

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/320344888>

# Dados abertos educacionais no Brasil e sua preparação para os dados abertos na web

Conference Paper · October 2017

DOI: 10.5753/cbie.sbie.2017.526

CITATIONS

0

READS

102

3 authors:



**Bruno Penteado**

12 PUBLICATIONS 21 CITATIONS

[SEE PROFILE](#)



**Ig Ibert Bittencourt**

Universidade Federal de Alagoas

173 PUBLICATIONS 807 CITATIONS

[SEE PROFILE](#)



**Seiji Isotani**

University of São Paulo

225 PUBLICATIONS 1,178 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Data-Informed Pedagogical Decision-Making [View project](#)



Pensamento Computacional - Ensino de Computação [View project](#)

# Dados abertos educacionais no Brasil e sua preparação para os dados abertos na web

Bruno Elias Penteado<sup>1</sup>, Ig Ibert Bittencourt<sup>2</sup>, Seiji Isotani<sup>1</sup>

<sup>1</sup>Instituto de Ciências Matemática e Computação – Universidade de São Paulo (USP)  
13566-590 – São Carlos – SP – Brasil

<sup>2</sup>Instituto de Computação – Universidade Federal de Alagoas (UFAL) – 57072-970 –  
Maceió – AL - Brasil

brunopenteado@usp.br, ig.ibert@ic.ufal.br, sisotani@icmc.usp.br

**Abstract.** *Open education data carry important information on the education landscape. Its publishing impacts transparency and economic potential to society. The Web brings many features that can extend this potential and, to achieve that goal, W3C – organization responsible for the standards in the Web - developed the DWBP recommendation, for the production and sharing of open data that realizes the potentials of Web architecture. So far, no research was found assessing open data based on this referential. In this work, we applied a case study with diverse categories of education data analyzing the DWBP practices. As result, most of the advocated practices are not met in the selected cases. In the end, it brings potential interventions either for data production and for data portals which host them.*

**Resumo.** *Os dados abertos educacionais trazem informações importantes sobre o cenário educacional. Sua publicação traz impactos em transparência e potencial econômico para a sociedade como um todo. A Web traz muitas funcionalidades que podem estender esse potencial e, para isso, a W3C - organização responsável pelas padronizações na Web - criou a recomendação DWBP, para a produção de dados abertos que tirem proveito da arquitetura da Web. Até o momento, não foram encontradas pesquisas que avaliem os dados de acordo com esse referencial. Neste trabalho, aplicamos estudo de caso com diferentes categorias de dados educacionais analisando as práticas da DWBP. Como resultado, a maior parte das práticas preconizadas não são atualmente atendidas pelos casos selecionados. Ao final, traz possibilidades de intervenções tanto para a produção dos dados quanto para os portais que os hospedam.*

## 1. Introdução

Na última década, diversos países pelo mundo têm apresentado iniciativas de abertura de seus dados governamentais, no chamado movimento de *governo aberto*, com objetivo de fornecer espaço para a abertura, transparência e diálogo contínuo entre o governo e seus cidadãos (Parycek e Sachs, 2015), viabilizado por meio dos *dados abertos*. Ao liberar os dados em formatos não proprietários e sem licenças restritivas, o governo permite que diferentes setores da sociedade se apropriem das informações e gerem análises, produtos e serviços que retornariam na forma de benefícios para a

própria sociedade (Santos, 2014). Neste contexto, *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)” (*opendefinition.org*). Em 2007, um grupo de ativistas definiu 8 princípios básicos que têm servido de diretriz 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 a tempo, acessíveis, processáveis por máquina, não discriminatórios, não proprietários e de licença de uso livre (*opendatagov.org*).

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).

Este trabalho busca entender *como os dados abertos educacionais estão prontos para tirarem máximo proveito do ecossistema da 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*) e levando em conta outras características endereçadas por uma das recentes recomendações da W3C (*World Wide Web Consortium*, organização responsável pelos padrões da Web): a *Data on the Web Best Practices* (W3C, 2017), usada neste trabalho para balizar a análise.

