Dados Abertos Educacionais Brasileiros: Um Mapeamento Sistemático da Literatura

Leandro A. Ferreira[1], Rodrigo L. Rodrigues[2], Rodrigo N. P. M. de Souza[3]

[1,2,3]Programa de Pós-Graduação em Tecnologia e Gestão em Educação a Distância (PPGTEG)
Universidade Federal Rural de Pernambuco

{leandro.aferreira, rodrigo.linsrodrigues, rodrigo.npmsouza} @ufrpe.br

Abstract. *Brazilian open educational data have important information for the entire educational context, but still little used. The area has been growing in recent years, but it still does not generate enough content for all levels of Brazilian education. This systematic mapping covered scientific papers that used Brazilian open educational data. The study identified that in the last three years the interest in this area of research has increased, but the majority focused on higher education, data from elementary and secondary education have been little explored. This article presents an overview of the use of open educational data in Brazil, demonstrating that there is still a lot to invest in this area of research.*

Resumo. *Dados abertos educacionais brasileiros carregam informações importantes para todo o contexto educacional, mas ainda pouco utilizados. A área vem crescendo nos últimos anos, mas ainda não gera conteúdo suficiente para todos os níveis da educação brasileira. Este mapeamento sistemático abrangeu artigos científicos que tratam de dados abertos educacionais brasileiros. O estudo conseguiu identificar que nos últimos três anos aumentou o interesse por esta área de pesquisa, mas a maioria focada no ensino superior, dados do ensino fundamental e médio vêm sendo pouco explorados. Este artigo traz um panorama da utilização de dados abertos educacionais no Brasil, demonstrando que ainda há muito o que investir nessa área de pesquisa.*

1. Introdução

O Governo brasileiro por meio do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP, vinculado ao Ministério da Educação - MEC é responsável por promover estudos, pesquisas e avaliações educacionais. É de responsabilidade do mesmo a produção e disponibilização de dados abertos educacionais em seu portal para gestores, educadores e sociedade em geral (BRASIL, 2011). O INEP[1] coleta, armazena e disponibiliza dados de diversos níveis de ensino, desde a Educação Básica até o Ensino Superior.

No Brasil a Lei de Acesso à Informação - LAI (nº 12.527) garante à sociedade o acesso a essas informações públicas, no entanto ainda há uma falta de maturidade e padronização por parte do governo para disponibilizar os dados para a sociedade, dessa forma os dados é disponibilizado, mas o cidadão não tem acesso total às informações, por não ter conhecimento técnico para a manipulação de tais bases de dados (PEDROSO *et al,* 2013).

[1]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos

Embora a capacidade de manipular base de dados, em busca de informação útil para tomada de decisão, ainda é uma habilidade pouco desenvolvida por profissionais da educação. Para Davenport e Patil (2012) trabalhar com dados será uma habilidade essencial dessa última década, as instituições estão gerando uma grande quantidade de dados, e a necessidade de minerar e analisar esses dados pode motivar tomadas de decisões e nortear a própria sociedade nas tomadas de decisões mais assertivas.

Apesar das informações contidas nos mesmos não estão estruturadas, sendo essa uma das dificuldades com a extração de informações provindas desses dados, ao ponto que todos tenham acesso às informações, é necessário um conhecimento técnico para poder acessá-los e manipulá-los (ISOTANI e BITTENCOURT, 2015).

Tendo em vista essa problemática, este trabalho buscou realizar um mapeamento sistemático, no sentido de proporcionar uma visão ampla de como a comunidade científica, interessada em dados abertos educacionais, está realizando pesquisas a partir da disponibilização de dados educacionais abertos brasileiros.

Este trabalho está dividido nas seguintes seções: na seção 2, são apresentados detalhes sobre dados abertos educacionais; na seção 3, a metodologia desenvolvida neste trabalho; na seção 4 descrevemos os resultados; por fim, nas seções 5 e 6 são apresentadas discussão e considerações finais sobre os resultados alcançados.

