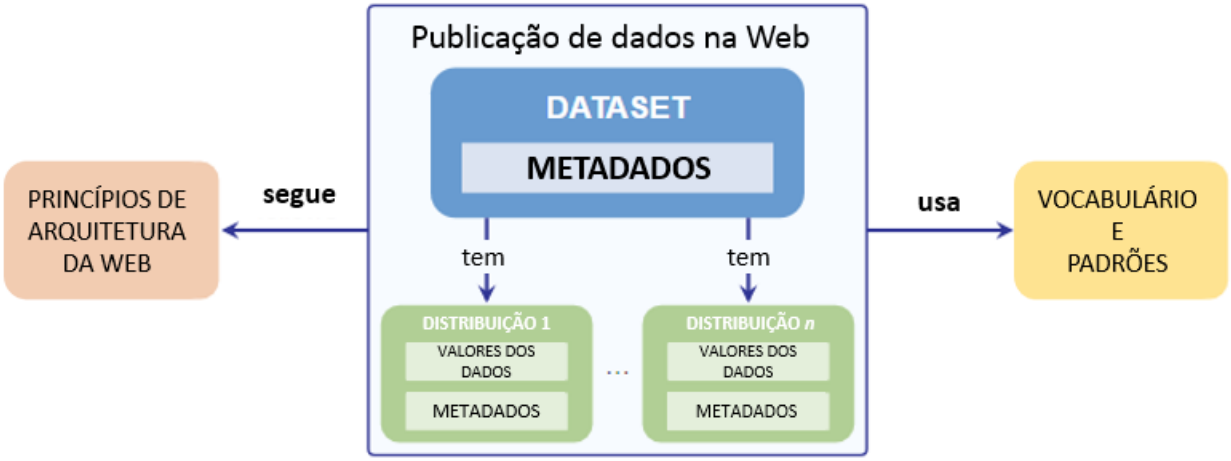

Figura 2. Modelo dos principais conceitos e seus relacionamentos dentro da DWBP (adaptado de W3C, 2017).

A recomendação DWBP é composta por 35 práticas distribuídas em 12 dimensões:

1. **Metadados** (como fornecer metadados para humanos e computadores?): fornecer metadados (BP1); fornecer metadados descritivos (BP2); fornecer metadados estruturais (BP3);
2. **Licença** dos dados (como permitir/restringir o acesso?): fornecer informações sobre as licenças para os dados (BP4);
3. **Proveniência e qualidade** (como adicionar qualidade?): fornecer informações sobre a proveniência dos dados (BP5); fornecer informação sobre a qualidade dos dados (BP6);
4. **Versionamento** dos dados (como localizar diferentes versões e seus históricos?): fornecer indicador de versões (BP7); fornecer histórico de versões (BP8);
5. **Identificação** dos dados (como identificar *datasets* e distribuições?): usar URIs persistentes como identificadores de *datasets* (BP9) e dentro dos *datasets* (BP10); atribuir URIs para versões de *datasets* e séries (BP11);
6. **Formato** dos dados (que formatos de dados usar?): usar formatos padronizados, processáveis por computador (BP12); usar representações neutras de localidade (BP13); fornecer os dados em múltiplos formatos (BP14);
7. **Vocabulário** para os dados (como melhorar a interoperabilidade dos dados?): reusar vocabulários, preferencialmente os padronizados (BP12); escolher o nível correto de formalização (BP14);
8. **Acesso** aos dados (como fornecer acesso aos dados?): fornecer download em massa (BP15); fornecer subconjuntos do *datasets* como um todo (BP16); usar negociação de conteúdo (BP17); fornecer acesso em tempo real (BP17); fornecer dados atualizados (BP18); fornecer explicação para dados que não estejam mais disponíveis (BP19); tornar os dados disponíveis por meio de API (BP20); usar padrões da Web como fundamento para a API (BP21); fornecer documentação completa para a API (BP22); evitar mudanças que causem quebras na API (BP23);

9. **Preservação** dos dados (como os dados podem ser arquivados?): preservar os identificadores (BP27); avaliar a cobertura do *dataset* (BP28);
10. **Feedback** (como engajar os usuários?): coletar feedback dos consumidores dos dados (BP29); tornar o feedback disponível (BP30);
11. **Enriquecimento** dos dados (como adicionar valor aos dados?): enriquecer dados ao gerar mais dados (BP31); fornecer apresentações complementares dos dados (BP32);
12. **Republicação** dos dados (como reusar responsavelmente os dados?): fornecer feedback para o publicador original (BP33); seguir os termos das licenças (BP34); citar a publicação original (BP35);

4 Trabalhos relacionados

Pesquisas recentes levantaram o estado da arte dos dados abertos no governo federal brasileiro, em termos de atendimento aos princípios de dados abertos (Brito et al., 2015), aspectos técnicos dos dados no geral (Oliveira et al., 2016), aplicativos desenvolvidos a partir deles, privacidade dos dados, dentre outros. Klein (2017) propõe um índice de transparência em portais brasileiros de dados abertos governamentais. Embora a literatura em dados abertos seja grande, são poucos os estudos voltados especificamente para os dados educacionais. Parte dos estudos se foca no reuso de dados abertos e visualizações para fins educacionais, como em experiências didáticas que usam dados abertos na prática pedagógica (e.g. Atenas e Havemann (2015)) ou de aplicações para dados abertos conectados (e.g. Sarker e Farhana (2014), Isotani e Bittencourt (2015)).

No cenário internacional, estudos buscaram compreender diversos aspectos dos dados abertos educacionais. Pereira et al. (2017) conduziram um mapeamento da área, discutindo ferramentas, vocabulários e *datasets* disponíveis, contribuindo para a compreensão do estado da arte na área e alguns dos desafios, como a integração de dados entre repositórios distribuídos, a qualidade dos dados disponíveis, a privacidade dos dados, a existência de ontologias e vocabulários, dentre outros. Van Schalkwyk, Willmers e Czerniewicz (2014) analisaram os dados educacionais do ensino superior na África do Sul usando dimensões mais amplas (contexto político, econômico, tecnologias, intermediários, impactos) para compreender a evolução de seu ecossistema de dados, analisando o papel dos intermediários a necessidade de dados mais interoperáveis para melhor absorção por novos usuários ou intermediários. Zuiderwijk e Janssen (2012) analisaram dados abertos educacionais da Holanda, comparando suas políticas e práticas com os dados produzidos por outro ministério, considerando 14 aspectos, desde os princípios adotados, os formatos de arquivos, a natureza dos dados abertos e não abertos, entre outros. Meijer (2007) analisa a abertura de dados educacionais na Holanda por meio da internet e descreve seus efeitos. Pelo outro lado, Cucos (2013) aponta para os benefícios dos dados abertos sobre as escolas para auxiliar pais na busca de escolas para seus filhos.