## **2. Dados abertos educacionais**

Os dados abertos educacionais são relativamente uma nova área de interesse, com potencial para o suporte ao desenvolvimento educacional e à melhoria da gestão (Bandeira et al., 2015). 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 *datasets* são disponibilizados para uso no ensino. No segundo caso, o interesse está na divulgação de dados de instituições acadêmicas sobre seu desempenho e de seus estudantes. Neste caso, Santos (2014b) classifica a geração e uso dos dados educacionais em 3 níveis: *macro* (dados em nível nacional ou regional), *meso* (nível escolar) e *micro* (nível de sala de aula ou aluno).

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 exploração econômica dos *datasets*. Pelo lado econômico, estima-se que a liberação de dados abertos educacionais pode gerar cerca de 1 trilhão de dólares na economia global (McKinsey, 2013), a partir do mapeamento de necessidades específicas de aprendizado em função do histórico de desempenho escolar de cada aluno. Sobre o uso, Meijer (2007) analisa a abertura de dados educacionais na Holanda por meio da internet e descreve seus efeitos. A

publicação de resultados de desempenho por terceiros estimulou as escolas a melhorarem suas notas nos indicadores, por sentirem os ‘olhos do público’ sobre elas, porém com o potencial de inflar seus números sem necessariamente melhorar sua efetividade educacional. 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.

### 3. Melhores Práticas para Dados na Web

Para fazer essa análise foi escolhido como referencial a recomendação da W3C *Data on the Web Best Practices* (DWBP), que traz um conjunto de práticas relacionadas à publicação e uso de dados na Web (W3C, 2017). Seu objetivo principal é fazer com que os dados sejam encontrados e compreendidos tanto por humanos quanto por computadores usando a 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.

A recomendação DWBP estende classificações como o esquema de 5 estrelas (Berners-Lee, 2006) e os princípios dos dados abertos governamentais, ao trazer práticas em um cenário mais amplo, baseada em padrões já estabelecidos na Web para a recuperação de informações. A Figura 1 demonstra um esquema geral dos principais conceitos relacionados à DWBP.



**Figura 1. Modelo dos principais conceitos e seus relacionamentos dentro da DWBP (adaptado de W3C, 2017).**

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. A distribuição representa uma forma específica do *dataset*, que contém os dados, podendo ser arquivos, APIs ou *feeds* RSS.

A DWBP é composta por 35 práticas distribuídas por 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 (BP15); escolher o nível correto de formalização (BP16);
8. Acesso aos dados (*como fornecer acesso aos dados?*): fornecer download em massa (BP17); fornecer subconjuntos do *datasets* como um todo (BP18); usar negociação de conteúdo (BP19); fornecer acesso em tempo real (BP20); fornecer dados atualizados (BP21); fornecer explicação para dados que não estejam mais disponíveis (BP22); tornar os dados disponíveis por meio de API (BP23); usar padrões da Web como fundamento para a API (BP24); fornecer documentação completa para a API (BP25); evitar mudanças que quebrem a API (BP26);
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**

Embora a literatura em dados abertos seja grande, são poucos os estudos voltados para 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. 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.

Além disso, alguns índices de acompanhamento e avaliação das iniciativas nacionais de dados abertos foram desenvolvidos 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 específica 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.

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 e Isotani (2017) trazem uma linha do tempo dos *datasets* educacionais disponíveis, indicando ano a ano quais *datasets* foram publicados e qual seu objetivo. No entanto, 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 foi adotada a metodologia de estudo de caso múltiplo, tendo como unidade de análise os *datasets*. A estratégia foi se aprofundar em poucos casos dentro do grande universo de *datasets* disponíveis, esperando que forneçam compreensão sobre a população em geral. Para tanto, deve existir o cuidado de selecionar casos representativos perante sua população. Para a seleção de casos, usamos a técnica de *casos diversos* (Gerring, 2009), em que é pretendido representar cada categoria, mas mantendo a variabilidade entre elas. Para validar o grau de abertura dos dados educacionais para as exigências da Web, usamos o *framework* de melhores práticas DWBP, da W3C. Foi utilizado como fonte de dados o Portal Brasileiro de Dados Abertos (*dados.gov.br*), portal 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. Ele disponibiliza uma API pela qual é possível recuperar informações desde o nível do portal até dados específicos das distribuições, organizações e usuários que interagem com o sistema.