2. Dados Abertos Educacionais

Para a *Open Knowledge International* (2012) um dado aberto é aquele que pode ser utilizado livremente, reutilizado e redistribuído. Berners-Lee (2006) propõe uma classificação em cinco estrelas, para que o dado seja considerado aberto, as estrelas classificam a qualidade do dado: uma estrela quando esse dado precisa está disponível na *Web*, tem que ser garantido uma licença aberta; duas estrelas neste nível precisam esta em formato estruturado, nesse caso necessitam de softwares proprietários (por exemplo, Excel); três estrelas, a disponibilidade será em formatos não estruturado um exemplo são arquivos CSV que podem ser manipulados por diversos softwares; quatro estrelas, quando é utilizam padrões recomendados pelo Consórcio *World Wide Web* (W3C), os dados precisam possuir identificadores URI (*uniform resource identifier*) utilizam identificadores para apontamentos na *Web*; cinco estrelas, quando dados estão conectados com outros dados, dessa forma gerando um *link* entre os mesmo.

O INEP utiliza o princípio da *Open Knowledge International* (2012) e até a terceira estrela dos propostos por Berners-Lee (2006), o mesmo disponibiliza um grande volume de dados em todos os níveis da educação brasileira, Rigotti *et tal.* (2015), destaca a atualização anual das bases do INEP e a importância desses dados para área educacional. As informações contidas nessas bases, podem agregar valor para a melhoria do processo ensino-aprendizagem e tomada de decisão.

Dourado *et tal.* (2017) destaca a importância de criar mecanismos em que as instituições possam verificar o desempenho escolar de forma contínua e em tempo real, por meio de técnicas de análises e extração de dados. O uso da mineração de dados na educação cresceu nos últimos anos, com o aumento considerável na quantidade de dados, avanços tecnológicos nas ciências da computação e ferramentas bem desenvolvidas de análise.

Desta forma a utilização de dados abertos abre um grande leque de possibilidades principalmente para o setor educacional. O Brasil produz várias bases de dados educacionais onde constam dados relacionados às avaliações, informações como matrículas, número total de estudantes, aspectos físicos da instituição, bem como características sócio-econômica dos envolvidos, entre outros. Nesse contexto é necessário investir em soluções para que toda a comunidade educacional possa ter acesso a tais dados, apesar da grande disponibilidade de dados, mas o consumo ainda não é feito de forma adequada para a construção de políticas públicas educacionais, bem como tomadas de decisões pedagógicas.

3. Metodologia

Diante desta realidade, esta pesquisa foi estruturada a partir de um mapeamento sistemático seguindo o processo descrito por Petersen *et al*. (2008), segundo o autor, há cinco passos importantes a serem seguidos nesse processo: primeiro a definição de pesquisa; segundo definição da *string* de busca; terceiro escolha das bases de dados que serão utilizadas na pesquisa e critérios de inclusão e exclusão; quarto análise dos trabalhos selecionados; quinto extração dos dados e conclusão do mapeamento.

As questões de pesquisas que norteiam este mapeamento são:

QP^1 - Quais os anos com a maior predominância de trabalhos publicados?

QP^2 - Quais bases de dados públicas foram utilizadas?

QP^3 - Quais tecnologias foram utilizadas para o processo de análise dos dados?

QP^4 - Quais os métodos ou metodologias específicas foram utilizadas?

QP^5 - Os estudos buscaram resolver problemas ao nível nacional ou local?

QP^6 - Os trabalhos focaram no desenvolvimento de alguma ferramenta/produto educacional?

QP^7 - Forma de publicações: anais de congresso, conferências ou revistas?

Após a definição das questões de pesquisa foram realizados diversos testes com palavras chaves para refinar a *String* de busca, a mesma foi dividida em duas categorias, uma relacionada a Mineração de Dados Educacionais - MDE e a outra relacionada a bases de Dados Abertos Educacionais do Brasil com os idiomas em Português e Inglês.