Além disso, alguns índices de acompanhamento e avaliação das iniciativas nacionais de dados abertos foram desenvolvidos por diferentes organizações que consideram diferentes aspectos dos dados abertos, como: o *Open Data Index* (Open Knowledge International, 2016), o *Open Data Barometer* (World Wide Web Foundation, 2017), o *E-Gov Survey* (Nações Unidas, 2016). Dentre esses índices, o único que traz avaliação especifica para educação é o *Open Data Barometer*, ao analisar a educação como um dos 15 setores essenciais para o funcionamento do governo. Em educação, o índice considera somente se existem dados de desempenho escolar em formato aberto, baseado nos princípios dos dados abertos governamentais. Dentre 92 países avaliados na quarta edição, lançada em 2017, apenas 9 apresentaram dados educacionais, sendo o Brasil um deles. Para o domínio educacional, o índice considera somente se existem dados de desempenho escolar em formato aberto, baseado nos princípios dos dados abertos governamentais. Embora seja uma

iniciativa válida, é insuficiente tanto no escopo educacional (já que julga apenas dados de desempenho) quanto o grau de abertura (apenas o formato em si, não o ecossistema).

No Brasil, Santos (2014) compara dados abertos de desempenho educacional entre Brasil e Inglaterra, avaliando dados do IDEB em relação aos 8 princípios do governo aberto, apontando aspectos não atendidos por esse *dataset*. Berberian, Mello e Camargo (2014) analisam os dados educacionais brasileiros com base em 3 pilares do governo aberto: transparência, participação e colaboração, ressaltando o papel da tecnologia para esta finalidade e o avanço deste setor na disponibilização dos dados. Penteado (2016) apresenta um estudo sobre a correlação entre o desempenho dos municípios brasileiro no IDEB com diversos indicadores socioeconômicos, cruzando dados de diferentes fontes e disponibilizando os dados em formato RDF conectado. Penteado e Isotani (2017) trazem uma linha do tempo dos *datasets* educacionais disponíveis, indicando ano a ano quais *datasets* foram publicados, qual seu objetivo e o formato adotado.

Apesar dos benefícios dos trabalhos apresentados para esclarecer o uso de dados abertos no contexto educacional, nenhuma pesquisa abordou como os dados educacionais estão atualmente estruturados para tirar proveito das capacidades técnicas da Web.

5 Metodologia

Para se responder à questão de pesquisa (*como os dados abertos educacionais estão prontos para serem usados no ecossistema de dados da Web?*) foi adotada a metodologia de estudo de caso (Yin, 2001), tendo como unidade de análise os *datasets.* Com isso, o objetivo foi se aprofundar em poucos casos[4] dentro do universo de *datasets* disponíveis.

Assim, obtivemos uma amostragem estratificada pelas categorias. Para a seleção de casos, usamos a técnica de *casos diversos* (Gerring, 2009), em que é pretendido representar os valores de cada categoria, mas mantendo a variabilidade entre elas. Para validar o grau de abertura dos dados educacionais para as exigências da Web, usamos a recomendação DWBP, da W3C. Ele traz um esforço para uma abertura mais ampla das bases de dados, não limitados somente aos aspectos técnicos, mas também uma visão mais ampla, melhor adequada ao caráter interdisciplinar das iniciativas de dados abertos.

Foi utilizado como fonte de dados o Portal Brasileiro de Dados Abertos (*dados.gov.br*) que centraliza as informações do governo federal brasileiro. Este portal é baseado no software de catálogo de dados CKAN (*ckan.org*), gratuito e de código aberto, utilizado em muitos países pelo mundo em iniciativas parecidas. Ele disponibiliza uma API pela qual é possível recuperar informações desde o nível do portal até dados específicos das distribuições, organizações que publicam os dados e usuários que interagem com o sistema. Assim, apenas dados educacionais de nível federal foram considerados. Neste portal há duas organizações que produzem dados educacionais em maior quantidade: o Ministério da Educação (MEC) e o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), cada qual com suas atribuições.

Como primeiro passo, foi criado um script para acessar a API do CKAN em *dados.gov.br* e recuperar os metadados dos *datasets* e distribuições do MEC e do INEP, num total de 49 e 17 *datasets*, respectivamente. Como critérios adicionais de análise foram analisados os *datasets*: cujas

[4] Estudos de caso se focam em poucos casos que forneçam *insights* sobre a população em geral. Para tanto, deve existir o cuidado de selecionar e tratar da representatividade dos casos selecionados perante sua população. Em pesquisas com população pequena, é necessário conseguir a representatividade dos casos selecionados.

distribuições estejam de fato disponíveis, não sejam duplicados e que tenham mais de uma distribuição, reduzindo para 25 e 10 *datasets*, respectivamente. Como a análise é muito dispendiosa, foi selecionada apenas uma amostra representativa do universo de bases de dados publicadas, baseados em estratos de conjuntos de dados de temas relacionados (técnica de casos diversos). Para isso, foi feita uma análise manual de classificação de cada *dataset* de acordo com as categorias propostas por (Santos, 2014): *dados censitários, dados orçamentários, dados acadêmicos e indicadores de desempenho*. Dentro de cada categoria foi selecionado o mais relevante e, dentro deste, a distribuição mais recente disponível para ser feita a verificação de aderência ao DWBP usando para a avaliação os seguintes critérios: "Sim", "Não", "Parcialmente" e "Não se Aplica" (codificado por S / N / P / NA, respectivamente). Aqui definimos o mais relevante como sendo o *dataset* mais longevo e que não sofreu interrupção em sua publicação, baseado em (Penteado & Isotani, 2017). A Figura 3 ilustra esse procedimento.