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 apenas do MEC e do INEP, num total de 66 *datasets*. Como critérios adicionais de exclusão foram analisados os *datasets*: cujas distribuições estejam de fato disponíveis, não estejam duplicados e que tenha mais de uma distribuição, caindo para 35 *datasets*. Depois, foi feita

classificação manual 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 *dataset* que apresentou mais distribuições ao longo do tempo. Para cada *dataset* foi selecionada sua distribuição mais recente para ser avaliada de acordo com a aderência a cada uma das práticas do DWBP, usando para a avaliação os critérios: “Sim”, “Não” e “Parcialmente”. Assim, buscou-se capturar os *datasets* mais duradouros e suas distribuições mais recentes, que reflitam as últimas mudanças na publicação dos dados, considerando a diversidade e especificidades de cada categoria.

## 6. Resultados

A Tabela 1 traz os resultados da aplicação do método, avaliando cada prática estabelecida no *framework* DWBP de acordo com os critérios estabelecidos.

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. 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. O mesmo se aplica à *procedência dos dados*, que estão disponíveis nas respectivas páginas, mas sem acesso automático. A divulgação da *qualidade dos dados* é uma prática que não foi atendida por nenhum dos *datasets*, nem mesmo em suas respectivas páginas.

Quanto ao *versionamento dos dados*, 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 GUIDs. 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.pgj.gov.br/api/1/fonte/155.json> referencia o gabinete do ministro do Ministério da Educação).

Em relação aos *formatos dos dados*, 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.

**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
Proveniência 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
	7	Fornecer indicador de versão	N	N	N	N
Versionamento dos dados	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
Vocabulários dos dados	15	Reusar vocabulários, de preferência padronizados	N	N	N	N
	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	?	?	?	?
	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	?	?	?	?
	28	Avaliar a cobertura do <i>dataset</i>	?	?	?	?
Feedback	29	Coletar feedback dos consumidores dos dados	P	P	P	P
	30	Tornar o feedback disponível	N	N	N	N
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 publicador original	?	?	?	?
	34	Seguir os termos de licença	S	S	S	S
	35	Citar a publicação original	N	N	N	N

Os *vocabulários de dados* são adotados em todos os casos selecionados, fazendo parte inclusive da regulamentação da INDA<sup>1</sup>, que especifica quais metadados são obrigatórios e quais são opcionais quando da publicação dos *datasets*. Porém, estão disponíveis somente nas páginas e não para processamento automático. Além disso, não seguem os requisitos semânticos para a anotação dos metadados.

<sup>1</sup> 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/>.



O *acesso aos dados* também traz grandes desafios. O download em massa e a API versionada e com documentação referenciada 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.

Quanto ao *feedback*, o portal fornece apenas a opção de compartilhar o endereço do *dataset* em redes sociais (Facebook, Twitter e Google+). No entanto, os comentários não são salvos no portal e exibidos para que outros usuários possam comentar em cima. O *enriquecimento dos dados* também não esteve presente para os *datasets*. Não existem campos calculados nos *datasets* analisados e nem formas complementares para a visualização dos dados. Em relação à *republicação* de dados, os *datasets* educacionais usam informações de identificação de municípios e suas regiões geográficas vindas do IBGE, replicando a licença de dados original, porém não é apontado qual a fonte original desses dados.

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. Outro ponto foi o fornecimento de feedback ao publicador original (BP33), uma vez que os órgãos analisados são os publicadores originais desses dados.

## 7. Discussão

Os *datasets* escolhidos são amostras relevantes do universo de dados educacionais, pois foram adotados 3 dos *datasets* mais requisitados via e-SIC, representando mais da metade das requisições de dados por parte dos cidadãos<sup>2</sup> e de diversidade quanto à sua aplicação (Santos, 2014). 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 W3C<sup>3</sup>, que aponta como os softwares de catálogo de dados mais comuns atendem às práticas desse framework. 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. Assim, são listadas possíveis frentes de trabalho para a adequação dos dados educacionais à DWBP.