Tabela 01 - String de busca

String de busca em Português	String de busca em Inglês
("mineração de dados") AND ("censo escolar" OR "enade" OR "enem" OR "prova brasil" OR "saeb" OR "encceja")	("educational data mining" OR "data mining") AND ("school census" OR "enade" OR "enem" OR "proof Brazil" OR "saeb" OR "encceja")

A extração dos artigos foi realizada nas seguintes bases de dados de artigos científicos: IEEE, ScienceDirect, ACM Digital, SpringerLink, Scopus, Periódicos da Capes e Google Acadêmico. Especificamente para o Google Acadêmico foi utilizado o filtro de avançar até as páginas 40 em Inglês e Português, foram selecionados pelos

títulos 178 artigos. Na etapa de seleção e classificação foram aplicados os seguintes critérios de exclusão e inclusão:

Critérios de Exclusão

Os seguintes tipos de estudos serão excluídos:

I. Artigos que não apresentam estudos referentes a dados educacionais brasileiros;
II. Artigos escritos em qualquer idioma, exceto em Inglês ou Português;
III. Trabalho de conclusão de graduação e pós-graduação (ou seja TCC´s, monografias, dissertações e teses);
IV. Artigos não acessíveis na íntegra;
V. Estudos duplicados: apenas o mais atual será incluído;
VI. Tutoriais, *keynote speech*, relatórios de workshop, relatórios técnicos, estudos secundários e terciários (ou seja, revisões sistemáticas da literatura e mapeamentos de estudos).

Critérios de Inclusão

Os seguintes tipos de estudos foram incluídos:

I. Artigos científicos que tratam sobre dados abertos educacionais brasileiros.
II. Estudos publicados de 2010 até junho de 2021.

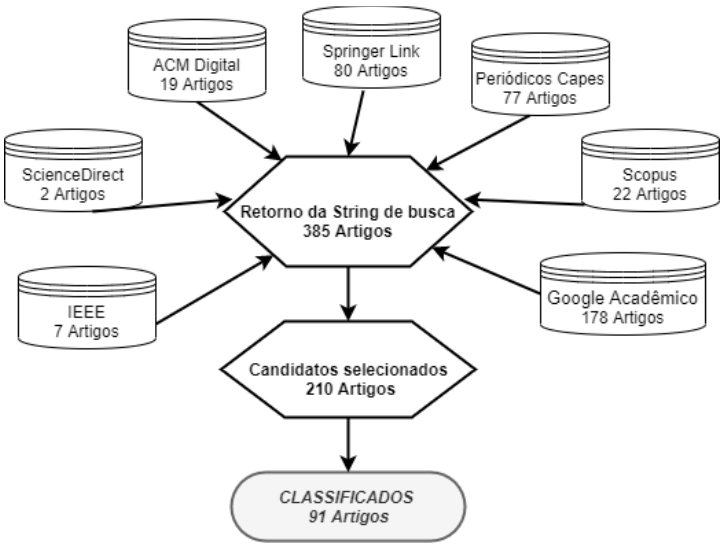

Figura 1. Seleção e classificação dos artigos

Após aplicar os critérios de exclusão e inclusão nas bases escolhidas foram realizadas as leituras dos principais campos título, resumo ou *abstract* e palavras chaves foram escolhidos 210 candidatos, classificando 91 artigos após a leitura do artigo completo.

4. Resultados e Discussões

Esta seção abordará os resultados das questões de pesquisas apresentadas na seção 3 metodologia.

QP[1] - Quais os anos com a maior predominância de trabalhos publicados?

Penteado e Isotani (2017), destacam que os pesquisadores brasileiros têm investido na publicação de dados abertos, o governo federal nos portais do MEC e INEP, tem disponibilizado e atualizado as bases de dados educacionais, os últimos anos nos mostram um interesse crescente na área, a **Figura 2** nos mostra que em 2019 tivemos o maior número de publicações, seguido por 2020 e 2018.