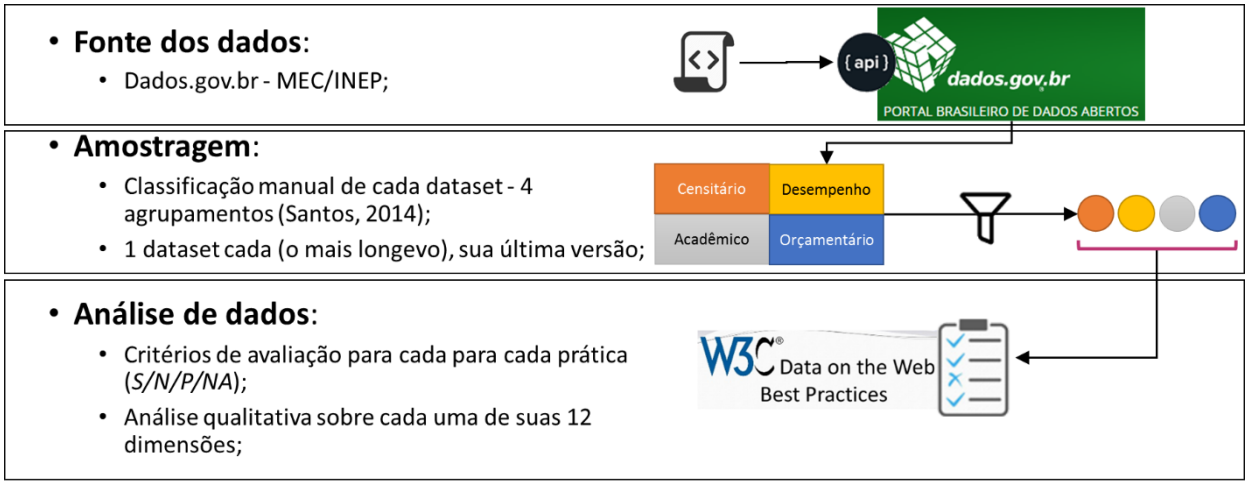

Figura 3. Procedimento metodológico usado neste estudo.

6 Resultados

A Tabela 1 traz os resultados da aplicação do método, avaliando cada prática estabelecida nas 12 dimensões do framework DWBP de acordo com os critérios estabelecidos.

1. Quanto aos *metadados*, o **atendimento foi parcial**, uma vez que eles existem, estão disponíveis em formato eletrônico, mas, em sua maioria, em formato PDF e não em formato semântico. Dois dos datasets não trouxeram os metadados estruturais, com as descrições dos campos. Embora sejam campos fáceis de serem interpretados pelos usuários, fica impossibilitado seu processamento automático.
2. Em relação a *licença dos dados*, o **atendimento também foi parcial** – todos os datasets apresentam a licença de uso dos dados na página do portal, mas não de forma automática.
3. O mesmo (**atendimento parcial**) se aplica à *procedência dos dados*, que estão disponíveis nas respectivas páginas, mas sem acesso automático.
4. A divulgação da *qualidade dos dados* é uma prática que **não foi atendida** por nenhum dos datasets, nem mesmo em suas respectivas páginas.

Tabela 1. Avaliação das práticas do framework DWBP sobre os *datasets* selecionados.

Dimensão	BP	Prática	Datasets selecionados			
			Censitário	*Acadêmico*	*Desempenho*	*Orçamentário*
			2015	*2014*	*2011*	*2014*

			microdados-do-**censo-escolar**	microdados-do-exame-nacional-do-ensino-medio-**enem**	ensino-basico-**ideb-por-municipios**	ensino-basico-pnae-**programa-nacional-de-alimentacao-escolar**
			zip+rar+csv	zip+csv	XML + JSON	XML + JSON
Metadados	1	Fornecer metadados	P	P	P	P
	2	Fornecer metadados descritivos	P	P	P	P
	3	Fornecer metadados estruturais	P	P	N	N
Licença dos dados	4	Fornecer informação de licença de dados	P	P	P	P
Procedência dos dados	5	Fornecer informação sobre procedência dos dados	P	P	P	P
Qualidade dos dados	6	Fornecer informação sobre qualidade dos dados	N	N	N	N
Versionamento dos dados	7	Fornecer indicador de versão	N	N	N	N
	8	Fornecer histórico de versão	N	N	N	N
	9	Usar URIs persistentes como identificadores dos datasets	S	S	S	S
	10	Usar URIs persistentes como identificadores dentro dos datasets	N	N	S	S
	11	Atribuir URIs para versões e séries dos datasets	N	N	N	N
Formatos dos dados	12	Usar formatos de dados processáveis por computador	P	P	S	S
	13	Usar representações de dados com neutralidade de localidade (datas, moedas, números)	N	P	S	S
	14	Fornecer dados em formatos múltiplos	N	N	S	S
	15	Reusar vocabulários, de	N	N	N	N

Vocabulários dos dados		preferência padronizados				
	16	Escolher o nível correto de formalização	N	N	N	N
Acesso aos dados	17	Fornecer download em massa	S	S	S	S
	18	Fornecer subconjuntos para datasets grandes	N	N	N	N
	19	Usar negociação de conteúdo para servir os dados em múltiplos formatos	N	N	N	N
	20	Fornecer acesso em tempo real	N	N	N	N
	21	Fornecer dados atualizados	N	N	N	N
	22	Fornecer explicação para os dados que não estiverem disponíveis	NA	NA	NA	NA
	23	Tornar os dados disponíveis por meio de uma API	N	N	S	S
	24	Usar padrões Web como fundamentos da API	N	N	N	N
	25	Fornecer documentação completa para a API	S	S	S	S
	26	Evitar quebras na API	S	S	S	S
Preservação dos dados	27	Preservar os identificadores	NA	NA	NA	NA
	28	Avaliar a cobertura do *dataset*	NA	NA	NA	NA
Feedback	29	Coletar feedback dos consumidores dos dados	P	P	P	P
	30	Tornar o feedback disponível	P	P	P	P
Enriquecimento dos dados	31	Enriquecer os dados ao gerar novos dados	N	N	N	N
	32	Fornecer apresentações complementares dos dados	N	N	N	N
Republicação	33	Fornecer feedback ao	N	N	N	N

		publicador original				
	34	Seguir os termos de licença	N	N	N	N
	35	Citar a publicação original	N	N	N	N