- Portal de dados: parte das lacunas levantadas podem ser endereçadas pelo software que hospeda os dados. Por exemplo, no caso brasileiro, extensões da plataforma CKAN, usada no portal *dados.gov.br*, as BPs 1, 2 e 3 (metadados) podem ser atendidas pela extensão *dcat*, que era parte da instalação nativa do CKAN, mas que foi extraído em suas versões mais recentes; BP6 com a

---

<sup>2</sup> [http://download.inep.gov.br/institucional/legislacao/2016/portaria\\_n370.pdf](http://download.inep.gov.br/institucional/legislacao/2016/portaria_n370.pdf)

<sup>3</sup> <http://w3c.github.io/dwbp/dwbp-implementation-report.html>

extensão *qa*, que traz informações automáticas sobre a qualidade dos dados; BP32 com os *viewers* para a visualização de dados, dentre outros. Alguns *datasets* mais recentes trazem algumas dessas funcionalidades, como a visualização de dados e API de consulta – porém não nos dados educacionais;

- Adequação na publicação dos dados: os *datasets* e suas distribuições trazem metadados que atendem parcialmente às práticas, por não os disponibilizarem em formato semântico (e.g. BP5, BP15, BP16) ou de forma automática (via API). Inclusive, alguns dos metadados são parte da INDA – sendo campos obrigatórios ou não, sendo uma questão de como apresentá-los, dado que já existem. Caso fossem transformados para o formato semântico, seria possível processá-los por computador mais facilmente;

De forma geral, os resultados sugerem que o portal *dados.gov.br* parece ter evoluído para o consumo de dados por humanos, trazendo os dados, metadados, formatos de arquivos e informações auxiliares para que pessoas possam baixar, compreender e usar os dados disponíveis. Por exemplo, temos o caso do censo escolar, que traz um arquivo zip, composto por outros arquivos de dados compactados, metadados em formato de planilha e informações sobre a metodologia em formato PDF. Para um usuário, é conveniente de se baixar todo o conjunto de uma vez. Pelo outro lado, isso torna muito complexo seu processamento automático e em tempo real (BP12).

Deste modo, espera-se que a análise feita neste trabalho e as intervenções apontadas possam trazer benefícios para a adequação dos dados abertos educacionais brasileiros para atingir o potencial que a Web oferece. Atualmente, o MEC e o INEP desenvolvem suas respectivas ações de dados abertos no escopo do Plano de Dados Abertos (o atual é de 2016/2018), com as estratégias, matriz de responsabilidades e os dados e metadados a serem publicados, e algumas dessas práticas poderiam estar presentes em uma próxima edição.

## Agradecimentos

Os autores agradecem à CAPES, CNPq e FAPESP pelo apoio financeiro dado ao projeto.

## Referências

- Atenas, J., Havemann, L. (2015). “Open Data as Open Educational Resources: Case studies of emerging practice”. London: Open Knowledge, Open Education Working Group. <http://dx.doi.org/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.
- 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.
- Berners-Lee, T. (2006). Linked data: design issues. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acessado em: 17 jun. 2017.
- Cucos, R. (2013). “Open Government Data: Helping Parents to find the Best School for their Kids”. The WorldBank, 2013. Disponível em: <http://blogs.worldbank.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:
- 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;
- Isotani, S., Bittencourt, I. I. (2015). “Dados abertos conectados”. Editora Novatec. DOI: <http://doi.org/10.13140/RG.2.1.4355.6329>.
- McKinsey. (2013). “Open data: Unlocking innovation and performance with liquid information”. McKinsey Global Institute. [S.l.]. 2013.
- 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: <http://doi.org/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.
- 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.
- Penteado, B. E., Isotani, S. (2017). “Dados abertos educacionais: que informações temos disponíveis?”. *Anais do VI Congresso Brasileiro de Educação*. Bauru, Julho de 2017.
- 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.
- Santos, O. A. R. (2014b). “What has open data got to do with education?”. *Open Education Working Group Community Session*, Open Knowledge Foundation. Disponível em: <https://pt.slideshare.net/otavios/open-education-data-research>. Acesso: 16 de junho de 2017.
- 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.
- 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.
- 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.
- 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”. *13th International Conference on Digital Government Research*. Jun 2012, Maryland, EUA. p. 84-89.