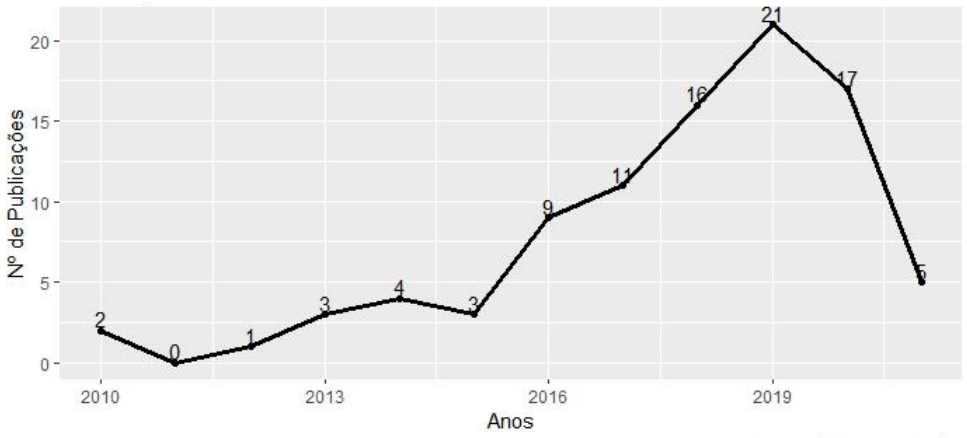

Figura 2. Publicações por ano

Apesar da **Figura 2** indicar um crescimento nos últimos três anos, é necessário avançar mais em pesquisas com dados abertos educacionais, as redes de ensino necessitam de respostas sobre as situações de ensino, socioeconômico, infraestrutura, corpo docente e etc.

QP[2] - Quais bases de dados públicas foram utilizadas?

Foi possível observar que no cenário brasileiro, de acordo com **Figura 3**, as bases de dados abertas educacionais mais utilizadas, foram o ENEM, seguido pelo ENADE e por fim entre os três mais bem colocados aparecem as bases da Prova Brasil e o Censo de Educação Básica.

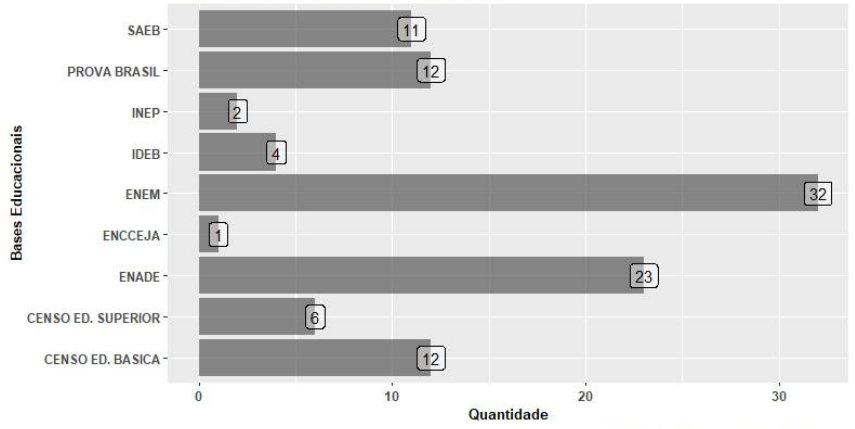

Figura 3. Bases de Dados Abertos Educacionais mais utilizadas no Brasil

Dessa forma podemos notar que os pesquisadores realizaram estudos focados no ensino superior, isso é algo preocupante tendo em vista que a rede de ensino fundamental e médio é extensa e necessita de ações para a melhoria dos processos educacionais. Com a extração desses dados, é possível tratar aspectos como por exemplo, a evasão, com técnicas específicas e tomadas de decisão é possível investir esforços para melhorar os diversos índices educacionais (MANHÃES et tal. 2011 e RIGO et tal. 2012).

QP^3 - Quais tecnologias foram utilizadas para o processo de análise dos dados?