5. Quanto *ao versionamento dos dados*, houve **atendimento parcial**. Todos os *datasets* trazem um identificador único e interpretável por humanos, já que é um requisito da plataforma CKAN. Porém, as distribuições não trazem nomes interpretáveis, apenas UUID (*Universally Unique Identifier*, *string* hexadecimal de 128 bits usado para identificar recursos de maneira probabilisticamente inequívoca). Nenhum dos *datasets* e suas distribuições trazem explicitamente um valor de versão e seu histórico, ficando implícita essa informação na nomenclatura adotada pelas distribuições (por ex.: ENEM 2009, ENEM 2010, etc.), sem disponibilidade semântica desses metadados e não garantindo assim uma 'última versão' do dataset, que seria a distribuição mais recente. Dois dos *datasets* apresentaram URIs persistentes como identificadores dentro de seus dados. Eles fazem referências a outras entidades e conceitos, representadas por uma URI em formato JSON (por ex.: a URI http://api.pgi.gov.br/api/1/fonte/155.json referencia o gabinete do ministro do Ministério da Educação).
6. Em relação aos *formatos dos dados*, houve **atendimento parcial**. Dois dos *datasets* estão disponíveis em formato XML e JSON e os outros, com muito mais dados, estão em formato compactado, incluindo outros arquivos compactados dentro de si. Por sua vez, esses arquivos estão em formato CSV. Embora os *datasets* que tenham mais de uma representação dos dados, cada um tem sua URI codificada de maneira diferente, o que impede a negociação de conteúdo ao requisitá-lo sob demanda via software. As representações de dados de moedas, datas e números são atendidas de modo diferente entre os casos selecionados. Os formatos em JSON e XML trazem esses dados decompostos, especificando dentro do próprio arquivo ou apontando para URIs externas o formato utilizado. Um dos casos foi considerado parcial, pois seus dados não se encaixaram em dados que necessitasse de localização apropriada.
7. Os *vocabulários de dados* tiveram **atendimento parcial**, pois são adotados em todos os casos selecionados, fazendo parte inclusive da regulamentação da INDA[5], que especifica quais metadados são obrigatórios e quais são opcionais quando da publicação dos datasets. Porém, estão disponíveis somente nas páginas do portal e não para processamento automático. Além disso, não seguem os requisitos semânticos para a anotação dos metadados.
8. O *acesso aos dados* também teve **atendimento parcial** e traz consigo grandes desafios. O download em massa e a API versionada e documentada são funcionalidades garantidas pela plataforma CKAN. No entanto, para os casos selecionados, não é possível fazer download parcial dos dados, as chamadas na API não estão em formato de padrões Web, e o *download* não está disponível para diferentes formatos por meio de negociação de conteúdo. A partir da API do CKAN é possível resgatar os metadados dos *datasets* mas não os dados em si. A atualização dos dados também é um problema encontrado em todos os casos, que não trazem suas versões mais recentes, conforme sua granularidade temporal.
9. Quanto ao *feedback*, ouve **atendimento parcial**. O portal fornece a opção de compartilhar o endereço do dataset em redes sociais (Facebook, Twitter e Google+). No entanto, os

[5] INDA (Infraestrutura Nacional para Dados Abertos): conjunto de padrões, tecnologias, procedimentos e mecanismos de controle necessários para atender às condições de disseminação e compartilhamento de dados e informações públicas no modelo de Dados Abertos. Fonte: http://wiki.dados.gov.br/.

comentários não são salvos no portal e exibidos para que outros usuários possam comentar em cima. Recentemente, foi adicionada uma extensão ao portal, integrada à ouvidoria geral da união[6], para que os usuários possam opinar (com 'sim' ou 'não') se os datasets apresentam os dados como esperado, tendo os resultados exibidos junto à página do *dataset* no portal.

10. O *enriquecimento dos dados* **não atendeu** às práticas, pois também não esteve presente para os *datasets* analisados. Não existem campos calculados ou conectados nos *datasets* e nem apresentam formas complementares para a visualização dos dados – embora esta funcionalidade esteja presente para outros *datasets* no portal.
11. Em relação à *republicação de dados*, **não houve atendimento** das práticas. Os *datasets* educacionais usam informações de identificação de municípios e suas regiões geográficas vindas do IBGE (Instituto Brasileiro de Geografia e Estatística), replicando a licença de dados original, porém não é apontado qual a fonte original desses dados (sua atribuição), o que dificulta o acesso à fonte original desses dados complementares. Além disso, não foram encontradas evidências de feedback (BP33) ao IBGE sobre o uso de sua base.

Não foi possível avaliar algumas das práticas, em especial as relativas à *preservação dos dados* (BP27 e BP28), acesso a dados não mais disponíveis (BP22), já que não foram identificados *datasets* ou distribuições que tivessem sido arquivados.

7 Discussão

Neste trabalho, os *datasets* escolhidos são amostras relevantes do universo de dados educacionais. Ao analisar dados do e-SIC[7], sistema responsável pelo acompanhamento de pedidos de acesso à informação do governo federal, foram adotados 3 dos *datasets* educacionais mais requisitados, representando mais da metade das requisições de dados por parte dos cidadãos. A seguir, são analisados 3 aspectos: a adequação à recomendação DWBP, discutindo as possíveis causas do não-atendimento das práticas estabelecidas; possíveis intervenções, tanto em nível de produção de dados quanto de organização do catálogo de dados; outras características que impactam a exploração das capacidades da Web pelos dados publicados.

7.1 Adequação à recomendação DWBP.

A Figura 4 traz uma visão geral de práticas agrupadas por dimensão do DWBP. Observando as práticas atendidas, apenas 3 dimensões tiveram ao menos uma prática contemplada; entretanto nenhuma completamente: *acesso*, *formato e versionamento* dos dados. Outras 4 não tiveram prática alguma atendida: *enriquecimento*, *vocabulário*, *qualidade* e *republicação* dos dados. Nesta subseção discutimos as prováveis causas desses resultados.

[6] http://dados.gov.br/noticia/governo-integra-portal-brasileiro-de-dados-abertos-e-sistema-da-ouvidoria-geral-da-uniao

[7] http://download.inep.gov.br/institucional/legislacao/2016/portaria_n370.pdf

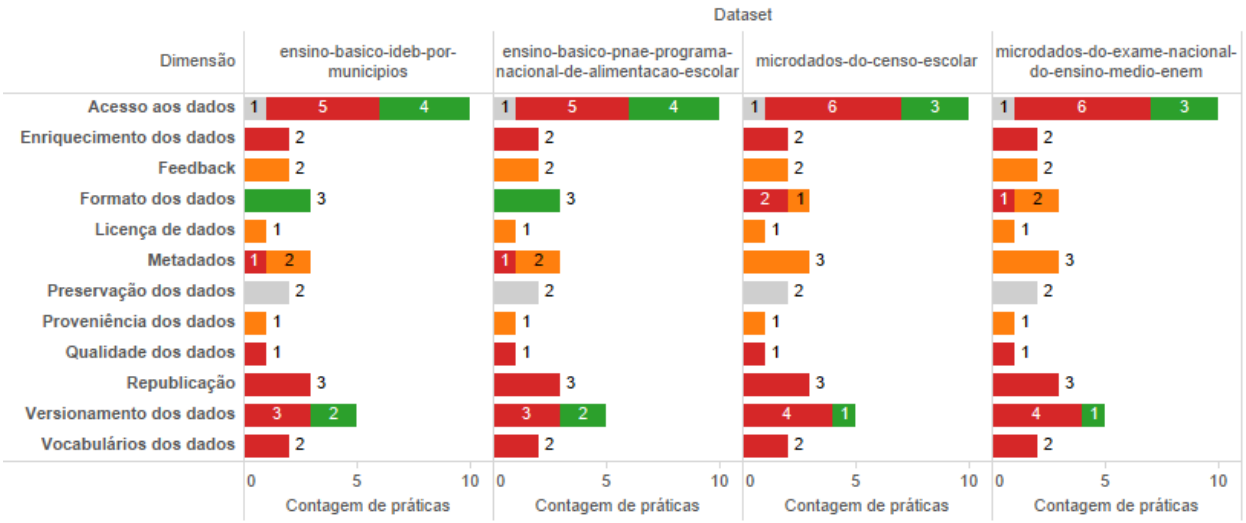

Figura 4. Atendimento das práticas agrupadas por dimensão.