Simon e Cazella (2017) utilizaram o Software WEKA com algoritmo J48 para minerar dados do ENEM de 2015 com a finalidade de extrair o desempenho dos indicadores médios em ciências da natureza e suas tecnologias dos estudantes, Braga e Drummond (2017) utilizaram o Software R para mineração de dados abertos da base do ENEM de 2013, onde concluíram que as disciplinas que os estudantes tinham mais dificuldades foram química, física e biologia. A **Figura 4** mostra os softwares mais utilizados na mineração de dados educacionais, WEKA, acompanhado do Software R, seguindo os mais bem colocados com o Software estatístico *Statistical Package for Social Sciences* (IBM SPSS).

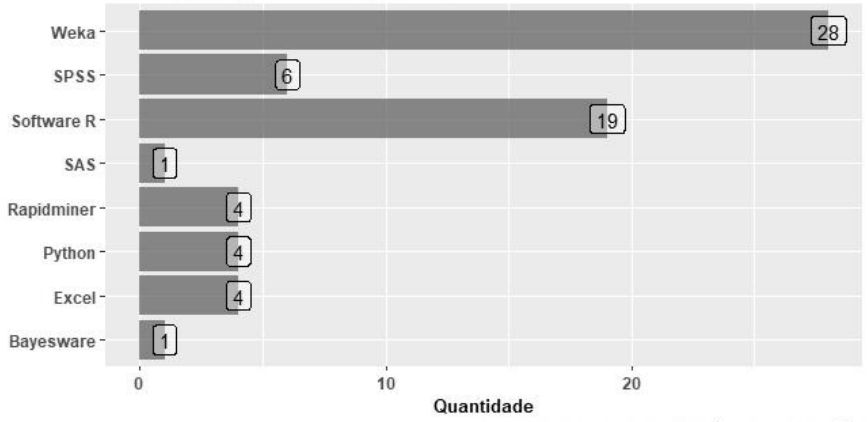

Figura 4. Software mais utilizados em MDE

Simon e Cazella (2017) concluíram em seu estudo utilizando o algoritmo J48 que o desempenho médio dos estudantes em ciências da natureza e suas tecnologias, com pontuação acima de 550 pontos, ocorrem nas escolas privadas, apenas estudantes com nível socioeconômico muito alto, e nas federais, com níveis muito alto, alto e médio alto, já em escolas estaduais, apenas com níveis muito alto, por fim nas municipais apenas com níveis médio alto.

A **Figura 5**, apresenta os algoritmos mais citados nas pesquisas, J48 e NaiveBayes foram os mais citados.

Figura 5. Nuvem de palavras dos Algoritmos mais citados

WEKA e o Software R têm um ponto em comum, por não serem privados, são amplamente utilizados nos estudos.

QP[4] - Quais os métodos ou metodologias específicas foram utilizadas nos estudos?

Camilo e Silva (2009), cita o *Knowledge Discovery in Databases* - KDD como um processo que identifica padrões válidos. Moro *et al*. (2011), define a metodologia *Cross-Industry Standard for Data Mining* - CRISP-DM como uma sequência de seis fases, a mesma permite que seja implementado um modelo de mineração que possa ser usado em um ambiente real, dessa forma ajudando as decisões de negócios. Conforme podemos observar os métodos ou metodologias mais utilizados em mineração de dados educacionais nos estudos, foram o KDD com 54% das pesquisas classificadas e o CRISP-DM com 46% dos estudos.

QP[5] - Os estudos buscaram resolver problemas ao nível nacional ou local?

A exposição das pesquisas foi bem maior no cenário nacional, atingindo um percentual de 64% dos trabalhos e os demais estudos procuram realizar pesquisas com problemas locais, ficando com 36% dos trabalhos. Isotani e Bittencourt (2015), destacar a importância da exposição de dados locais, tais dados auxiliam a comunidade na tomada de decisões e comparação de determinados índices.

Outro fator preocupante é a falta de pesquisas para resolver problemas locais, o questionamento a ser feito será pela preocupação com o consumo de tais dados? Mas partindo de outra visão o impacto em instituições locais é de suma importância para o seu crescimento regional. Desta forma, ainda é carente a quantidade de pesquisas que buscam resolver problemas em uma cidade ou região específica, através da utilização de dados educacionais abertos.