Dessas, a mais atendida foi a de *formato* dos arquivos, sinalizando a direção recente de fornecer dados em formato aberto e processáveis (BP12), estabelecido pelo Decreto 8.777/2016 (BRASIL, 2016) e detalhado pela INDA, em sua cartilha técnica para publicação de dados abertos[8], sugerindo os formatos JSON, XML, CSV, ODS e RDF como preferenciais, como demonstrado pelos *datasets* selecionados neste artigo. Dois deles apresentaram dados tanto em formato JSON quanto XML (BP14) e os outros dois em formato CSV compactado, em suas versões mais recentes.

Na dimensão de *acesso*, as práticas atendidas foram de *download* dos dados em massa (BP17) e o de gerenciamento de API (BP 25 e 26). A primeira prática é a forma mais simples de disponibilização de dados, sendo usada desde os primórdios da divulgação de dados abertos governamentais (Penteado & Isotani, 2017). Já as outras decorrem da escolha da ferramenta de catálogo de dados adotada pelo governo federal (CKAN), que tem essas práticas adotadas por padrão. Pelo outro lado, a adoção desta ferramenta não traz outras funcionalidades que atendam a outras práticas de acesso aos dados, como *download* de subconjuntos dos dados (BP18), negociação de conteúdo para baixar os dados em diferentes formatos (BP19), acesso aos dados por meio de API (BP23) e do uso de padrões Web nas API (BP24).

Na dimensão de *versionamento*, o uso de URI persistentes para identificar os *datasets* (BP9) é uma funcionalidade padrão para a ferramenta que publica os dados. No entanto, a forma com que o CKAN especifica seu padrão viola o que está estabelecido na política nacional de URIs para publicação de dados no governo, parte da e-VoG (Brasil, 2014) – que adota padrões Web para o acesso e o versionamento aos dados publicados. O indicador (BP7) e a série (BP11) de versões não estão explicitamente presentes no acesso às distribuições. Em todos os casos selecionados isso é feito por meio de arquivos diferentes com essas informações contidas no título da distribuição e ordenada na ferramenta (ex.: Censo Escolar 2015, Censo Escolar 2014 e assim por diante), cada qual com sua URI. Esta é uma escolha, que, no entanto, dificulta formas mais automáticas de acesso a esses metadados. Além do mais, não é mantido o rastreamento de modificações nos dados (BP8), em casos de erros e ratificações no conteúdo da publicação, como o ocorrido na divulgação de dados do ENEM que não continham institutos federais e que foram posteriormente corrigidos[9]. O uso de URI nos dados em si (BP10) está presente em dois *datasets*, como parte da Plataforma de

[8] http://dados.gov.br/pagina/cartilha-publicacao-dados-abertos

[9] https://g1.globo.com/educacao/noticia/inep-admite-equivoco-com-federais-e-diz-que-vai-refazer-enem-2015-por-escola.ghtml

Gestão de Indicadores (PGI), alinhado com os conceitos de dados conectados. Entretanto, a PGI foi desativada em 2015 e outros *datasets* mais recentes não parecem ter continuado essa prática.

As dimensões de *enriquecimento* e *vocabulário* de dados não tiveram prática alguma atendidas. Isso sinaliza que os dados disponíveis ainda carecem de um nível maior de abertura, conforme definido por Berners-Lee (2006), em que os dados devem estrar anotados semanticamente por vocabulários padronizados e enriquecidos por dados de outras fontes. O uso de vocabulários e ontologias é algo recomendado no e-PING[10] (arquitetura que define as políticas e especificações técnicas que regulamentam as TIC nos serviços de governo eletrônico no âmbito federal, como forma de incentivar a interoperabilidade entre diferentes entidades da Federação) e no e-VOG[11] (Vocabulários e Ontologias do Governo Eletrônico – conjunto de padrões, ferramentas e metodologias para possibilitar o intercâmbio de informações no formato semântico vinculado à e-PING). No escopo do e-PING, isso significa que esse atendimento não é obrigatório, mas sugerido para a adoção em novos projetos de TI. Além do mais, são poucas as ontologias e vocabulários estáveis presente no e-VoG, nenhuma delas no domínio educacional.

Outra dimensão com atendimento pobre foi a de *republicação* dos dados. Nela, foram identificados dados de outras fontes que não as selecionadas. No caso, informações do IBGE relativas à localização das unidades escolares referenciadas nos *datasets* selecionados. Embora os dados do IBGE são licenciados via ODbL (*Open Database License*), o mesmo que o PNAE e do IDEB por municípios, e ser uma forma mais moderna do *Creative Commons Attributions* usado pelo Censo Escolar, não existe a atribuição, a citação ou mesmo feedback explícitos nos dados – termos que constituem as licenças adotadas. Por se tratarem de entidades sob o mesmo comando, o governo federal, supõe-se que tais práticas não sejam obrigatórias para esse tipo de formalização.

A dimensão de *qualidade* dos dados, embora contenha apenas uma prática (BP6), não foi atendida por quaisquer das distribuições analisadas. A DWBP se mostra bem flexível quanto ao atendimento desta prática, não especificando quaisquer tipos de métricas ou aspectos de qualidade. Em seu documento, traz como exemplos a disponibilidade da distribuição para download e a completude dos dados no arquivo (isto é, a proporção de registros existentes dentro do número total de registros esperados para o escopo do *dataset*). No entanto, em nenhuma das distribuições analisadas foram encontradas evidências de informações sobre a qualidade dos dados disponibilizados. Em relação à normatização desse aspecto, a Lei 8.777/2016 (Brasil, 2016) traz que os conjuntos de dados devem obedecer a "padrões mínimos de qualidade" (art. 2) e apresentar "informação suficiente para a compreensão de eventuais ressalvas quanto à sua qualidade e integridade" (art. 3). Do mesmo modo, a INDA sugere em sua página[12] diversos modelos de avaliação de qualidade dos datasets, em diferentes aspectos. Outra métrica de qualidade pode ser o índice de transparência apresentado por Klein (2017). No entanto, não existem materiais que especifiquem formas mais concretas de determinação da qualidade ou mesmo normas que obriguem sua adoção.

7.2 Possíveis intervenções na oferta dos dados abertos educacionais.

A seguir são listadas possíveis frentes de trabalho, com sugestões de ações para a adequação dos dados educacionais à DWBP.

- *Portal de dados*: foi observado durante a análise que algumas das melhores práticas são mais dependentes do software de catálogo de dados que hospeda os *datasets*, do que dos dados em si. Isso também é notado no relatório de implementação disponibilizado pela

[10] http://eping.governoeletronico.gov.br/

[11] http://vocab.e.gov.br/

[12] http://wiki.dados.gov.br/Produto-GT-2-Modelo-de-Maturidade.ashx

W3C[13], que aponta como os softwares de catálogo de dados mais comuns atendem às práticas dessa recomendação. Parte das lacunas levantadas são endereçadas por extensões da plataforma CKAN. Por exemplo, as BP 1, 2 e 3 podem ser atendidas pela extensão *dcat*, que era parte da instalação nativa, mas que foi extraída da instalação nativa do CKAN em suas versões mais recentes; as BP29 e 30 (*feedback*) podem ser atendidas pela extensão *ytp-comment*; que adicionam formulários de feedback para que os consumidores dos dados possam opinar sobre o conteúdo dos dados; BP6 com a extensão *qa*, que traz informações automáticas sobre a qualidade dos dados; BP32 com os *viewers* para a visualização de dados, dentre outros. Alguns datasets publicados mais recentemente trazem algumas funcionalidades, como a visualização de dados e a API de consulta de dados – porém não nos dados educacionais dos órgãos selecionados;

- *Produção dos dados*: pelo outro lado, certas ações podem fazer parte do processo de produção dos dados a serem publicados, de modo que atendam às práticas devidas da DWBP independentemente de onde estiverem catalogadas. Neste trabalho, os datasets e suas distribuições trazem metadados que atendem parcialmente às práticas, por não os disponibilizarem em formato semântico (e.g. BP1 a 5, BP15, BP16) e de forma automática (via API), o que facilitaria o processo de enriquecimento, republicação e adoção de vocabulários. Alguns dos metadados são parte da INDA – sendo campos obrigatórios ou opcionais, o que inclui o uso de vocabulários já existentes, como o VCGE[14]. A BP16 diz respeito a neutralidade dos dados, ou seja, a explicitação de unidades de medida sempre que cabível (datas, valores financeiros, entre outros), o que implica modificar a estrutura dos dados já existentes. Em outra prática (BP10) o uso de URI como identificadores dentro dos dados (BP10) também modifica a estrutura de publicação atual, visando a conexão semântica de dados entre diferentes fontes. Para tanto, deve existir uma organização dos termos e conceitos, representados por URI, que possam ser reutilizados em diferentes pontos dentro dos arquivos de dados, de forma permanente, como foi adotado pela PGI. Barbosa et al. (2017) trazem uma compilação de ferramentas que dão suporte a alguns dos passos para a publicação e o consumo de dados abertos conectados e que podem ser considerados neste processo.

7.3 Outras características dos dados abertos educacionais

Como demonstrado previamente, houve um grande crescimento na disponibilização de *datasets* educacionais no âmbito do governo federal brasileiro (*cf.* Figura 1). Nos últimos anos, houve a consolidação de leis e de estruturas de governança para que políticas e planos de dados abertos fossem estendidos a todos os órgãos federais. E, com isso, houve a uniformização em termos de formatos, periodicidade, definição de papéis e responsáveis, dentre outros aspectos. No entanto, isso gerou, por consequência, a existência de um grande contingente de dados legados que não atendem a diversas dessas práticas já adotadas nas distribuições dos últimos anos. Além disso, a governança adotada para publicação dos dados governamentais é descentralizada, ou seja, a entidade governamental produz os dados, podendo disponibilizá-la em seu site, mas que deve encaminhar para o portal *dados.gov.br* para centralização do acesso, conforme instrução normativa da INDA (Brasil, 2012). Se, por um lado, temos as instituições educacionais produzindo e disponibilizando os dados, temos outra responsável pela centralização e catalogação dos mesmos. Isso exemplifica a complexidade das iniciativas de dados abertos, que devem possuir um plano de governança bem coordenado entre todos os envolvidos.

Um problema foi a descontinuidade da publicação dos *datasets* no portal *dados.gov.br*. Grande parte dos *datasets* foi concebida para atender solicitações pontuais de dados, tendo até mesmo

[13] http://w3c.github.io/dwbp/dwbp-implementation-report.html

[14] http://governoeletronico.gov.br/documentos-e-arquivos/VCGE_2_1_0.pdf/download

apenas uma distribuição publicada ao longo dos anos. Outros, como o Censo Escolar e dos microdados ENEM estão atualizados na página do INEP - o órgão responsável por sua manutenção - mas não foram disponibilizados no portal *dados.gov.br*. Além disso, no ano de 2017 e até o momento da escrita deste trabalho não foram publicadas novas distribuições nem novos *datasets* no escopo do MEC e do INEP, em desacordo com o estabelecido no Plano de Dados Abertos do INEP para o período 2016-2018.

Outra constatação foi a grande quantidade de dados duplicados presentes nos *datasets* do MEC e INEP. A partir de um conjunto de dados originais, são criadas diversas visões sobre esses dados, agregados de forma a facilitar a compreensão em diferentes unidades de análise. Por exemplo, os dados do IDEB. Os microdados do SAEB são usados para o cálculo do IDEB. Depois do cálculo, são disponibilizados *datasets* que agregam os resultados por município, estado e do Brasil como um todo, cada um deles de forma duplicada. Além disso, dentro de cada dataset existem diversas divisões, como a divisão por escolas municipais, estaduais, federais e privadas, por sua vez divididas em distribuições para o ensino fundamental 1 e 2. Supõe-se que isso seja o efeito de diferentes requisições de dados por parte da sociedade que, para facilitar o acesso, tenham sido criados arquivos compilados e independentes e para download. No entanto, do ponto de vista técnico, existe um alto grau de duplicidade nos dados e a consequente uso de espaço em disco. Outra consequência decorrente deste fato é a inflação nos números de publicações, ao enfocar somente na quantidade de *datasets* publicados e não em sua natureza.

Em geral, as informações disponibilizadas têm um formato de indicadores temporais, ou seja, que apresentam um determinado valor em um determinado instante de tempo. Isso influencia na estrutura dos arquivos. Enquanto alguns contém todo o conteúdo de sua série histórica em um único arquivo, outros datasets ‘fatiam’ o conteúdo em diferentes arquivos. Por exemplo, nos casos selecionados neste trabalho, temos o Censo Escolar, que a cada ano gera um arquivo de dados diferente. Já o PNAE contém todos os dados de sua série histórica em um único arquivo. Esta diferença na estruturação influencia em como os dados podem ser processados automaticamente, necessitando a indicação para o correto processamento dos dados.