QP[6] - Os trabalhos focaram no desenvolvimento de alguma ferramenta/produto educacional?

Lima et al. (2009), propõe em seu estudo a criação de um protótipo aliado ao uso de dados abertos para apoiar a tomada de decisão, dessa forma possibilitará acessar vários dados como minera-los com filtros, que se ajustem às necessidades do gestor,

para que as decisões tomadas sejam as mais exatas possíveis. A **Tabela 2** mostra que poucos trabalhos desenvolveram ferramentas ou produtos para visualização de dados com bases abertas no Brasil.

Tabela 02 - Ferramenta/Produto Educacional

Autor(es)	Ferramenta/Produto Educacional	Base Educacional
VICTORINO, Marcio et al.	Protótipo de um Sistema de Apoio à Decisão (dashboards)	Censo de Educação Superior
LIMA, Priscila da Silva Neves et al.	Protótipo é denominado SysEnade	ENADE
LEMOS, Robson Rodrigues et al.	Aplicação WEB - VisDadosEnem	ENEM
LIMA, Priscila et al.	Protótipo em JAVA	ENADE e ENEM
MARTINS, Cristina; DE FARIA, Derlei E.	Ferramenta BI	ENEM
PINHEIRO, Rodrigo Guedes Pereira; ELIA, Marcos; SAMPAIO, Fábio F.	Ferramenta ACHA	Prova Brasil
RIVEROS, Lilian Jeannette Meyer; FERNÁNDEZ, Carlos Manuel Reyes; JUNIOR, Nereu Mokwa.	Protótipo de um Sistema de Apoio à Decisão (dashboards)	ENADE

A **Tabela 2** deixa claro a necessidade de pesquisas com dados abertos educacionais voltadas para a educação básica, das 7 pesquisas que criaram alguma solução tecnológica, somente 1 desenvolveu projeto com a educação básica. Essa rede de educação é a que mais sofre por falta de pesquisas, apesar de ter tantos dados disponíveis, mas não são utilizados para tomadas de decisões.

Isso é muito preocupante e deixa alguns questionamentos, por que não "olhar" para a educação básica? Por que não desenvolver soluções e exposição de dados para essa rede? A resposta pode ter relação ao consumo dos dados gerados, se observamos a rede de educação básica ela tem um grande potencial para consumir tais dados, é uma rede ensino muito maior que a do ensino superior.

QP[7] - Forma de publicações: anais de congresso, conferências ou revistas?

A maior parte de publicações estão concentradas em revistas com 51% dos estudos publicados, os outros 49% se dividem entre anais, congressos, conferências e simpósios. Acreditamos que outra forma de divulgação são os relatórios locais das

secretarias de educação. Esses ambientes de divulgação não foram englobados no escopo desta pesquisa, embora reconhecemos que pode ser uma limitação deste estudo.

5. Considerações finais

Esse trabalho apresentou um mapeamento sistemático sobre os Dados Abertos Educacionais no Brasil. Buscou-se analisar quais as bases educacionais abertas utilizadas entre 2010 até junho de 2021, o mapeamento trouxe grandes contribuições, o estudo conseguiu identificar as bases mais utilizadas, ENEM e ENADE estão no topo da lista dos pesquisadores, seguido da Prova Brasil. Uma preocupação é com a falta de pesquisas na Educação Básica e Encceja, autores chegaram a trabalhar com duas bases no mesmo estudo. Outro fator observado é a falta de ferramentas ou produtos destinados a visualização dos dados educacionais, é necessário investir mais em soluções que possam dar aos gestores, professores ou comunidade em geral condições de analisar os dados e tomar decisões para a melhoria do processo ensino aprendizagem.

A maioria dos estudos procuraram expor os dados utilizando o cenário nacional, isso mostra pouca expansão em estudos de caso locais, foi observado que entre 2010 a 2017 se tinha poucas pesquisas com dados abertos educacionais, 2018 teve um aumento significativo chegando ao ápice em 2019 mas voltando a cair em 2020, dessa forma demonstrando se a área está ou não aquecida.

Ficou claro com este mapeamento, que esta área de pesquisa ainda precisa avançar, trazendo soluções para que a comunidade educacional possa visualizar os dados abertos e tomar decisões para melhoria dos seus processos educacionais.

6. Agradecimentos

Ao Programa de Pós-Graduação em Tecnologia e Gestão em Educação a Distância (PPGTEG) da Universidade Federal Rural de Pernambuco e ao Grupo de Pesquisa Laboratório de Evidências Analíticas em Tecnologias Educacionais - EVANTE.

7. Referências

BERNERS-LEE, Tim. (2006) Linked Data. https://www.w3.org/DesignIssues/LinkedData.html, Junho.

BRAGA, L. C. C.; DRUMMOND, I. N. (2017) Uma Abordagem de Mineração Descritiva Aplicada a Dados Abertos Governamentais Empregando a Ferramenta R. Anais do Computer on the Beach, p. 51–60.

BRASIL. (2011) Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). PDE/PROVA BRASIL Plano de Desenvolvimento da Educação 2011, Brasília, http://www.portal.mec.gov.br/dmdocuments/prova%20 brasil_matriz2.pdf, Junho.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. (2009) Mineração de dados: Conceitos, tarefas, métodos e ferramentas, Universidade Federal de Goiás (UFC), v. 1, n. 1, p. 1-29.

DAVENPORT, Thomas H.; PATIL, D. J. (2012) Data scientist. Harvard business review, v. 90, n. 5, p. 70-76.

DOURADO, Raphael et al. (2017) Novas possibilidades de avaliação em larga escala na educação básica através do uso de EDM e Learning Analytics. In: *Anais do VI Workshop de Desafios da Computação aplicada à Educação. SBC*.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. (2015) Dados Abertos Conectados: Em busca da Web do Conhecimento. Novatec Editora.

LIMA, Cristovão et al. (2019) SADLABI: Proposta de um Sistema de Apoio à Decisão para a gerência de Laboratórios de Informática. In: *Anais do XXV Workshop de Informática na Escola. SBC*, p. 1479-1483.

MANHÃES, Laci Mary Barbosa et al. (2012) Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In: *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)*.

MORO, S., LAUREANO, R., CORTEZ, P. (2011) Using data mining for bank direct marketing: An Application of the crisp-dm methodology. EUROSIS. *Proceedings of European Simulation and Modelling Conference-ESM*, page 117–121.

OPEN KNOWLEDGE. (2012) Open data handbook, http://opendatahandbook.org/, Junho.

PEDROSO, Louise; TANAKA, Asterio; CAPPELLI, Claudia. (2013) A Lei de Acesso à Informação brasileira e os desafios tecnológicos dos dados abertos governamentais. *In*: *SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO (SBSI)*, 9., João Pessoa. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2013 . p. 523-528.

Penteado, B. E., Isotani, S. (2017) Dados abertos educacionais: que informações temos disponíveis? VI Congresso Brasileiro de Educação, vol. 4, p. 1933-1938.

PETERSEN, K., FELDT, R., MUJTABA, S., and MATTSSON, M. (2008) Systematic mapping studies in software engineering. In Ease, volume 8, pages 68–77.

RIGO, Sandro J.; CAZELLA, Silvio C.; CAMBRUZZI, Wagner. (2012) Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In: Anais do Workshop de Desafios da Computação Aplicada à Educação. p. 168-177.

RIGOTTI, José Irineu Rangel; CERQUEIRA, Cézar Augusto. (2015) As bases de dados do INEP e os indicadores educacionais: conceitos e aplicações. Livros, p. 71-88.

SIMON, Augusto; CAZELLA, Sílvio. (2017) Mineração de Dados Educacionais nos Resultados do ENEM de 2015. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. p. 754.