Outro fator que complica o processamento automático dos dados é a natureza dos arquivos disponibilizados para download. Talvez por limitação da ferramenta CKAN, arquivos complementares aos dados, como dicionários de dados, ontologias, documentos de metodologia são disponibilizados como arquivos de dados. Enquanto que essa centralização é conveniente para usuários humanos no portal, dificulta muito a recuperação automática de informações via API.

A falta de documentação descritiva dos dados publicados também pode ser considerada um problema. Dentre os *datasets* analisados, dois deles não contém dicionários de dados, nem sequer em formato para leitura humana. Apesar de apresentar informações relativamente simples, a existência de um documento descritivo, criado pelo produtor dos dados, pode evitar problemas relacionados ao entendimento sobre as propriedades de um campo de dados

8 Conclusões

De forma geral, os resultados sugerem que o portal *dados.gov.br* parece ter evoluído para o consumo de dados por humanos, em forma de catálogos dos conjuntos de dados, trazendo os dados, metadados, formatos de arquivos e informações auxiliares para que pessoas possam baixar, compreender e usar os dados disponíveis. A recomendação DWBP auxiliou a evidenciação desta constatação por meio da checagem de suas práticas, consideradas como essenciais para o aproveitamento das potencialidades da Web. Poucas práticas foram plenamente atendidas pelos *datasets* analisados e boa parte foi parcialmente atendida por não disponibilizar mecanismo para acesso automática aos dados ou metadados.

Para efeito ilustrativo, peguemos o caso do Censo Escolar da Educação Básica, disponibilizado em um arquivo compactado de 1.7 GB de tamanho (considerando o censo de 2017, publicado em fevereiro de 2018), composto por outros arquivos de dados compactados, metadados em formato de planilha e informações sobre a metodologia em formato PDF, como ilustrado na Figura 5. Para uma pessoa, é algo relativamente conveniente de se baixar todo o conjunto de uma vez. Pelo outro lado, isso torna muito complexo seu processamento automático e em tempo real por agentes de software (BP12). O acesso a um determinado ponto de dado nesta distribuição fica muito dificultado.

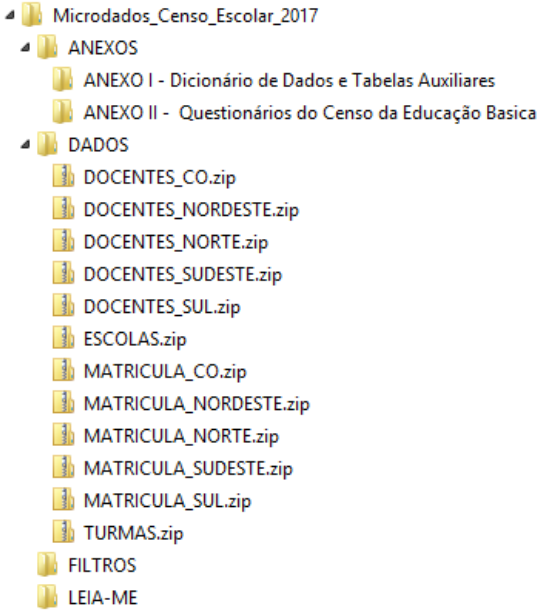

Figura 5. Estrutura de arquivos no *dataset* compactado do censo escolar da educação básica de 2017.

Atualmente, o MEC e o INEP desenvolvem suas ações de dados abertos no escopo do Plano de Dados Abertos de 2016/2018, com seu escopo já fechado. Os *datasets* escolhidos são amostras representativas do universo, pois foram adotados 3 dos datasets mais requisitados entre as entidades selecionadas, representando mais da metade das requisições de dados por parte da população[15] e de diversidade quanto à sua natureza. Deste modo, espera-se que a análise feita e as intervenções apontadas possam trazer benefícios para a adequação dos dados abertos educacionais brasileiros para atingir o potencial que a Web oferece.

Como trabalhos futuros, poderão ser sugeridos processos, métodos e ferramentas para que haja a ampliação na oferta de dados abertos, de preferência já em seu formato conectado (Web de dados). Com isso, espera-se que exista todo aparato para o suporte na tomada de decisão e geração de políticas públicas baseadas em evidências.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, CNPq e FAPESP (Processo 2015/24507-2) pelo apoio financeiro dado ao projeto.

Referências

[15] http://download.inep.gov.br/institucional/legislacao/2016/portaria_n370.pdf

Atenas, J., Havemann, L. (2015). "Open Data as Open Educational Resources: Case studies of emerging practice". London: Open Knowledge, Open Education Working Group. DOI: 10.6084/m9.figshare.1590031.

Bandeira, J., Ávila, T., Alcantara, W., Sobrinho, A., Bittencourt, I. I., Isotani, S. (2015). "Dados abertos conectados para a Educação". Jornada de Atualização em Informática na Educação 4.1 (2015): 47-69. http://br-ie.org/pub/index.php/pie/article/viewFile/3551/2937.

Barbosa, A., Bittencourt, I. I., Siqueira, S. W., Silva, R. D., & Calado, I. (2017). The Use of Software Tools in Linked Data Publication and Consumption: A Systematic Literature Review. International Journal on Semantic Web and Information Systems (IJSWIS), 13(4), 68-88. DOI:10.4018/IJSWIS.2017100104.

Berberian, C. F. Q., Mello, P. J. S. M., Camargo, R. M. P. (2014). "Governo Aberto: a tecnologia contribuindo para maior aproximação entre o Estado e a Sociedade". Revista do TCU, n. 131. Dez 2014. P 30-39. https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/60/67.

Berners-Lee, T. (2006). Linked data: design issues. Disponível em: http://www.w3.org/DesignIssues/LinkedData.html. Acessado em: 17 jun. 2017.

Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011. "Lei de Acesso à Informação". Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm.

Brasil (2012). Instrução normativa nº4, de 12 de Abril de 2012. "Infraestrutura Nacional de Dados Abertos – INDA". Disponível em: https://www.governoeletronico.gov.br/documentos-e-arquivos/IN%204%202014_compilada.pdf.

Brasil (2014). Política de URIs para Publicação de Dados no Governo. Disponível em: https://www.governoeletronico.gov.br/documentos-e-arquivos/Politica_URIs_Publicacao_Dados_Governo.pdf.

Brasil (2016). Lei nº 8.777, de 11 de maio de 2016. Política de Dados Abertos do Poder Executivo federal. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm.

Brito, K. S., Costa, M. A. S., Garcia, V. C. & Meira, S. R. L. (2015). "Is brazilian open government data actually open data?" An analysis of the current scenario. International Journal of E-planning Research, Vol.4 (2). DOI: 10.4018/ijepr.2015040104.

Cucos, R. (2013). "Open Government Data: Helping Parents to find the Best School for their Kids". The WorldBank, 2013. Disponível em: http://blogs.w/orldbank.org/ic4d/open-government-/data-helping-parents-find-best-school-their-kids. Acessado em: 14 jun. 2017.

Gerring, J. (2009) "Case selection for case-study analysis: qualitative and quantitative techniques", The Oxford Handbook of Political Methodology, J. M. Box-Steffensmeier, H. E. Brady e D. Collier, England, Oxford University Press, p. 645-684. DOI: 10.1093/oxfordhb/9780199286546.003.0028

Guy, M. (2016). "The Open Education Working Group: Bringing People, Projects and Data Together". Open Data for Education: Linked, Shared, and Reusable Data for Teaching and Learning, Springer International Publishing. Eds: Mouromtsev, D. d'Aquin, M. p. 166-187; DOI: 10.1007/978-3-319-30493-9_9.

International Budget Partnership. (2015). "Open Budget Index Rankings". Disponível em: http://www.internationalbudget.org/opening-budgets/open-budget-initiative/open-budget-survey/publications-2/rankings-key-findings/rankings/. Acesso em: 15 de fevereiro de 2017.

Isotani, S., Bittencourt, I. I. (2015). "Dados abertos conectados". Editora Novatec. DOI: DOI:10.13140/RG.2.1.4355.6329.

Klein, R. H. (2017). "Mecanismos de ampliação da transparência em portais de dados abertos governamentais brasileiros à luz da *Accountability Theory*", Tese (Doutorado em Administração) – Escola de negócios, Pontifícia Universidade Católica Do Rio Grande Do Sul, Porto Alegre, 2017. Disponível em: http://tede2.pucrs.br/tede2/handle/tede/7724.

McKinsey. (2013). "Open data: Unlocking innovation and performance with liquid information". McKinsey Global Institute. [S.l.]. 2013. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information

Meijer, A. J. (2007). "Publishing public performance results on the Internet Do stakeholders use the Internet to hold Dutch public service organizations to account?" Government Information Quarterly, v. 24, p. 165-185. DOI: 10.1016/j.giq.2006.01.014.

Nações Unidas. (2016). "United Nations E-government survey 2016". Disponível em: https://publicadministration.un.org/egovkb/en-us/reports/un-e-government-survey-2016. Acessado em: 10 de junho de 2017.

OECD (2017). "Open government data". Disponível em: http://www.oecd.org/gov/digital-government/open-government-data.htm. Acesso: 03 de Março de 2018.

OGP (2011). "Open Government Partnership". Disponível em: https://www.opengovpartnership.org/about/about-ogp. Acessado em: 10 de Março de 2017.

Oliveira, M. I. S, Oliveira, H. R., Oliveira, L. A., Lóscio, B. F. (2016). Open government data portals analysis: the Brazilian case. International Conference on Digital Government Research. Junho, 2016, Xangai, China. P. 415-424. DOI:10.1145/2912160.2912163.

Open Knowledge International (2016). "Global Open Data Index". Disponível em: https://index.okfn.org/place. Acesso: 25 maio 2017.

Parycek, P., Sachs, M. (2010). "Open government-information flow in Web 2.0". European Journal of ePractice, Vol. 9(1), p. 1–12. [GS SEARCH]

Penteado, B. E. (2016). "Correlational Analysis Between School Performance and Municipal Indicators in Brazil Supported by Linked Open Data". Proceedings of the 25th International Conference Companion on World Wide Web, p. 507-512. DOI: 10.1145/2872518.2890459.

Penteado, B. E., Isotani, S. (2017). "Dados abertos educacionais: que informações temos disponíveis?". Anais do VI Congresso Brasileiro de Educação, vol. 4. Bauru, Julho de 2017, ISBN 978-85-5444-002-2, p. 1933-1938. Disponível em: http://www.cbe-unesp.com.br/2017/pages/anais_cbe_v01.pdf.

Penteado, B. E., 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, 2017, p. 526-535. Recife. DOI: 10.5753/cbie.sbie.2017.526.

Pereira, C. K., Siqueira, S., Nunes, B. P., Dietze, S. (2017). Linked data in Education: a survey and a synthesis of actual research and future challenges. IEEE Transactions on Learning Technologies. Vol. PP (99), p 1-15. DOI: 10.1109/TLT.2017.2787659.

Santos, O. A. R. (2014) "Minha escola transparente: uma análise comparativa do uso de dados governamentais abertos na educação básica no Brasil e Inglaterra", Dissertação (Mestrado profissional em Adm. Pública) – EBAP, FGV, Rio de Janeiro. 2014. Disponível em: http://bibliotecadigital.fgv.br/dspace/handle/10438/12927.

Sarker, Farhana (2014) "Linked data technologies to support higher education challenges: student retention, progression and completion". University of Southampton, Physical Sciences and Engineering. Tese de doutorado, 273pp. Disponível em: https://eprints.soton.ac.uk/374317/.

Siqueira, S.; Bittencourt, I. I.; Isotani, S.; Nunes, B. (2017). Information Systems based on (Linked) Open Data: From Openness to Innovation. In: Boscarioli, C.; Araujo, R. M.; Maciel, R. S. P. (Org.). I GranDSI-BR - Grand Research Challenges in Information Systems in Brazil 2016-2026. 1ed. Rio Grande do Sul: SBC, 2017, v. 1, p. 52-61. Disponível em: http://www2.sbc.org.br/ce-si/arquivos/GranDSI-BR_Ebook-Final.pdf.

Ubaldi, B. (2013), "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", OECD Working Papers on Public Governance, No. 22, OECD Publishing, Paris. DOI:10.1787/5k46bj4f03s7-en.

Van Schalkwyk, F., Willmers, M., Czerniewicz, L. (2014). "Open data in the governance of South African higher education". Disponível em: http://eprints.bbk.ac.uk/13353/1/Book-Open-Data-as-Open-Educational-Resources1.pdf. Acesso em: 22 junho 2017.

Yin, R. K. (2001) Estudo de caso: planejamento e métodos, 2ª edição. Bookman: Porto Alegre. [GS SEARCH]

W3C (2017). "Data on the Web Best Practices". W3C Recommendation 31 January 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acessado em: 10 junho 2017.

Wood, D., Zaidman, M., Ruth, L., Hausenblas, M. (2014). "Linked Data: Structured data on the Web". New York: Manning Publications Co. [GS SEARCH]

World Wide Web Foundation. (2017). Open Data Barometer: Fourth Edition. Maio, 2017. Disponível em: http://opendatabarometer.org/doc/4thEdition/ODB-4thEdition-GlobalReport.pdf. Acesso em: 17 julho 2017.

Zuiderwijk, A., Janssen, M. (2012). "A comparison of open data policies and their implementation in two Dutch ministries". Digital Government Research, Maryland, EUA. p. 84-89. DOI: 10.1145/2307729.2307744.